The Price of Forgetting: Data Redemption Mechanism Design for Machine Unlearning

Yue Cui and Man Hon Cheung

***Abstract*—Nowadays, technology companies spend efforts collecting datasets from massive users and training machine-learning models to enable innovative artificial intelligence (AI) applications. However, current data protection policies, such as General Data Protection Regulation (GDPR), enforce *the right to be forgotten* and require the server to perform *machine unlearning* and eliminate the effect of users' data on the trained model once receiving the data redemption requests. Such privacy regulations are unfair to the server, as it incurs extra costs for unlearning but suffers from a degraded model. In this paper, we propose the first incentive mechanism in machine unlearning to compensate for the server's cost to the best of our knowledge. We first characterize the accuracy degradation and consumed time as a function of the unlearning ratio by conducting experiments on three popular datasets and two widely used unlearning algorithms. Then, we model the interaction between the server and users as a two-stage Stackelberg game. We propose an iterative algorithm to optimize the unit price for data redemption by characterizing the convexity of server's profit maximization problem. The experimental results on real datasets show that our mechanism can achieve the largest server's profit and social welfare, compared with the GDPR and no redemption schemes.**

I. Introduction

A. Motivations

Technology companies, such as Facebook and Google, spend significant efforts to collect large datasets from a massive number of users, invest a substantial amount of training resources (e.g., time, hardware, and software), and train sophisticated artificial intelligence (AI) models to benefit our lives these days [1], [2]. However, some private information can be embedded in the well-trained models in a complex way, such as users' medical records [3] and personal emails [4]. Therefore, recent data protection regulations, such as General Data Protection Regulation (GDPR) [5] in E.U. and California Consumer Privacy Act (CCPA) [6] in the U.S., enforce *the right to be forgotten*. They require the companies to delete the collected data samples and remove their effects on the models immediately to eliminate the potential privacy leakage risks once the users submit their data redemption requests. To meet their requirements, *machine unlearning* [7], [8] has arisen as a compelling approach that facilitates the removal of data's effect from well-trained machine-learning models in practice.

Following these strict privacy regulations may cause financial loss for AI companies [9] as the company incurs costs for performing machine unlearning while suffering from the accuracy degradation of models [7], [10]. First, fully retraining the model from scratch [8] is the most legitimate way to remove the targeted data completely from the AI model but is extremely *time-consuming* [11]. Sharded, Isolated, Sliced, and Aggregated training (SISA) proposed by Bourtoule *et al.* [7] is an efficient and general way to speed up the machine unlearning in an ensemble way, but the consumed time is still significant, especially for large models and frequent requests [12]. Additionally, the *accuracy degradation* of models is another key factor preventing the server from performing machine unlearning. The accuracy degradation can be exponential [10] when more data is unlearned, which is referred to as catastrophic unlearning [13], and how to mitigate it naturally is still an open question [14].

Clearly, the server has no incentive to perform the machine unlearning honestly and consciously without sufficient compensation from users who request to redeem data. A survey [15] shows that over 27% of companies from the U.S. and E.U. have yet to begin work on making their organization GDPR-compliant more than two years after the deadline has passed, and 60% of technology companies are not ready for compliance. If we force the companies to follow existing regulations, it may hinder innovative data usage and AI investments since the companies foresee the unlearning cost in the future. Therefore, designing an economic mechanism to compensate for the server's unlearning cost sufficiently is essential.

B. Contributions

To the best of our knowledge, we propose the first incentive mechanism for machine unlearning in this paper. First, we characterize the accuracy degradation in an exponential function and model the consumed time as a linear function by fitting the experimental results obtained from three real-world datasets (i.e., MNIST [16], Cifar-10 [17], and Adult Income [18]) and two popular unlearning algorithms (i.e., Retraining from Scratch [8] and SISA [7]).

Next, we model the data redemption mechanism of the server and users as a two-stage Stackelberg game. Stage I is the server's profit maximization problem that optimizes the unit price for redeeming data, and Stage II is the users' payoff maximization problem that optimizes the data redemption amount. To design a fair mechanism in machine unlearning, the most important question to answer is *how much should the server charge the users for their data redemption*? If the price is too low, users may redeem a significant amount (or even all) of their data, resulting in a significant drop in the server's model accuracy. However, if the price is too high, it may not lead to a

Y. Cui and M. H. Cheung are with the Department of Computer Science, City University of Hong Kong, HKSAR; E-mails: yuecui8-c@my.cityu.edu.hk, mhcheung@cityu.edu.hk. This work is supported by the City University of Hong Kong's Research Grant under Project 7005994. It is also supported by the Early Career Scheme (Project Number CityU 21206222) established under the University Grant Committee of the Hong Kong Special Administrative Region, China.

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622287

larger profit for the server since unlearning an extremely small amount of data is also time-consuming.

Analyzing the Stackelberg game is highly non-trivial as it involves the two-stage interactions between the server and users. Nevertheless, we characterize the conditions under which Stage I is a convex problem, decompose the problem into several sub-problems based on their convexities, and propose an algorithm to solve the problem based on the gradient descent method [19].

We summarize our major contributions as follows.

- *Incentive mechanism design for machine unlearning*: To the best of our knowledge, this is the first incentive mechanism to compensate the server for machine unlearning, while the existing policies (such as GDPR and no redemption) can be considered as special cases of our mechanism.
- *Experimental characterization of unlearning costs*: We characterize the accuracy degradation and consumed time in data redemption amount with three real-world datasets (i.e., MNIST [16], Cifar-10 [17], and Adult Income [18]) and two widely adopted machine unlearning algorithms (i.e., Retraining from Scratch [8] and SISA [7]).
- *Challenging two-stage optimization problem*: We formulate the decisions of the server and users as a two-stage Stackelberg game. The analysis is non-trivial given the complexity of their interactions. We derive a closed-form solution for users in Stage II, substitute it into Stage I, characterize the convexity of Stage I, and propose an iterative algorithm to obtain the optimal unit price for the server.
- *Insights and performance evaluation from real data*: The experimental results show that our mechanism achieves the largest server's profit and social welfare compared with GDPR and No Redemption baselines. Moreover, we find that it is profitable for the server when the informed ratio increases.

C. Related Work

Machine unlearning [7], [8], [12] aims to adapt the trained model to a new one that achieves equivalent (or at least comparable) performance on its task but trained on the dataset that excludes the redeemed data. Existing machine unlearning studies (e.g., [10] and the references therein) mainly focus on accelerating the unlearning process or improving the performance of the unlearned model from a technical perspective. However, none of them consider an economic mechanism to incentivize the server to perform the machine unlearning. Ding *et al.* [20] proposed an incentive mechanism for federated unlearning, where the server allocates the rewards and motivates the clients to perform the unlearning algorithm collaboratively. Different from [20], we consider the centralized machine unlearning, where server is the unlearning executor that requires incentives.

Another related line of research is the economics of privacy [21]–[23], which discussed the trade-off to share and protect users' data under various technical backgrounds (e.g., search engines and social media). However, none of them considered the economic mechanism under machine unlearning. Whether *the right to be forgotten* can support users' privacy without hampering social welfare is still an open problem [24].

II. System Model

In this section, we first introduce the machine unlearning setting in Section II-A. Next, we formulate the user's utility in Section II-B and server's unlearning cost in Section II-C. Then, we formulate the interaction between the server and users as a two-stage Stackelberg game in Section II-D.

A. Machine Unlearning Setting

We consider a machine unlearning setting in which multiple users participated in the model training by contributing their own data to a central server. Now, some users request to redeem their data simultaneously. Let $\mathcal{I}$ be the set of all users with $\mathcal{I} = \mathcal{I}_r \cup \mathcal{I}_n$, where $\mathcal{I}_r$ is the set of informed users that make rational decisions to redeem data, while $\mathcal{I}_n$ is the set of uninformed users who do not redeem any data [25]. We define the informed ratio as $\gamma = \frac{\mathcal{I}_r}{\mathcal{I}} \in [0, 1]$. Let $\mathcal{D} = \{\mathcal{D}_i\}_{i\in\mathcal{I}}$ be the contributed dataset by each user and $\mathcal{D}^u = \{\mathcal{D}_i^u\}_{i\in\mathcal{I}}$ be the subsets of data that each user redeems with $\mathcal{D}_i^u \subseteq \mathcal{D}_i, \forall i \in \mathcal{I}_r$ and $\mathcal{D}_i^u = \emptyset, \forall i \in \mathcal{I}_n$. Denote $d_i = |\mathcal{D}_i|$ as the number of contributed data samples and $x_i = |\mathcal{D}_i^u|$ as the number of data samples to redeem for each user $i \in \mathcal{I}$. Thus, we have $x_i \in [0, d_i]$ for $i \in \mathcal{I}_r$ and $x_i = 0$ for $i \in \mathcal{I}_n$. We assume that the data to be redeemed for each user is uniformly sampled from $\mathcal{D}_i$.

The unlearning aims at optimizing the unlearned model parameter w_u to minimize the loss of the model trained on the dataset that excludes the redeemed set [14]:

$$w_u^* = \arg\min_{w_u} \sum_{i\in\mathcal{I}} \frac{d_i - x_i}{|\mathcal{D}| - |\mathcal{D}^u|} L(w_u; \mathcal{D}_i \backslash \mathcal{D}_i^u), \tag{1}$$

where $L(w; \mathcal{D})$ is the loss function with model parameter w for dataset $\mathcal{D}$ and its form depends on the specific training task. Equation (1) shows that users' decisions can significantly affect the performance of the unlearned model.

B. User's Utility

Let $U_i(x_i)$ be the *utility function* of user i if x_i amount of data is redeemed from the server. In general, users can obtain a higher degree of utility from the larger amount of redeemed data because of the enhanced privacy level. However, when users have already redeemed sufficient data, further data redemption cannot provide significant privacy improvement because of the diminishing marginal utility of the additional data [26]. Thus, we model $U_i(x_i)$ as a concave increasing function in x_i with $U_i(0) = 0$. Following the widely adopted choice of utility function [27], [28], we define it as

$$U_i(x_i) = \lambda_i \ln(x_i + 1), \tag{2}$$

where $\lambda_i \geq 0$ is user i's marginal privacy gain. A larger λ_i implies a higher concern about privacy.

C. Server's Unlearning Cost Characterization on Real Data

We consider two important costs of the server resulting from the data redemption: *accuracy degradation* in Section II-C1 and *consumed time* in Section II-C2.

To investigate the properties of these two costs in machine unlearning, we conduct experiments on three popular datasets (i.e., MNIST [16], Cifar-10 [17], and Adult Income [18]) and use two widely adopted machine unlearning algorithms (i.e., Retraining from Scratch [8] and SISA [7]). Retraining from Scratch [8] works as the most legitimate way that retrains the model completely after excluding the redeemed dataset. SISA [7] first shards and slices the dataset to incrementally train the model, and only retrains the affected shards/slices. We first train a model until convergence on each dataset as the original model. For MNIST and Cifar-10, we use the convolution neural network (CNN) with two convolutional layers and two fully connected layers. For Adult Income, we use the multi-layer perceptron (MLP) with two layers. Then, we range the unlearning ratio (i.e., $\sum_{i\in\mathcal{I}_r}\frac{x_i}{d_i}$) from zero to one with a step size of 0.1 and perform the unlearning algorithms to obtain the accuracy of each unlearned model and its corresponding consumed time. Here, we assume that $\gamma = 1$ to model the complete trend when all users are informed.

1) Accuracy Degradation Formulation: We define the *accuracy degradation* as the server's model accuracy reduction after performing machine unlearning. Note that the accuracy is evaluated on a hold-out test set. The accuracy degradation of Retraining from Scratch [8] and SISA [7] against the various unlearning ratios are shown in Figure 1a and Figure 1b, respectively. Each dot represents the accuracy reduction in percentage under an unlearning ratio and one of the datasets. We can observe that the accuracy degradation is *convex increasing* in the unlearning ratio. Such a trend confirms the commonly observed diminishing marginal model accuracy increment when we include more data in model training if we observe the results conversely (i.e., when we decrease the unlearning ratio). Therefore, we adopt an exponential function to model the impact of the users' aggregated data redemption amount $s = \sum_{i\in\mathcal{I}_r} x_i$ on the accuracy degradation as

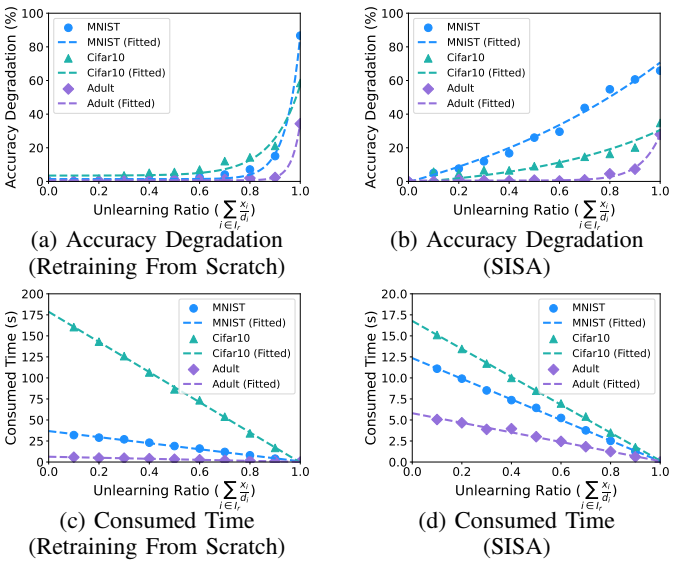
$$A(s) = A_1 e^{A_2 s} - A_3, \tag{3}$$

where $A_1, A_2, A_3 \geq 0$ are the curve-fitting parameters. Then, we use Equation (3) to fit the dots in Figure 1a and Figure 1b, and it shows good approximations of the experimental results.

2) Consumed Time Formulation: We define the *consumed time* as the duration of time required for performing the unlearning algorithm. The consumed time of Retraining from Scratch [8] and SISA [7] against the various unlearning ratios are shown in Figure 1c and Figure 1d, respectively. Similarly, each dot represents the time duration in seconds, and we can observe that the consumed time is *linearly decreasing* in the unlearning ratio. It is because we use the same number of training rounds to build the unlearned model as we trained the original model. When we linearly increase the unlearning ratio, the remaining amount of data is linearly decreasing. Thus, the consumed time of Retraining from Scratch is linearly decreasing. Such reason also applies to SISA because almost all shards and slices need to be retrained even if a very small ratio of data is redeemed (e.g., 1%) in typical SISA settings [7]. We define the consumed time of the server resulting from unlearning $s = \sum_{i\in\mathcal{I}_r} x_i$ amount of data as

$$T(s) = \begin{cases} 0, & \text{if } s = 0, \\ T_0(\sum_{i\in\mathcal{I}} d_i - s), & \text{if } s > 0, \end{cases} \tag{4}$$

where $T_0 > 0$ is the curve-fitting parameter. We use Equation (4) to fit the dots in Figure 1c and Figure 1d, which matches well with the experimental results.[1]



(a) Accuracy Degradation (Retraining From Scratch)
(b) Accuracy Degradation (SISA)
(c) Consumed Time (Retraining From Scratch)
(d) Consumed Time (SISA)

Figure 1: Server's Unlearning Cost.

Next, we define the server's unlearning cost as the weighted summation of the accuracy degradation and consumed time, which can be formulated as

$$C(s) = \alpha A(s) + \beta T(s), \tag{5}$$

where $\alpha, \beta > 0$ are scaling factors between the accuracy degradation and consumed time of the server.

D. Two-Stage Stackelberg Game Formulation

Enforcing the current privacy policies is unfair to the server since it spends extra costs in performing machine unlearning but suffers from model corruption [13]. Therefore, we introduce a pricing mechanism to provide incentives for server's machine unlearning, where the interaction of users and server is formulated as a two-stage Stackelberg game.

1) (Stage I) Server's Profit Maximization: To reimburse for the degradation of model accuracy and the time spent on performing machine unlearning as stated in Section II-C, the server charges each unit of redeemed data a price for the compensation. For fairness with all users, we consider homogeneous pricing, where the server announces only one unit price b for all users. In Stage I, the server optimizes the unit price b to maximize its profit (i.e., received payment minus unlearning cost):

[1]We exclude the dots and fitted curves in Figure 1c and Figure 1d when $s = 0$ for simplicity.

$$\begin{aligned} \max_{b} \quad & \phi(b) = \textstyle\sum_{i\in\mathcal{I}_r} b x_i(b) - C(\sum_{i\in\mathcal{I}_r} x_i(b)) \\ \text{s.t.} \quad & b \geq 0, \end{aligned} \tag{6}$$

where $x_i(b)$ denotes the amount of data to redeem decided by user i based on the unit price b, which is the solution of Problem (7) of Stage II below. The scenario where $b < 0$ is not practical to consider. In that case, the server needs to pay for the users while suffering from the unlearning costs, resulting in a negative profit.

2) (Stage II) Users' Payoff Maximization: After receiving the unit price b from the server, each informed user $i \in \mathcal{I}_r$ decides the data redemption amount independently in Stage II to maximize its payoff (i.e., obtained utility by redeeming data minus the payment to be spent), which can be defined as:

$$\begin{aligned} \max_{x_i} \quad & \delta(x_i) = U_i(x_i) - b x_i \\ \text{s.t.} \quad & 0 \leq x_i \leq d_i, \forall i \in \mathcal{I}_r, \end{aligned} \tag{7}$$

while the other users in $\mathcal{I}_n$ do not redeem their data (i.e., $x_i = 0, \forall i \in \mathcal{I}_n$).

III. Solutions of Two-stage Optimization

In this section, we analyze the two-stage optimization problem using backward induction. Specifically, we first analyze the optimal solution of Stage II in Section III-A. Next, in Section III-B, we substitute the solution of Stage II into Stage I and compute the optimal unit price.

A. Solution of Stage II

We derive the optimal data redemption amount x_i^* for each user in $\mathcal{I}_r$ when the unit price function parameter b is given by the server in Proposition 1.

Proposition 1. *The optimal data redemption amount for informed user $i \in \mathcal{I}_r$ in Stage II under unit price b is*

$$x_i^*(b) = \left[\frac{\lambda_i}{b} - 1\right]_0^{d_i}, \tag{8}$$

where $[z]_p^q = \min\{\max\{p, z\}, q\}$.

The proof of Proposition 1 is given in Online Appendix [29]. Based on Proposition 1, we can divide the informed users $\mathcal{I}_r$ into three non-overlapping subsets as follows.

Corollary 1. *The informed users can be divided into three non-overlapping subsets according to the value of b with $\mathcal{I}_r = \mathcal{I}_1(b) \cup \mathcal{I}_2(b) \cup \mathcal{I}_3(b)$:*

- *$\mathcal{I}_1(b)$ who think the unit price is too high to afford (i.e., $b > \lambda_i$) and their solution is $x_i^* = 0, \forall i \in \mathcal{I}_1(b)$*
- *$\mathcal{I}_2(b)$ who think the unit price is too low (i.e., $b < \frac{\lambda_i}{d_i+1}$) and their solution is $x_i^* = d_i, \forall i \in \mathcal{I}_2(b)$*
- *$\mathcal{I}_3(b)$ who think the unit price is fair (i.e., $\frac{\lambda_i}{d_i+1} \leq b \leq \lambda_i$) and their solution is $x_i^* = \frac{\lambda_i}{b} - 1, \forall i \in \mathcal{I}_3(b)$*

B. Solution of Stage I

Based on Equation (8) and Corollary 1, we can further limit

$$0 \leq b \leq b_{\max}, \tag{9}$$

where $b_{\max} = \max_{\forall i\in\mathcal{I}_r}\{\lambda_i\}$ to reduce the domain since when $b > b_{\max}$, all users are $\mathcal{I}_1(b)$, and no unlearning happens.

To solve Stage I, we substitute the solution of $x_i^*(b)$ in Proposition 1 into the objective function of Stage I (i.e., Problem (6)). Given the complexity of directly solving Problem (6) after the substitution, we characterize the conditions (in Theorems 1 and 2) under which Problem (6) a *convex optimization problem* [19] such that our proposed gradient-based algorithm (i.e., Algorithm 1) can obtain the *optimal* solution of Problem (6). Otherwise, Problem (6) is a *non-convex* problem, and we simplify its first derivative (in Proposition 2) to speed up the computation of our algorithm. We further assume the user's marginal privacy gain λ_i follows a uniform distribution and the amount of contributed data by each user is the same so that the optimal price b^* can be obtained by solving some equations (in Theorem 3).

We present the first condition about the convexity of Problem (6) in the following Theorem 1. The detailed proofs of all the analytical results in this paper can be found in the Online Appendix [29].

Theorem 1. *$\phi(b)$ is a concave function in b if*

$$\alpha A_1 A_2 - \beta T_0 \geq 0. \tag{10}$$

Therefore, if Condition (10) is satisfied, Problem (6) is a *convex optimization problem*, which can be solved optimally by the gradient descent method [19].

On the other hand, when Condition (10) is not satisfied, we can divide Problem (6) into two sub-problems:

$$\max_{0\leq b\leq b_{\max}} \phi(b) = \max\{\max_{0\leq b\leq \tilde{b}} \phi(b), \max_{\tilde{b}\leq b\leq b_{\max}} \phi(b)\}, \tag{11}$$

where the threshold $\tilde{b} \in [0, b_{\max}]$ is given in the following Theorem 2 and its proof can be found in Online Appendix [29].

Theorem 2. *When*

$$\alpha A_1 A_2 - \beta T_0 < 0, \tag{12}$$

there exists a threshold $\tilde{b} \in [0, b_{\max}]$ that satisfies

$$\sum_{i\in\mathcal{I}_2(\tilde{b})} d_i + \frac{\sum_{i\in\mathcal{I}_3(\tilde{b})}\lambda_i}{\tilde{b}} - |\mathcal{I}_3(\tilde{b})| = \frac{1}{A_2\gamma}\ln\frac{\beta T_0}{\alpha A_1 A_2}, \tag{13}$$

such that $\phi(b)$ is a concave function in b for $b \in [0, \tilde{b}]$.

Thus, the first sub-problem in Problem (11) is a *convex optimization problem* with linear constraint (i.e., $0 \leq b \leq \tilde{b}$), and its optimal solution denoted as b_1^* can be obtained by the gradient descent method [19]. However, in general, the second sub-problem is a non-convex optimization problem with linear constraint (i.e., $\tilde{b} \leq b \leq b_{\max}$). To solve the second sub-problem, we analyze its first derivative $\phi'(b)$ that can be presented as

$$\begin{aligned} \phi'(b) \quad = & \frac{\alpha\gamma A_1 A_2 \sum_{i\in\mathcal{I}_3(b)}\lambda_i}{b^2} e^{A_2\gamma(\sum_{i\in\mathcal{I}_2(b)} d_i + \frac{\sum_{i\in\mathcal{I}_3(b)}\lambda_i}{b} - |\mathcal{I}_3(b)|)} \\ & + \gamma\textstyle\sum_{i\in\mathcal{I}_2(b)} d_i - \gamma|\mathcal{I}_3(b)| - \frac{\beta T_0 \gamma \sum_{i\in\mathcal{I}_3(b)}\lambda_i}{b^2}, \end{aligned} \tag{14}$$

and find that it can be simplified as shown in Proposition 2 to speed up the computation of the gradient.

Proposition 2. *When Condition (10) is not satisfied, the first derivative of the objective function of Problem (6) is*

$$\phi'(b) = \gamma\left(\sum_{i\in\mathcal{I}_2(b)} d_i - |\mathcal{I}_3(b)|\right), \text{ for } b \in [\tilde{b}, b_{\max}]. \tag{15}$$

Given the above theorems and proposition, we propose an iterative algorithm for solving Problem (6) in Algorithm 1.

Algorithm 1: Iterative Algorithm for Solving Problem (6)

Input: step size $\Delta b > 0$, tolerance $\epsilon > 0$, system parameters $\alpha, \beta, A_1, A_2, A_3, T_0$
Output: Optimized unit price b^*

1 **if** *Condition (10) is satisfied* **then**
2 **return** b^* = GD $(0, b_{\max}, \phi^{'}(b))$ using Equation (14)
3 **else**
4 Compute $\tilde{b}$ using Equation (13)
5 Obtain b_1^* = GD $(0, \tilde{b}, \phi^{'}(b))$ using Equation (14)
6 Obtain b_2^* = GD $(\tilde{b}, b_{\max}, \phi^{'}(b))$ using Equation (15)
7 **return** $b^* = \arg\max_{b \in \{b_1^*, b_2^*\}} \phi(b)$
8 **end**
9 **Function** GD $(\check{b}, \hat{b}, g(b))$ **:**
10 **while** *abs*$(g(b_k)) > \epsilon$ **do**
11 Decide $\mathcal{I}_1(b_k), \mathcal{I}_2(b_k), \mathcal{I}_3(b_k)$ by Corollary 1
12 $b_{k+1} = [b_k + g(b_k)\Delta b]_{\check{b}}^{\hat{b}}$
13 **end**
14 **return** b_k

In Algorithm 1, we define a function (lines 9 to 14) to perform the gradient descent method [19] in $[\check{b}, \hat{b}]$. The function has three inputs varying to the conditions: $\check{b}$ and $\hat{b}$ are the lower and upper bound of the domain, while $g(b)$ is the gradient of $\phi(b)$. In the function, the unit price b is iteratively updated. In the body of Algorithm 1, when Condition (10) is satisfied (lines 1 to 2), the domain is $b \in [0, b_{\max}]$ and the solution can be optimally obtained by gradient descent method [19]. When Condition (10) is not satisfied, we compute the threshold (line 4), obtain the solutions for two sub-problems (lines 5 to 6), and choose one of them as the final solution (line 7).

Furthermore, we analyze a practical case that assumes the users' marginal privacy gain λ_i follows a uniform distribution in $[\lambda_{\min}, \lambda_{\max}]$ and the amount of contributed data by each user d_i is the same. In this special case, we can prove that even the second sub-problem in Equation (11) is a *convex optimization problem* such that we can obtain b^* directly by solving some equations in Theorem 3 instead of running Algorithm 1.

Theorem 3. *Suppose* $\{\lambda_i\}_{\forall i \in \mathcal{I}_r}$ *follows a uniform distribution* $\mathbb{U}(\lambda_{\min}, \lambda_{\max})$ *and* $d_i = d, \forall i \in \mathcal{I}_r$. *When Condition* (10) *is satisfied, the solution for Problem (6) is* $b^* = [b_1]_0^{b_{\max}}$, *where* b_1 *can be numerically obtained by solving*

$$e^{M_2 b_1} = \frac{M_3}{M_1} + \frac{M_4}{M_1} b_1, \tag{16}$$

with $M_1 = e^{A_2 \gamma d \left(\frac{\lambda_{\max}|\mathcal{I}|}{\lambda_{\max} - \lambda_{\min}} + \frac{1}{2} d(d+2) \right)}$, $M_2 = -\frac{(d+2)|\mathcal{I}|}{\lambda_{\max} - \lambda_{\min}}$, $M_3 = \frac{\beta T_0}{\alpha A_1 A_2} - \frac{2|\mathcal{I}|\lambda_{\max}}{\alpha A_1 A_2 (\lambda_{\max} - \lambda_{\min})(d+2)}$ *and* $M_4 = \frac{1}{\alpha A_1 A_2 (\lambda_{\max} - \lambda_{\min})}$. *Otherwise, when Condition* (10) *is not satisfied, both sub-problems in Equation* (11) *are convex optimization problems with a closed-form threshold*

$$\tilde{b} = \frac{\frac{1}{A_2 \gamma} \ln \frac{\beta T_0}{\alpha A_1 A_2} - \frac{\lambda_{\max}|\mathcal{I}| d}{\lambda_{\max} - \lambda_{\min}}}{(\frac{1}{2} - \frac{|\mathcal{I}|}{\lambda_{\max} - \lambda_{\min}}) d(d+2)}, \tag{17}$$

such that the optimal unit price can be obtained by

$$b^* = \begin{cases} [b_1]_0^{\tilde{b}}, & \text{if } \phi([b_1]_0^{\tilde{b}}) \geq \phi([\frac{\lambda_{\max}}{d+2}]_{\tilde{b}}^{b_{\max}}), \\ [\frac{\lambda_{\max}}{d+2}]_{\tilde{b}}^{b_{\max}}, & \text{otherwise}. \end{cases} \tag{18}$$

IV. Performance Evaluations

In this section, we conduct experiments to prove the effectiveness of our mechanism compared with two baselines. Since the results on each dataset and unlearning algorithm are similar, we use the results on MNIST [16] dataset and SISA [7] algorithm to demonstrate our findings.

We consider $|\mathcal{I}| = 6000$ users and assume the MNIST dataset with $|\mathcal{D}| = 60000$ is uniformly and randomly contributed by each user such that $d_i = 10, \forall i \in \mathcal{I}$. We set the marginal privacy gain of each user $\{\lambda_i\}_{\forall i \in \mathcal{I}}$ follows a uniform distribution in the internal $(0.5, 30)$. The server uses the fitting parameters obtained from the above-mentioned experiments in Section II-C to optimize the unit price b. The scaling factors for accuracy degradation and consumed time are set to $\alpha = 3000$ and $\beta = 1$ to align with the realistic condition [30], [31]. Such settings satisfy Condition (10). We vary the informed ratio γ from zero to one with step size 0.1. For each γ, we randomly select γ of all users to decide their data redemption amount rationally given the unit price b, perform SISA according to the total redemption amount, and record the accuracy degradation and consumed time. We set up two baselines to compare with:

- **GDPR** as a representative of current data protection policies, where the informed users redeem all of their data (i.e., $x_i = d_i, \forall i \in \mathcal{I}_r$). It corresponds to Stage II when the server charges nothing from users (i.e., $b = 0$)
- **No Redemption** as the case without any data protection policy, where none of the users redeem their data (i.e., $x_i = 0, \forall i \in \mathcal{I}$). It corresponds to Stage II when $b = \infty$.

Firstly, we compare the server profit $\phi(b)$ of each mechanism under various γ, and the results are shown in Figure 2. The results show that only our mechanism can provide profits for the server because of receiving sufficient compensation from the users. Nevertheless, the No Redemption baseline cannot obtain any profit since no machine unlearning happens. The results also show that if we enforce the current data protection policies, such as GDPR, the server suffers from huge losses. Therefore, companies are not willing to comply with the policies in reality. Moreover, we observe that when the fraction of informed users γ increases, server's profit under our mechanism increases. This is because the server can raise the unit price to obtain a larger profit when more users are informed, as stated in Theorem 3. Thus, the policymaker should inform more people to participate in our mechanism.

Secondly, we compare the social welfare of each mechanism, which is the sum of server profit and users' payoffs (i.e., $\phi(b) + \sum_{i \in \mathcal{I}} \delta(x_i)$). The results under various γ are shown in Figure 3. We find that GDPR can not maintain sufficient social welfare when a large number of users is informed (e.g., $\gamma > 0.3$) since those informed users redeem all of their data, resulting in a catastrophic unlearning [13]. Also,

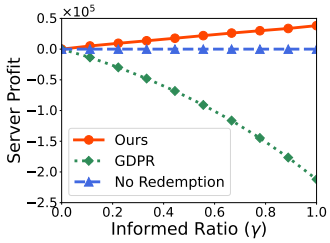


Figure 2: Informed Ratio versus Server Profit.

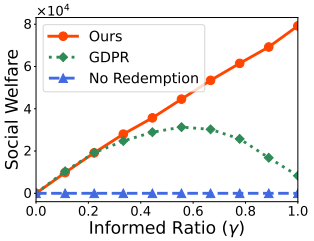


Figure 3: Informed Ratio versus Social Welfare.

the No redemption baseline results in zero social welfare since no data redemption happens. In contrast, our mechanism can provide the highest social welfare compared with other baselines, especially when γ is large. Combining Figure 2 and 3, we can find that our mechanism can improve the server's profit while not hampering social welfare, which reaches a balance between the server and users.

V. Conclusion

To the best of our knowledge, we proposed the first incentive mechanism for machine unlearning in this paper. We conducted experiments on three real-world datasets with two popular machine unlearning algorithms and characterized two major costs of the server: accuracy degradation and consumed time. We modeled the interaction between the server and users as a two-stage Stackelberg game, where the server optimizes the unit price for compensation and users optimize their data redemption amounts. To solve the problem, we first derived the optimal data redemption amount for users given the received unit price from the server. We proposed an iterative algorithm to optimize the unit price for the server. Our simulation results based on real data demonstrated that our proposed mechanism achieves the largest server's profit and social welfare compared with two realistic baselines. For future work, we plan to consider a more general scenario in which the users not only contribute and redeem their data but also use the trained model.

References

[1] H. Liu, W. Wang, and H. Li, "Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4995–5006.

[2] Y. Cui, Z. Li, L. Liu, J. Zhang, and J. Liu, "Privacy-preserving speech-based depression diagnosis via federated learning," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 1371–1374.

[3] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[4] M. X. Chen, B. N. Lee, G. Bansal, S. Cao, J. Lu, J. Tsay, Y. Wang, A. M. Dai, Z. Chen *et al.*, "Gmail smart compose: Real-time assisted writing," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[5] "General Data Protection Regulation." https://gdpr-info.eu/, 2016.

[6] "California Consumer Privacy Act." https://oag.ca.gov/privacy/ccpa, 2018.

[7] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021.

[8] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 896–911.

[9] "GDPR cost businesses 8 percent of their profits, according to a new estimate." https://techmonitor.ai/policy/privacy-and-data-protection/gdpr-cost-businesses-8-of-their-profits-according-to-a-new-estimate, 2022.

[10] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," *arXiv preprint arXiv:2209.02299*, 2022.

[11] V. Codreanu, D. Podareanu, and V. Saletore, "Scale out for large minibatch sgd: Residual network training on imagenet-1k with improved accuracy and reduced time to train," *arXiv preprint arXiv:1711.04291*.

[12] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites, "Adaptive machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 319–16 330, 2021.

[13] Q. P. Nguyen, R. Oikawa, D. M. Divakaran, M. C. Chan, and B. K. H. Low, "Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten," in *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, 2022, pp. 351–363.

[14] A. H. Chuan Guo, Tom Goldstein and L. van der Maaten, "Certified data removal from machine learning models," *arXiv preprint arXiv:1911.03030*, 2020.

[15] "What is GDPR and how does it impact your business?" https://www.superoffice.com/blog/gdpr/, 2023.

[16] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research," *IEEE Signal Processing Magazine*, vol. 29, 2012.

[17] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[18] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[19] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[20] N. Ding, Z. Sun, E. Wei, and R. Berry, "Incentive mechanism design for federated learning and unlearning," in *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2023.

[21] A. F. Daughety and J. F. Reinganum, "Public goods, social pressure, and the choice between privacy and publicity," *American Economic Journal: Microeconomics*, vol. 2, no. 2, pp. 191–221, 2010.

[22] Y. Jeong and M. Maruyama, "Commitment to a strategy of uniform pricing in a two-period duopoly with switching costs," *Journal of Economics*, vol. 98, pp. 45–66, 2009.

[23] H. R. Varian, "Economic aspects of personal privacy," *Internet Policy and Economics: Challenges and Perspectives*, pp. 101–109, 2009.

[24] A. Acquisti, C. Taylor, and L. Wagman, "The economics of privacy," *Journal of economic Literature*, vol. 54, no. 2, pp. 442–492, 2016.

[25] "How concerned are europeans about their personal data online?" http://fra.europa.eu/en/news/2020/how-concerned-are-europeans-about-their-personal-data-online, 2020.

[26] Z. Wu, Y. Shu, and B. K. H. Low, "DAVINZ: Data valuation using deep neural networks at initialization," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022.

[27] D. Popescu and A. Chronopoulos, "Power control and utility optimization in wireless communication systems," in *2005 IEEE 61st Vehicular Technology Conference*, vol. 1, 2005, pp. 314–318 Vol. 1.

[28] M. H. Cheung, H. Mohsenian-Rad, V. W. Wong, and R. Schober, "Utility-optimal random access for wireless multimedia networks," *IEEE Wireless Communications Letters*, vol. 1, no. 4, pp. 340–343, 2012.

[29] "Online Appendix." https://github.com/ycui26/Incentive-Mechanism-for-Machine-Unlearning, 2023.

[30] S. S. Kadam, A. C. Adamuthe, and A. B. Patil, "CNN model for image classification on MNIST and fashion-MNIST dataset," *Journal of scientific research*, vol. 64, no. 2, pp. 374–384, 2020.

[31] A. Palvanov and Y. Im Cho, "Comparisons of deep learning algorithms for MNIST in real-time environment," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 18, no. 2, pp. 126–134, 2018.