

# Verifiable and Provably Secure Machine Unlearning

Thorsten Eisenhofer<sup>+†</sup>, Doreen Riepel<sup>‡</sup>, Varun Chandrasekaran<sup>\*†</sup>,  
Esha Ghosh<sup>\*</sup>, Olga Ohrimenko<sup>§</sup>, Nicolas Papernot<sup>¶||</sup>

Ruhr University Bochum<sup>‡</sup>, Microsoft Research<sup>\*</sup>, The University of Melbourne<sup>§</sup>,  
University of Illinois Urbana-Champaign<sup>†</sup>, University of Toronto<sup>¶</sup>, Vector Institute<sup>||</sup>

**Abstract**—Machine unlearning aims to remove points from the training dataset of a machine learning model after training; for example when a user requests their data to be deleted. While many machine unlearning methods have been proposed, none of them enable users to audit the procedure. Furthermore, recent work shows a user is unable to verify if their data was unlearned from an inspection of the model alone. Rather than reasoning about model parameters, we propose to view verifiable unlearning as a security problem. To this end, we present the first cryptographic definition of verifiable unlearning to formally capture the guarantees of a machine unlearning system. In this framework, the server first computes a proof that the model was trained on a dataset  $D$ . Given a user data point  $d$  requested to be deleted, the server updates the model using an unlearning algorithm. It then provides a proof of the correct execution of unlearning and that  $d \notin D'$ , where  $D'$  is the new training dataset. Our framework is generally applicable to different unlearning techniques that we abstract as *admissible functions*. We instantiate the framework, based on cryptographic assumptions, using SNARKs and hash chains. Finally, we implement the protocol for three different unlearning techniques (retraining-based, amnesiac, and optimization-based) to validate its feasibility for linear regression, logistic regression, and neural networks.

## I. INTRODUCTION

The right to be forgotten entitles individuals to self-determine the possession of their private data and compel its deletion. In practice, this is now mandated by recent regulations like the GDPR [2], CCPA [3], or PIPEDA [4]. Consider the case where a company or service provider collects data from its users. These regulations allow users to request a deletion of their data, and legally compels the company to fulfil the request. However, this can be challenging when the data is used for downstream analyses, *e.g.*, training machine learning (ML) models, where the relationship between model parameters and the data used to obtain them is complex [44]. In particular, ML models are known to memorize information from their training set [22], [17], resulting in a myriad of attacks against the privacy of training data [19], [47].

Thus, techniques have been introduced for *unlearning*: a trained model is updated to remove the influence a training point had on the model’s parameters and predictions [18]. Yet, regardless of the particular approach, existing techniques [75], [15], [37], [30], [33], [7], [58] suffer from one critical limitation: they are unable to provide the user with a proof that their data was indeed unlearned. Put another way, the user is asked to

blindly trust that the server executed the unlearning algorithm to remove their data with no ability to verify this. This is problematic because dishonest service providers may falsify unlearning to avoid paying the large computational costs or to maintain model utility [59], [29].

Additionally, verifying that a point is unlearned is non-trivial *from the user’s perspective*. A primary reason is that users (or third-party auditors) cannot determine whether a data point is unlearned (or not) by comparing the model’s predictions or parameters before and after claimed unlearning. The complex relationship between training data, models’ parameters, and their predictions make it difficult to isolate the effects of any training point. In fact, prior work [65], [68] demonstrates that a model’s parameters can be identical when trained with or without a data point.

To address these concerns, we propose a *cryptographic approach to verify unlearning*. Rather than trying to verify unlearning by examining changes in the model, we ask the service provider (*i.e.*, the server) to present a cryptographic proof that an *agreed-upon* unlearning process was executed. This leads us to view unlearning as a security problem that we aim to solve with formal guarantees.

In this paper we propose *the first formal security definition* of verifiable machine unlearning. Our framework describes an iteration-based protocol and requires the server to prove that it has honestly updated the model and dataset in each iteration, either due to training with new data or unlearning previously used data. Only then does the user have sufficient guarantees about deletion of their data. Under this definition, we can instantiate protocols using any unlearning technique and any cryptographic primitives that have appropriate security guarantees.

We identified several challenges while developing the framework that we believe are inherent to unlearning.

- 1) Verifying unlearning cannot be solved by naive one-shot verifiable computation as it requires a user to be able to verify that their data was not re-added at later stages. Hence, the definition has to capture all model updates due to new points added or points being deleted.
- 2) The relationship between an updated model and the evolving dataset needs to be formally captured for verification. For example, a naive way would be to define this relationship as a re-training function, *i.e.*, the updated model is the result of training on the evolved dataset. This can be viewed as “exact unlearning”. However,

<sup>+</sup>Work done while the author was interning at Vector Institute.

since other (approximate) unlearning techniques exist, we define this relationship as a set of functions that we call *admissible functions*. This abstraction captures the relationship between models and datasets via initialization, training and unlearning functions.

- 3) As we observe above, the security definition needs to capture consistency of data during training and unlearning, *and* across model updates and evolving datasets. Though this can be done by passing the whole dataset between the verification steps (training and unlearning) and sending data to the user, we aim to verify consistency in a succinct manner. To this end, we define a strong notion of extractor-based security, capturing that the server must know some underlying dataset in order to compute a valid proof.

Our framework is general and we later demonstrate its applicability to three different unlearning techniques. Notably, none of these have been proved using verifiable computation before. We focus our discussion below on re-training based unlearning, one of the unlearning techniques, and give an overview of our framework. In this approach, the server retrains a model without using a data point  $d$  that needs to be unlearned.

In our framework, we identify the following guarantees that need to be satisfied: (a) the model was trained from some dataset and (b) the user’s data point is not present in this dataset. Thus, the framework has two major components. First, the server computes a *proof of training* whenever data points are added to the model’s training data: this establishes that it trained the model on a particular dataset. Second, when a user submits a request to unlearn a specific data point, the server computes a *proof of unlearning*. It proves that the model was updated addressing the request, and additionally provides the user with a *proof* that their data point was indeed unlearned and not part of the training set. These proofs should also ensure that no data point can be added back to the training set *after* it was unlearned.

We present a fully instantiated protocol in our framework. This instantiation uses SNARK-based verifiable computation for the proof of model updates induced by training or unlearning and hash chains for the proof of non-membership in a training set since the time unlearning request was made.

Finally, we provide the first implementation of verifiable unlearning based on cryptographic primitives. In particular, we use Spartan [60] for verifiable training and unlearning. Spartan is a transparent SNARK, *i.e.*, it does not require a trusted setup, which is desirable for our purpose. We consider three unlearning techniques: retraining-based unlearning, amnesiac unlearning [33] and optimization-based unlearning [40], [72]. We demonstrate the versatility and scalability of our construction on a variety of binary classification tasks from the PMLB benchmark suite [56], using linear regression, logistic regression, and simple neural networks.

**Contributions.** We make the following contributions:

- *Formal framework.* We introduce a general framework to construct protocols for verifiable machine unlearning. Our framework is designed to be general enough to capture different unlearning algorithms and secure primitives.
- *Security definition of verifiable machine unlearning.* We then propose a formal security definition of a verifiable machine unlearning scheme. This game-based definition allows one to prove security of their instantiation of the unlearning protocol. Our framework models verifiable unlearning as a 2-party protocol (executed between the server and the users).
- *Instantiation.* We present a fully instantiated protocol in our framework. This construction is based on a generic interface for training and unlearning and thus applicable to any training and unlearning algorithm (as captured by our admissible functions abstraction).
- *Practical implementation.* We implement the protocol’s main functionality, study its applicability to three unlearning techniques, different classes of ML models and benchmark datasets. We observe that compared to training verification, verifying unlearning adds a small cost and efficiency of unlearning depends on the unlearning technique.

## II. BACKGROUND

In this section, we discuss the preliminaries needed to understand the contributions of our work.

**Notation.** Throughout the paper, let  $\lambda$  denote the security parameter. We call a function negligible in  $\lambda$ —denoted by  $\text{negl}(\lambda)$ —if it is smaller than the inverse of any polynomial for all large enough values of  $\lambda$ .  $[m : n]$  denotes the set  $\{m, m + 1, \dots, n\}$  for integers  $m < n$ . For  $m = 1$ , we simply write  $[n]$ .  $y \leftarrow M(x_1, x_2, \dots)$  denotes that on input  $x_1, x_2, \dots$ , the probabilistic algorithm  $M$  returns  $y$ . An adversary  $\mathcal{A}$  is a probabilistic algorithm, and is *efficient* or Probabilistic Polynomial-Time (PPT) if its run-time is bounded by some polynomial in the length of its input. We will use code-based games, where  $\text{Pr}[G \Rightarrow 1]$  denotes the probability that the final output of game  $G$  is 1.

### A. Machine Learning Preliminaries

**Supervised Machine Learning.** Supervised machine learning (ML) is the process of learning a parameterized function  $f_\theta$  (often called a *model*) that is able to predict an output (from the space of outputs  $\mathcal{Y}$ ) given an input (from the space of inputs  $\mathcal{X}$ ), *i.e.*,  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ . Commonly learnt functions include linear regression, logistic regression, and feed-forward neural networks.

The parameters of this function are typically optimized using methods such as stochastic gradient descent (SGD). Let  $\theta_{\text{initial}}$  be randomly initialized parameters and  $D = \{d_1, \dots, d_n\}$  a set of training data points, where each  $d = (d_x, d_y) \in \mathcal{X} \times \mathcal{Y}$ . During training we iteratively update

parameters as  $\theta' := \theta - \eta \nabla_{\theta} \mathcal{L}(f_{\theta}(d_x), d_y)$  for points  $d \in D$  where  $\mathcal{L}$  is a suitably chosen loss function (e.g., cross-entropy loss) and  $\eta$  the *learning rate*.

In SGD, the update is calculated for a randomly chosen input  $d \in D$  in each step. In practice, this is often extended to *batches* of data points in order to reduce the variance of each update. Often, multiple passes (called *epochs*) are repeated through the dataset. We can describe the full process of *training* a model  $m$  (interchangeably used with  $\theta_m$ ) by

$$\theta_m := \theta_{initial} + \sum_{e \in [E]} \sum_{d \in D} \Delta_{e,d}$$

with  $E$  being the number of epochs and  $\Delta_{e,d}$  the update on the model's parameter from data point  $d$  in epoch  $e$ .

**Machine Unlearning.** In machine unlearning the goal is to design algorithms that enable an ML model (specifically, its parameters) to forget the contribution of a (subset of) data point(s). The canonical approach for this is to naively retrain the model from scratch. Hence, removing a data point  $d^*$  from a model  $m$  with *retraining-based* unlearning can be described as

$$\theta_{m'} := \theta_{initial} + \sum_{e \in [E]} \sum_{d \in D \setminus \{d^*\}} \Delta_{e,d}$$

As the resulting model  $\theta_{m'}$  is completely devoid of data point  $d^*$  (by construction), this is an example for *exact unlearning* [15], [18], [76], [50], which is desirable but often prohibitively expensive.

More practical unlearning techniques where the contribution of a data point cannot be *completely* removed and the guarantees tolerate some error [37], [20], [67], [7], [58], [30], [33] are commonly referred to as *approximate unlearning*. An example for this is *amnesiac unlearning* [33]. Given a model  $m$ , removing a data point  $d^*$  with amnesiac unlearning means that we compute

$$\theta_{m'} := \theta_m - \sum_{e \in [E]} \Delta_{e,d^*}$$

In other words, to unlearn data point  $d^*$ , we remove the updates to the model's parameters that were *directly* computed on that data point for all training epochs  $E$ . Yet, amnesiac unlearning only provides approximate guarantees since updates from unlearned data points indirectly also influence updates from later points during the iterative nature of the training [67].

Other approaches for approximate unlearning formulate unlearning as an optimization problem [40], [72] (similar to training). In every step, instead of reducing the loss of a data point, we increase it. We refer to this as *optimization-based* unlearning. Formally, we iteratively compute an update  $\Delta_{e,d^*}$  for the current model that is *subtracted* from its parameters:

$$\theta_{m'} := \theta_m - \sum_{e \in [\hat{E}]} \Delta_{e,d^*}$$

where  $\hat{E}$  denotes the number of *unlearning epochs* and  $\Delta_{e,d^*}$  the update from data point  $d^*$  in epoch  $e$ . For model parameters  $\theta$  (in epoch  $e$ ), we define  $\Delta_{e,d^*} := -\hat{\eta} \nabla_{\theta} \mathcal{L}(f_{\theta}(d_x^*), d_y^*)$  with *unlearning rate*  $\hat{\eta}$  and loss function  $\mathcal{L}$ .

## B. Cryptographic Preliminaries

**Collision-Resistant Hash Functions.** A function  $\text{Hash} : \{0, 1\}^n \rightarrow \{0, 1\}^{\kappa}$  is collision-resistant if

- It is length-compressing, i.e.,  $\kappa < n$ .
- It is hard to find collisions, i.e., for all PPT adversaries  $\mathcal{A}$ , and for all security parameters  $\lambda$ ,

$$\Pr \left[ \begin{array}{l} (x_0, x_1) \leftarrow \mathcal{A}(1^{\lambda}, \text{Hash}) : \\ x_0 \neq x_1 \wedge \text{Hash}(x_0) = \text{Hash}(x_1) \end{array} \right] \leq \text{negl}(\lambda).$$

**Proof Systems.** An interactive proof system describes a protocol between a *prover* and a *verifier*, where the prover wants to convince the verifier that some statement  $\phi$  for a given polynomial time decidable relation  $R$  is true. The prover holds a witness  $\omega$  for the statement. We can then express  $R$  with a circuit that takes public and private inputs (statement and witness) and returns **true** if the input is in the relation. In this work we are concerned with *non-interactive* proof systems. A Succinct Non-Interactive Argument of Knowledge (SNARK) allows the prover to non-interactively prove the statement with a short (or succinct) cryptographic proof which can be verified in time sublinear in the size of the statement. We denote a SNARK by  $\Pi$  and define it by the three algorithms ( $\Pi$ .Setup,  $\Pi$ .Prove,  $\Pi$ .Vrfy). More formally,

- $\text{pp} \leftarrow \Pi$ .Setup( $1^{\lambda}, R$ ): The setup algorithm outputs public parameters  $\text{pp}$  for a polynomial-time decidable relation  $R$ .
- $\pi \leftarrow \Pi$ .Prove( $R, \text{pp}, \phi, \omega$ ): The prover algorithm takes as input the  $\text{pp}$  and  $(\phi, \omega) \in R$  and returns an argument  $\pi$ , where  $\phi$  is termed the statement and  $\omega$  the witness.
- $b \leftarrow \Pi$ .Vrfy( $R, \text{pp}, \phi, \pi$ ): The verification algorithm takes the  $\text{pp}$ , a statement  $\phi$  and an argument  $\pi$  and returns a bit  $b$ , where  $b = 1$  indicates success and  $b = 0$  indicates failure.

**Perfect Completeness.** Given any true statement, an honest prover should be able to convince an honest verifier. More formally, let  $\mathcal{R}$  be a sequence of families of efficiently decidable relations  $R$ . For all  $R \in \mathcal{R}$  and  $(\phi, \omega) \in R$

$$\Pr \left[ \Pi$$
.Vrfy( $R, \text{pp}, \phi, \pi$ )  $\left| \begin{array}{l} \text{pp} \leftarrow \Pi$ .Setup( $1^{\lambda}, R$ ); \\ \pi \leftarrow \Pi.Prove( $R, \text{pp}, \phi, \omega$ ) \end{array} \right. \right] = 1.

**Computational Soundness.** We say that  $\Pi$  is sound if it is not possible to prove a false statement. Let  $L_R$  be the language consisting of statements for which there exists corresponding witnesses in  $R$ . For a relation  $R \sim \mathcal{R}$ , we require that for all non-uniform PPT adversaries  $\mathcal{A}$

$$\Pr \left[ \begin{array}{l} \phi \notin L_R \text{ and} \\ \Pi$$
.Vrfy( $R, \text{pp}, \phi, \pi$ ) \end{array} \left| \begin{array}{l} \text{pp} \leftarrow \Pi.Setup( $1^{\lambda}, R$ ); \\ (\phi, \pi) \leftarrow \mathcal{A}(R, \text{pp}) \end{array} \right. \right] \leq \text{negl}(\lambda).

We further define the notion of witness extractability or knowledge soundness.

**Computational Knowledge Soundness.**  $\Pi$  satisfies computational knowledge soundness if there exists an extractor that can compute a witness whenever the adversary produces a valid argument. More formally, for a relation  $R \sim \mathcal{R}$ , we require

that for all non-uniform PPT adversaries  $\mathcal{A}$  there exists a non-uniform PPT extractor  $\mathcal{E}$  such that

$$\Pr \left[ \begin{array}{l} (\phi, \omega) \notin R \text{ and } \\ \Pi.\text{Vrfy}(R, \text{pp}, \phi, \pi) \end{array} \middle| \begin{array}{l} \text{pp} \leftarrow \Pi.\text{Setup}(1^\lambda, R); \\ ((\phi, \pi); \omega) \leftarrow (\mathcal{A} \parallel \mathcal{E})(R, \text{pp}) \end{array} \right] \leq \text{negl}(\lambda).$$

We say a SNARK  $\Pi$  is secure if it satisfies perfect completeness, computational soundness and knowledge soundness.

### III. VERIFIABLE MACHINE UNLEARNING

We consider the following ecosystem: there are many *users*, each of whom has access to a set of data points (or dataset). They share their data with a *server* which uses it to learn an ML model. Users can send requests to either delete or add new data. We focus on ensuring that users can verify that their deletion requests are met.

**Threat Model.** We assume the server is malicious, *i.e.*, it may not execute unlearning. Reasons for this include the server being unwilling to tolerate a degradation in the model’s performance after data deletion [59], [29], or pay the computational penalty associated with updating the model [15], [33]. To this end, our focus is to develop a method to verify that the server adheres to users’ requests.

#### Scope and Assumptions.

- 1) *Out-of-band authentication.* We assume that the users have some (out-of-band) mechanism to authenticate to the server. A server will honor unlearning requests from legitimate users only. Since this can be done using standard authentication mechanisms, *e.g.*, digital signatures, we do not model this explicitly. This would prevent a malicious user deleting a point belonging to a different user.
- 2) *Uniqueness of data points.* We need a mechanism to attribute data points to users: when a user requests their data point to be deleted, it should be clearly identifiable. In practice, multiple users could have the same data point making identification challenging. To resolve this problem, we assume the server prepends a unique identifier to each data point. We refer to such a unique representation as a *data record*.
- 3) *Sybil data-records.* A malicious server can re-add a data point that was deleted (*e.g.*, by creating a fake user or colluding with an existing user). This would result in a different data record as it would have a different identifier. We note that, in principle, detecting this behaviour is possible, for example, by recording all added data points in a transparency log (similar to Certificate Transparency [1]) and then comparing the new points with previous points. However, differentiating this behaviour from benign behaviour is difficult: there could be an honest user with the same (or similar) data point. This is an interesting problem which we consider out-of-scope for this work. We note, however, that any solutions developed for it could be integrated into our framework.

**Necessity of Auditable Algorithmic Definitions.** In the status quo, there is no rigorous way for the user to verify if the model being used is devoid of their data. One naive solution would be to provide the user with the trained parameters (including random seeds) of the model and all the data, and request them to locally re-run the training and compare their model parameters to the server’s. Another method could be based on influence techniques [44] to understand if their data contributes to these parameters. Both of these naive solutions suffer from a fundamental problem: it is possible to arrive to same model parameters even if data was deleted. For example, recent work by Shumailov *et al.* [65] and Thudi *et al.* [68] describe how a user’s contribution (towards model parameters) can be approximated from other entries in a dataset, rendering such approaches insufficient: the server can claim to have obtained the exact same model parameters from a number of different datasets. Therefore, our approach relies on *proving the execution of an unlearning algorithm*. To prove that a particular record was unlearned correctly, we further need to trust that the model upon which unlearning is performed was derived from that particular record. Thus, we also require to prove that the training procedure was executed correctly. To capture this formally, we propose a generic interface for *admissible functions* to describe training and unlearning procedures (cf. Section IV).

**Desiderata.** From the discussion thus far, we unearth two main requirements to achieve our goals.

- D1.** Given an ML model that was trained on a dataset  $D$  and data points are added, we require a *proof of training* to establish that the *updated* model was obtained from the updated dataset by executing a training function all participating parties agreed on.
- D2.** In a similar fashion, we want to prove the removal of data points. Thus, given an ML model that was trained on a dataset  $D$ , machine unlearning mechanisms update this model by (conceptually) *removing* data points from the set  $D$ . Then, we require a *proof of unlearning* to establish that the *updated* model was obtained from the updated dataset using an agreed-upon unlearning function. Further, for each removed data point  $d$ , we require a proof that  $d$  is not part of the (updated) training data  $D'$  (*i.e.*,  $d \notin D'$ ), which we denote by a *proof of non-membership*.

To guarantee the absence of deleted data points even after future model updates, it is also necessary to take into account *all* updates to the model, and not only those referring to unlearning (*i.e.*, to prevent a server from re-adding an unlearned data record).

Additionally, it needs to be ensured that the server uses the most recent model for inference. However, we consider proving and verifying inference an interesting problem which is related to, but also independent of unlearning. We note that there exist several works addressing this problem which we discuss in Section VIII.

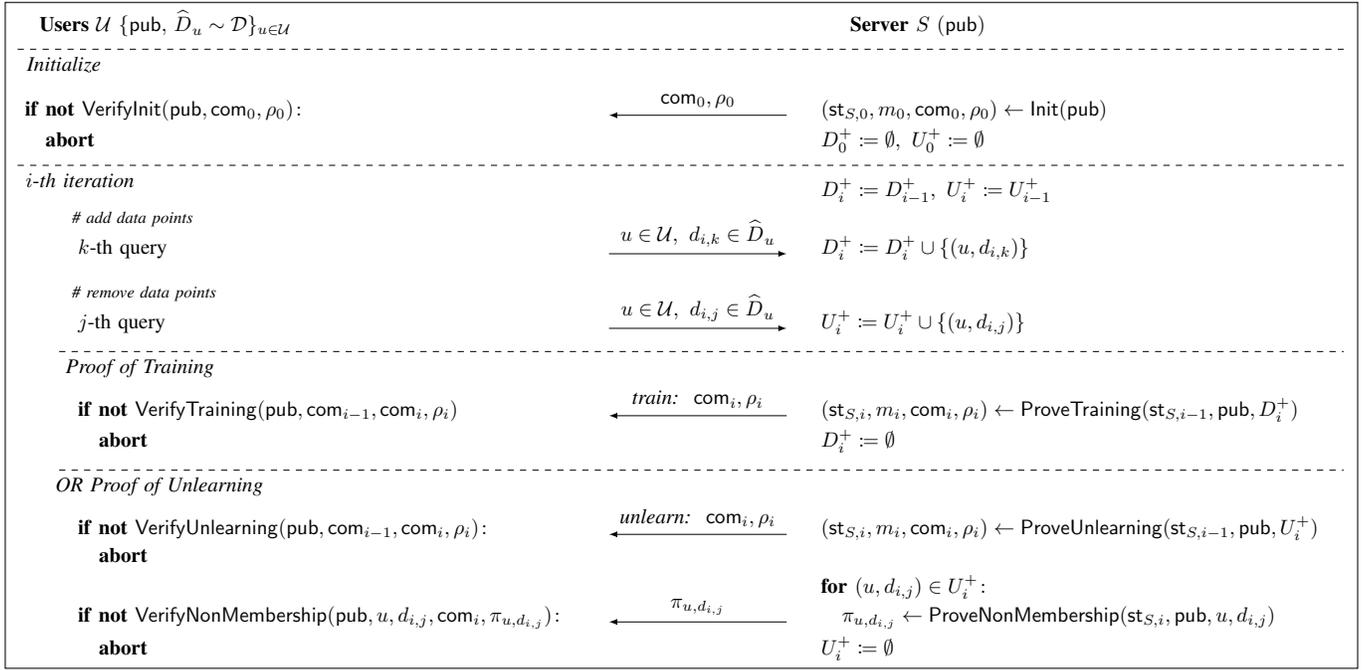


Fig. 1: *Unlearning Framework*. We describe protocols in this framework based on an admissible functions  $f$ . After initialization, execution proceeds in iterations. In the beginning of each iteration  $i$ , users  $\mathcal{U}$  can issue requests for data to be added or deleted. After this phase, the server  $S$  either performs a *proof of training* by adding the requested data records in  $D_i^+$  to the model or a *proof of unlearning* by removing the requested data records in  $U_i^+$ . It computes a commitment  $\text{com}_i$  on the updated model  $m_i$  and updated training dataset. Furthermore, the server computes a proof  $\rho_i$  that  $m_i$  was obtained from this dataset. The users verify this proof and the commitment. In each iteration of unlearning the server additionally creates a *proof of non-membership* for every unlearned data point conforming to a user that it has complied with a data deletion request. This proof can be verified by the user against  $\text{com}_i$ .

#### IV. OUR FRAMEWORK

We now present our formal framework for verifiable machine unlearning. This framework defines a generic interface for verifiable machine unlearning capturing desiderata **D1** and **D2** (cf. Section III) to verify training and unlearning. It allows various instantiations, e.g., using different unlearning techniques or cryptographic primitives. We consider protocols for verifiable machine unlearning that are executed interactively by the two roles introduced in the previous section: a set of users  $\mathcal{U}$  and a server  $S$ .

**Dataset.** Let  $\mathcal{D}$  be the distribution of data points. Each user  $u \in \mathcal{U}$  possesses a set of data points  $\widehat{D}_u \sim \mathcal{D}$ . At the server side, there is an initially empty dataset  $D_0 = \emptyset$  (which is later populated for training an ML model). During the execution of the protocol, users can request to add or delete their data points. Different versions of the server’s dataset (as it develops during the execution of the protocol) are denoted by their corresponding index (i.e.,  $D_0, D_1, \dots$ ). Recall that we consider data records such that each data point is distinctly identifiable and unique (refer Section III). Therefore, entries in  $D_i$  are tuples of the form  $(u, d) \in \mathcal{U} \times \widehat{D}_u$ .

**Admissible Functions.** We consider triples of functions  $(f_I, f_T, f_U)$ , where  $f_I$  is an initialization function,  $f_T$  is a

training function and  $f_U$  is an associated unlearning function. The set of all admissible functions is denoted by  $\mathcal{F}$ . We let  $\text{pp}_f$  denote public hyperparameter which are used for initialization. W.l.o.g., we assume the functions to be deterministic. A random seed may be contained in the hyperparameter and stored in the state to derive randomness deterministically. More explicitly:

- $(\text{st}_f, m) := f_I(\text{pp}_f)$ : The initialization function  $f_I$  takes as input hyperparameter  $\text{pp}_f$  and outputs the initial state  $\text{st}_f$  and model  $m$  (e.g., its initial weights).
- $(\text{st}_f, m) := f_T(\text{st}_f, D^+)$ : The training function  $f_T$  takes as input the current state  $\text{st}_f$  and a set  $D^+$  of data points to be added. It outputs the updated state and new model  $m$ .
- $(\text{st}_f, m) := f_U(\text{st}_f, U^+)$ : The unlearning function  $f_U$  takes as input the current state  $\text{st}_f$  and a set  $U^+$  of data points to be deleted. It outputs the updated state and new model  $m$ .

These functions allow us to establish an abstraction to track the relation between a model and its underlying dataset; we refer to this as the *conceptual* dataset. If  $D$  is the current conceptual dataset, removing data points from  $U^+$  with  $f_U$  updates the dataset as  $D := D \setminus U^+$ . We assume that server and

users agree on  $f$  before executing the protocol (e.g., similar to the TLS handshake protocol).

### A. Framework Overview

An overview of our framework is depicted in Figure 1. We denote a protocol in this framework by  $\Phi_f$ , where  $f = (f_I, f_T, f_U) \in \mathcal{F}$  is the triplet deployed by the protocol. We then describe the execution of the protocol with two phases (executed in an iterative manner):

**P1. Data Addition/Deletion:** Any user can issue an addition/deletion request to the server at any iteration. The server can batch multiple addition/deletion requests within a single iteration. For this, the server stores all requests in intermediate datasets  $D_i^+$  (addition) and  $U_i^+$  (deletion), respectively.

**P2. Proof of Training (resp. Unlearning):** At the end of each iteration  $i$ , the server updates its dataset by adding (resp. deleting) the data record stored in  $D_i^+$  (resp.  $U_i^+$ ), respectively. For training (resp. unlearning), the server needs to update the model using function  $f_T$  (resp.  $f_U$ ) on all records requested to be added (resp. deleted). It then computes a proof of training (resp. unlearning). This proof is verified by all users.

For unlearning, the server additionally needs to provide each user who requested a point to be unlearned with an individual proof that their data record was unlearned and removed from the server’s dataset (i.e., a proof of non-membership). After an update was performed, the dataset  $D_i^+$  (resp.  $U_i^+$ ) is reset to the empty set.

In the following, we describe the interfaces, including all algorithms, in more detail. The completeness properties of these algorithms are formally presented in Section IV-B.

**1. Setup and Initialization.** A global setup procedure Setup generates public parameters  $\text{pub}$ , i.e.,  $\text{pub} \leftarrow \text{Setup}(1^\lambda)$ , where  $\lambda$  is the security parameter. We assume that  $\text{pub}$  additionally includes the admissible functions  $f$  and the hyperparameter  $\text{pp}_f$ . This procedure can be executed either by the server or some external entity, depending on the application.  $\text{pub}$  is given to all actors. During initialization, the ML model and a state (which captures information needed for subsequent iterations) between the server and users are initialized using the  $\text{Init}$  and  $\text{VerifyInit}$  algorithms. Formally, the algorithms are defined as follows:

**Server:**  $(\text{st}_{S,0}, m_0, \text{com}_0, \rho_0) \leftarrow \text{Init}(\text{pub})$

$\text{Init}$  takes as input the public parameters  $\text{pub}$  and outputs the initial state  $\text{st}_{S,0}$ , the model  $m_0$ , the commitment  $\text{com}_0$ , and the proof  $\rho_0$ . The algorithm runs as follows: the server first suitably initializes model  $m_0$  using initialization algorithm  $f_I$  and additional hyperparameter  $\text{pp}_f$  contained in  $\text{pub}$ . It stores the resulting state  $\text{st}_f$  in  $\text{st}_{S,0}$ . The set of training records is initialized to be empty, i.e.,  $D_0 := \emptyset$ . It then commits to  $m_0$  and  $D_0$  with  $\text{com}_0 := (\text{com}_0^m \parallel \text{com}_0^D)$ . We assume that the

commitment to the initial dataset (and all its updates) is computed deterministically from  $\text{pub}$  using a function  $\text{Commit}$ , i.e.,  $\text{com}_0^D := \text{Commit}(\text{pub}, D_0)$ . Finally, proof  $\rho_0$  attests the initialization of model  $m_0$ .

**User:**  $0/1 \leftarrow \text{VerifyInit}(\text{pub}, \text{com}_0, \rho_0)$

$\text{VerifyInit}$  takes as input the public parameters  $\text{pub}$ , a commitment  $\text{com}_0$ , and a proof  $\rho_0$ . If the verification is successful, it outputs 1. On failure, it outputs 0. All users verifies (a) commitment  $\text{com}_0$  with  $D_0 = \emptyset$ , and (b) model initialization  $m_0$  against the proof  $\rho_0$ .

**2A. Proof of Training.** In each iteration  $i$  in which the server executes a proof of training, it updates the model with any newly added data records using the training function  $f_T$  and proves that this was performed correctly. The users verify the resulting proof. Formally, we define the two algorithms as follows:

**Server:**  $(\text{st}_{S,i}, m_i, \text{com}_i, \rho_i) \leftarrow \text{ProveTraining}(\text{st}_{S,i-1}, \text{pub}, D_i^+)$

$\text{ProveTraining}$  takes as input the previous state  $\text{st}_{S,i-1}$ , public parameters  $\text{pub}$ , the set of new data records  $D_i^+$ . It outputs the updated state  $\text{st}_{S,i}$ , the model  $m_i$ , the commitment  $\text{com}_i$ , and the proof  $\rho_i$ . The algorithm runs as follows: the server computes an updated model  $m_i$  obtained by executing training function  $f_T$  on state  $\text{st}_f$  and  $D_i^+$ . We define the resulting training set of  $m_i$  as the union of the previous dataset and all newly added data records, i.e.,  $D_i := D_{i-1} \cup D_i^+$ . The server commits to both the model and training data with  $\text{com}_i := (\text{com}_i^m \parallel \text{com}_i^D)$ . The server then computes the proof  $\rho_i$  that (a) model  $m_i$  was updated by applying  $f_T$ , and (b) training data  $D_i$  does not contain any unlearned record, i.e.,  $D_i \cap U_i = \emptyset$ , where  $U_i := \bigcup_{k \in [i]} U_k^+$  is the set of all unlearned data records so far. The proof also attests that (c) the set of unlearned data records has not changed, i.e.,  $U_{i-1} = U_i$ .

**User:**  $0/1 \leftarrow \text{VerifyTraining}(\text{pub}, \text{com}_{i-1}, \text{com}_i, \rho_i)$

$\text{VerifyTraining}$  takes as input the public parameters  $\text{pub}$ , two commitments  $\text{com}_{i-1}$ ,  $\text{com}_i$  and a proof  $\rho_i$ . It outputs 1 if the verification is successful, and 0 otherwise. All users validate properties (a)-(c) (as described above in  $\text{ProveTraining}$ ) and the update on commitment  $\text{com}_i$  by verifying the proof  $\rho_i$  against the previous commitment  $\text{com}_{i-1}$  and the new commitment  $\text{com}_i$ .

**2B. Proof of Unlearning.** In each iteration  $i$  in which the server performs a proof of unlearning, it first updates the model using function  $f_U$  and thus unlearning all data records requested to be unlearned. The server proves that this was performed correctly and then computes a proof of non-membership for each individual user who requested to delete their data point, proving that the corresponding records are *absent* in the training data of the updated model. All users verify the correct update. Those users who requested unlearning verify their proof of non-membership. Formally,

we define the following four algorithms:

**Server:**  $(st_{S,i}, m_i, com_i, \rho_i)$   
 $\leftarrow \text{ProveUnlearning}(st_{S,i-1}, \text{pub}, U_i^+)$

$\text{ProveUnlearning}$  takes as input the previous state  $st_{S,i-1}$ , public parameters  $\text{pub}$ , and the set  $U_i^+$  which contains data records to be removed. It outputs the updated state  $st_{S,i}$ , the updated model  $m_i$ , the commitment  $com_i$  and the proof  $\rho_i$ . Here, the server unlearns all records collected in  $U_i^+$  and computes the updated  $m_i$  by executing function  $f_U$ . Thus, conceptually, the new training set of  $m_i$  is defined as  $D_i := D_{i-1} \setminus U_i^+$ . Similar to  $\text{ProveTraining}$ , the server commits to both the model and training data with  $com_i$  and computes the proof  $\rho_i$  that (a) model  $m_i$  was updated by applying  $f_U$ , and (b) training data  $D_i$  does not contain any unlearned records, i.e.,  $D_i \cap U_i = \emptyset$ , where  $U_i := U_{i-1} \cup U_i^+$ . The proof also attests that (c) the previous set of unlearned data records is a subset of the updated set  $U_{i-1} \subset U_i$ . This ensures that an unlearned data record is never re-added into the training data.

**User:**  $0/1 \leftarrow \text{VerifyUnlearning}(\text{pub}, com_{i-1}, com_i, \rho_i)$

$\text{VerifyUnlearning}$  takes as input the public parameters  $\text{pub}$ , two commitments  $com_{i-1}$ ,  $com_i$  and a proof  $\rho_i$ . It outputs 1 if the verification is successful, and 0 otherwise. All users validate properties (a)-(c) (as described above in  $\text{ProveUnlearning}$ ) and the update on commitment  $com_i$  by verifying the proof  $\rho_i$  against the previous commitment  $com_{i-1}$  and the new commitment  $com_i$ .

**Server:**  $\pi_{u,d_{i,j}} \leftarrow \text{ProveNonMembership}(st_{S,i}, \text{pub}, u, d_{i,j})$

$\text{ProveNonMembership}$  takes as input the current state  $st_{S,i}$ , public parameters  $\text{pub}$  and a data record  $(u, d_{i,j})$ . Then, for each record  $(u, d_{i,j}) \in U_i^+$ , the server computes a proof  $\pi_{u,d_{i,j}}$  that this record is not part of the training set of model  $m_i$ , i.e.,  $(u, d_{i,j}) \notin D_i$ .

**User:**  $0/1 \leftarrow \text{VerifyNonMembership}(\text{pub}, u, d_{i,j}, com_i, \pi_{u,d_{i,j}})$

$\text{VerifyNonMembership}$  takes as input the public parameters  $\text{pub}$ , the user identifier  $u$ , their unlearned data point  $d_{i,j}$ , the commitment of the iteration where  $d_{i,j}$  was unlearned and the proof of unlearning  $\pi_{u,d_{i,j}}$ . It outputs the result of the verification. The user verifies with both  $\pi_{u,d_{i,j}}$  and  $com_i$  that  $(u, d_{i,j})$  was not part of the training data  $D_i$  of model  $m_i$ .

**Redundant computation.** The framework, as described above, requires all users to verify proofs generated at every iteration, including those iterations where they do not perform any data addition/deletion. This results in a lot of redundant computation performed across users. If the users trust a third party (e.g., an auditor), then the  $\text{VerifyTraining}$  and the  $\text{VerifyUnlearning}$  algorithm can be executed by this entity; the output can be shared with all users (cf. Section VII-B for further discussion).

## B. Completeness

For completeness, we require that an honest execution of the protocol yields the expected outputs. In particular, if the server is honest, then the users successfully verify the initialization of the model and the proofs for all updates—training and unlearning—performed by the server. Further, a proof of non-membership that was generated for an unlearned data record is also successfully verified by the corresponding user. In the following, we give a formal definition of computational completeness.

**Definition 1** (Completeness). *Let  $\lambda$  be the security parameter. A protocol  $\Phi_f$  is complete if for all  $\text{pub} \leftarrow \text{Setup}(1^\lambda)$ , the following properties are satisfied:*

1) *Let  $(st_{S,0}, m_0, com_0, \rho_0) \leftarrow \text{Init}(\text{pub})$ . Then*

$$\Pr[\text{VerifyInit}(\text{pub}, com_0, \rho_0) = 0] \leq \text{negl}(\lambda) .$$

2) *Let  $mode_i \in \{\text{train}, \text{unlearn}\}$  indicate whether proof of training or proof of unlearning has been performed in iteration  $i$ . Let  $\mathcal{A}$  be a PPT adversary that outputs a valid sequence of datasets either to be added  $\{\text{train}: D_i^+\}$  or to be deleted  $\{\text{unlearn}: U_i^+\}$  for all  $i \in [\ell]$ .*

*For all  $i \in [\ell]$ , if  $mode_i = \text{train}$ , let  $(st_{S,i}, m_i, com_i, \rho_i) \leftarrow \text{ProveTraining}(st_{S,i-1}, \text{pub}, D_i^+)$  and if  $mode_i = \text{unlearn}$ , let  $(st_{S,i}, m_i, com_i, \rho_i) \leftarrow \text{ProveUnlearning}(st_{S,i-1}, \text{pub}, U_i^+)$ .*

*Then for all  $mode_i = \text{train}$ :*

$$\Pr[\text{VerifyTraining}(\text{pub}, com_{i-1}, com_i, \rho_i) = 0] \leq \text{negl}(\lambda) ,$$

*and for all  $mode_i = \text{unlearn}$ :*

$$\Pr[\text{VerifyUnlearning}(\text{pub}, com_{i-1}, com_i, \rho_i) = 0] \leq \text{negl}(\lambda) ,$$

*where validity is defined via the following conditions:*

$$\forall i, j \text{ s.t. } i \neq j: D_i^+ \cap D_j^+ = \emptyset \text{ and } \forall i, j \text{ s.t. } j < i: D_i^+ \cap U_j^+ = \emptyset .$$

3) *For all  $i \in [\ell]$  s.t.  $mode_i = \text{unlearn}$ : for all  $(u, d) \in U_i^+$ , let  $\pi_{u,d} \leftarrow \text{ProveNonMembership}(st_{S,i}, \text{pub}, u, d)$ , then*

$$\Pr[\text{VerifyNonMembership}(\text{pub}, u, d, com_i, \pi_{u,d}) = 0] \leq \text{negl}(\lambda) .$$

We require computational completeness here to allow for a wide range of instantiations. For example, an instantiation that works on hash values of data records cannot achieve perfect completeness because of hash collisions. By allowing for computational completeness, however, we only require that it should be hard for a PPT adversary to find such collisions (i.e., with a negligible probability).

## C. Security Definition

In this section, we present the security definition for unlearning in a game  $\text{GameUnlearn}$ . The adversary, described by a probabilistic algorithm  $\mathcal{A}$ , takes the role of the server. Intuitively, the definition captures that a malicious server cannot add (and train on) a data point that a user requested to delete in a previous iteration. However, we go a step further

```

GameUnlearn $\mathcal{A}, \mathcal{E}, \Phi_f, \mathcal{D}(1^\lambda)$ 
00 pub  $\leftarrow$  Setup( $1^\lambda$ )
01  $(k, (u, d), \pi_{u,d}, \{mode_i: com_i, \rho_i\}_{i \in [0:\ell]}; \{D_i\}_{i \in [0:\ell]}) \leftarrow (\mathcal{A} \parallel \mathcal{E})(pub, aux)$ 
02 # Pre-processing
03  $U_k^+ := D_{k-1} \setminus D_k$ 
04 Parse  $com_i$  as  $(com_i^m \parallel com_i^D) \forall i \in [0 : \ell]$ 
05 # Evaluate winning condition
06 if Commit(pub,  $D_i$ ) =  $com_i^D \forall i \in [0 : \ell]$  # Datasets
07 and VerifyInit(pub,  $com_0, \rho_0$ ) # Initialization
08 and VerifyTraining(pub,  $com_{i-1}, com_i, \rho_i) \forall i: mode_i = train$  # Training
09 and VerifyUnlearning(pub,  $com_{i-1}, com_i, \rho_i) \forall i: mode_i = unlearn$  # Unlearning
10 and VerifyNonMembership(pub,  $u, d, com_k, \pi_{u,d}$ ) # Non-Membership
11 and  $k < \ell$  and  $(u, d) \in U_k^+$  and  $(u, d) \in D_\ell$ : # Point unlearned & re-added
12 return 1
13 return 0

```

Fig. 2: *Security Game*. We define the security of an unlearning protocol  $\Phi_f$  in terms of game GameUnlearn. The notation  $(\mathcal{A} \parallel \mathcal{E})$  denotes that both algorithms are run on the same input and random coins and assigning their results to variables before resp. after the semicolon. Input aux refers to auxiliary input.

and let the server choose *which* data records it will add or delete. Thus, the goal of the adversary is to find a data record for which it can prove deletion, but which is re-added in some subsequent iteration.

In order to capture this setting, we propose an *extractability-based* security definition as used in the context of hash functions or SNARKs [14], [24], [34]. That is, we require the existence of an extractor, modelling that the adversary cannot forge a transcript without knowing the underlying datasets. Thus, the adversary in our game has to provide the protocol outputs (*i.e.*, the commitments and proofs), whereas the extractor outputs the corresponding inputs that the adversary used (*i.e.*, the underlying datasets). We give a formal description of game GameUnlearn in Figure 2, which is divided into the following two stages:

**S1. Simulation.** The game draws the public parameters pub using Setup and runs the adversary  $\mathcal{A}$  on input pub. The extractor  $\mathcal{E}$  is run on the same input and random coins. Additionally, we provide benign auxiliary input aux, which captures any extra information that the adversary may have (possibly obtained prior to the start of executing the current protocol). At some point,  $\mathcal{A}$  will terminate and output a sequence of tuples  $(k, (u, d), \pi_{u,d}, \{mode_i: com_i, \rho_i\}_{i \in [0,\ell]})$  for some  $\ell \in \mathbb{N}$ , where  $(u, d)$  is a data record that was proved to be deleted in the  $k$ -th iteration, and  $mode_i \in \{train, unlearn\}$ . At the same time, the extractor outputs a sequence of datasets  $(D_0, \dots, D_\ell)$ .

**S2. Finalize.** After the adversary has terminated, the game uses the extractor’s output to compute the set of data points unlearned in the  $k$ -th iteration based on the datasets  $D_k$  and  $D_{k-1}$ . Recall that the commitment in the framework consists of two parts  $com_i^m$  and  $com_i^D$ , where we need the second part for verification. The game checks for the following conditions: (a)  $com_i^D$  was obtained from  $D_i$ , (b) the initial proof  $\rho_0$  verifies for the initial

commitment  $com_0$ , (c) each proof of training  $\rho_i$  verifies for commitments  $com_{i-1}$  and  $com_i$ , (d) each proof of unlearning  $\rho_i$  verifies for commitments  $com_{i-1}$  and  $com_i$ , (e) the proof of non-membership  $\pi_{u,d}$  verifies for  $(u, d)$  and  $com_k$ , (f)  $k < \ell$  and  $(u, d)$  was unlearned in iteration  $k$  and re-added in iteration  $\ell$ . If all these properties are satisfied, then the game outputs 1 and  $\mathcal{A}$  wins.

We summarize this in the following definition.

**Definition 2 (Unlearning).** Let  $\lambda$  be the security parameter and consider game GameUnlearn in Figure 2. Protocol  $\Phi_f$  for data distribution  $\mathcal{D}$  is *unlearning-secure* if for all PPT adversaries  $\mathcal{A}$  there exists an extractor  $\mathcal{E}$  such that for all benign auxiliary inputs aux:

$$\Pr[\text{GameUnlearn}_{\mathcal{A}, \mathcal{E}, \Phi_f, \mathcal{D}}(1^\lambda) \Rightarrow 1] \leq \text{negl}(\lambda).$$

## V. INSTANTIATION

Our framework defines a general interface to construct protocols for verifiable unlearning. In the following, we present such a protocol based on cryptographic building blocks. We use SNARKs and hash functions where the execution of the admissible functions is proved inside the SNARK and data records are stored in hashed form. By using SNARKs, we can keep the instantiation generic and universally prove its completeness and security for any triplet  $(f_I, f_T, f_U)$ . A detailed description of the protocol is in Figure 4 in Appendix A.

**Data Representation.** We internally split the data into training data  $D$  and unlearned data  $U$ . Therefore, the server stores two ordered sets  $\mathcal{H}_D$  and  $\mathcal{H}_U$  of hashed training data records and unlearned data records. From both sets, we additionally compute a hash value in form of a hash chain. This allows for efficient caching of intermediate hashes and, for  $\mathcal{H}_U$ , enables us to easily prove that entries are only appended to the chain as well as fast membership verification for unlearned data points. To account for the partition of training and unlearned data and the admissible function used, we instantiate the commitment com as a tuple of four elements: (a) hash of the state  $h_{st_f}$  (defined by admissible function  $f$ ), (b) hash of the model  $h_m$ , (c) hash of the training data  $h_D$ , and (d) hash of the unlearned data  $h_U$ . A formal description of how exactly the hash values are computed is given in Appendix A.

**Proof System.** In order to prove the correct execution of  $f_I$ ,  $f_T$ , and  $f_U$ , we use proof systems and more specifically SNARKs. To this end, we define the verification of the initialization, training updates and unlearning updates in terms of a polynomial decidable binary relation  $R_I$ ,  $R_T$  and  $R_U$  (respectively) over circuits  $C_I$ ,  $C_T$  and  $C_U$  (respectively). These circuits are outlined in Figure 3 and described further below.

**1. Initialization.** During the protocol’s initialization, function  $f_I$  is run to obtain the initial state  $st_{f,0}$  and initial model  $m_0$ . Also, the sets of hashed training data and unlearned data records are initialized, *i.e.*,  $\mathcal{H}_{D_0} = \emptyset$  and  $\mathcal{H}_{U_0} = \emptyset$ . The commitment consists of hashes to these four values,

```

CI(public hstf,0, hm0, hD0, hU0, private stf,0, m0)
00 # Check input for initialization
01 if hstf,0 ≠ HashState(stf,0) or
   hm0 ≠ HashModel(m0) or
   hD0 ≠ HashData(∅) or
   hU0 ≠ HashData(∅):
02   return false
03 return true

CT(public hstf,i, hstf,i-1, hmi, hDi, hDi-1, hUi, hUi-1, private stf,i-1, HUi-1, Di+)
04 # Check input set of hashed unlearned data records
05 if hUi-1 ≠ HashData(HUi-1):
06   return false
07 # Update and check set of hashed training data records and unlearned data records
08 HDi+ := {HashDataRecord(u, d)}(u,d)∈Di+
09 if hDi ≠ AppendHashData(hDi-1, HDi+) or hUi ≠ hUi-1 or HUi-1 ∩ HDi+ ≠ ∅:
10   return false
11 # Check input state, perform training and check outputs
12 if hstf,i-1 ≠ HashState(stf,i-1):
13   return false
14 (stf,i, mi) := fT(stf,i-1, Di+)
15 if hstf,i ≠ HashState(stf,i) or hmi ≠ HashModel(mi):
16   return false
17 return true

CU(public hstf,i, hstf,i-1, hmi, hDi, hDi-1, hUi, hUi-1, private stf,i-1, HDi-1, Ui+)
18 # Check input set of hashed training data records
19 if hDi-1 ≠ HashData(HDi-1):
20   return false
21 # Update and check set of hashed unlearned data records and training data records
22 HUi+ := {HashDataRecord(u, d)}(u,d)∈Ui+
23 HDi := HDi-1 \ HUi+
24 if hUi ≠ AppendHashData(hUi-1, HUi+) or hDi ≠ HashData(HDi):
25   return false
26 # Check input state, perform unlearning and check outputs
27 hstf,i-1 ≠ HashState(stf,i-1):
28   return false
29 (stf,i, mi) := fU(stf,i-1, Ui+)
30 if hstf,i ≠ HashState(stf,i) or hmi ≠ HashModel(mi):
31   return false
32 return true

```

Fig. 3: Circuits  $C_I$ ,  $C_T$  and  $C_U$ . Based on this circuits, we prove correct execution of admissible functions for initialization, proof of training and proof of unlearning. The hash algorithms are further specified in Appendix A.

*i.e.*,  $\text{com}_0 = (h_{\text{st}_{f,0}}, h_{m_0}, h_{D_0}, h_{U_0})$ . Correct initialization is proved using the SNARK for relation  $R_I$  captured by circuit  $C_I$  (cf. Figure 3). The proof of training  $\rho_0$  consists of the statement  $\phi_0$  and resulting SNARK proof  $\pi_0$ , which can be verified by the user using  $\text{com}_0$ .

**2A. Proof of Training.** The server starts by executing `ProveTraining`. In the  $i$ -th iteration, it first performs the model update by running function  $f_T$  on the previous state  $\text{st}_{f,i-1}$  and new data records  $D_i^+$ , the result being an updated state  $\text{st}_{f,i}$  and a new model  $m_i$ . Then the server updates the set of hashed training data records  $\mathcal{H}_{D_i}$  with  $D_i^+$  and computes the new commitment  $\text{com}_i = (h_{\text{st}_{f,i}}, h_{m_i}, h_{D_i}, h_{U_{i-1}})$ , where the commitment to the unlearned data records is the same as in the previous iteration since no data was deleted.

The proof  $\rho_i$  is computed using the SNARK for relation  $R_T$  captured by circuit  $C_T$  (cf. Figure 3). The corresponding statement  $\phi_i$  and proof  $\pi_i$  attest that (a) the model and state were updated correctly with  $D_i^+$ , (b) the set of hashed unlearned data was not changed, and (c) no data record that was previously unlearned is added. The server sends  $(\rho_i, \text{com}_i)$  to the users. Subsequently, the users execute `VerifyTraining` and

verify  $\rho_i$  using  $\text{com}_i$  and the previous commitment  $\text{com}_{i-1}$ .

**2B. Proof of Unlearning.** The proof of unlearning consists of two parts: the model update for deleting data records and the proof of non-membership. The server first runs `ProveUnlearning` which is similar to `ProveTraining`. In the  $i$ -th iteration, it performs the model update by running function  $f_U$  on the previous state  $\text{st}_{f,i-1}$  and the set  $U_i^+$  of data records to be deleted. The result is the updated state  $\text{st}_{f,i}$  and model  $m_i$ . The set  $\mathcal{H}_{U_i}$  is computed by appending hashed records of  $U_i^+$  to  $\mathcal{H}_{U_{i-1}}$ . At the same time,  $\mathcal{H}_{D_i}$  is computed from  $\mathcal{H}_{D_{i-1}}$  by removing those entries. The commitment  $\text{com}_i$  consists of the hash values  $(h_{\text{st}_{f,i}}, h_{m_i}, h_{D_i}, h_{U_i})$ . The whole procedure is proved using circuit  $C_U$  (cf. Figure 3) for relation  $R_U$ , producing a SNARK proof  $\pi_i$  for the corresponding statement  $\phi_i$ , which can be verified by the user using  $\text{com}_i$  and  $\text{com}_{i-1}$ .

The second part of the proof of unlearning is to provide a proof on non-membership to all users that requested their data record  $(u, d) \in U_i^+$  to be deleted. We prove this by proving its membership in  $\mathcal{H}_{U_i}$ . If  $\mathcal{H}_{U_i} \cap \mathcal{H}_{D_i} = \emptyset$ , it follows that  $(u, d) \notin D_i$  (which we show to hold when proving completeness). Specifically, we use the hash chain for  $\mathcal{H}_{U_i}$ : for a given data record, we compute a membership proof as a path in this chain; this path can be verified by recomputing the hash chain and comparing the final result with the hash in the commitment (*i.e.*, hash value  $h_{U_i}$ ).

Thus, in our protocol, the server performs `ProveNonMembership` by computing the chain path to a data record  $(u, d) \in U_i^+$  using the procedure `ComputeChainPath` (cf. Appendix A). It outputs this as the proof  $\pi_{u,d}$  which is then sent to the user. The user uses the hash  $h_{U_i}$  from the commitment to verify membership with `VerifyChainPath`. If the path leads to that hash, the user will accept, and will abort otherwise.

#### A. Completeness and Security

We first show that our instantiation is complete according to Definition 1.

**Theorem 1.** *Let  $\Pi$  be a complete SNARK and Hash a collision-resistant hash function. Then the instantiated protocol in Figure 4 satisfies completeness.*

We give a proof sketch; refer to Appendix A for the full proof.

*Proof (Sketch).* Completeness of the initialization (first property) is easy to observe since the two hashed datasets are initialized as empty and the execution of function  $f_I$  is proven with the SNARK for relation  $R_I$ . By completeness of the SNARK, the users can successfully verify the proof, additionally using the commitments to state, model and datasets. The second property follows from the completeness of the SNARKs for relations  $R_T$  and  $R_U$  and collision-resistance of the hash function. However, note that if a hash collision occurs, it is not possible to provide the proof of training. Thus, only computational completeness can be achieved. Given that the

proofs of training and unlearning are successful, completeness of the proof of non-membership (third property) follows from the construction and correctness of the hash chain.  $\square$

Now we want to prove that our instantiation is a *secure* unlearning protocol according to Definition 2.

**Theorem 2.** *Let Hash be a collision-resistant hash function and  $\Pi$  be a secure SNARK. Then the instantiated protocol in Figure 4 satisfies unlearning security.*

We give a proof sketch; refer to Appendix A for the full proof.

*Proof (Sketch).* Let  $\mathcal{A}$  be an adversary in the unlearning security game (cf. Figure 2). By knowledge soundness of the SNARK, there exists an extractor which outputs the witness and thus the datasets  $D_i$  corresponding to the outputs of the adversary. We then use the soundness of the SNARK. That is,  $\mathcal{A}$  must have computed the proof using a witness, *i.e.*, the state  $\text{st}_{f,i}$  and the dataset  $D_i^+$  (in the proof of training) or dataset  $U_i^+$  (in the proof of unlearning), which also determine the model  $m_i$  and must correspond to the hash values in the commitment. By collision-resistance of the hash function, the adversary cannot find another state, model or dataset for the same commitment. Thus, applying function  $f_T$  (or  $f_U$ ) to the previous state and datasets results in same state and model as used by  $\mathcal{A}$ .

Since all proofs as well as the proof of non-membership of data record  $(u, d)$  must verify successfully, the hash of  $(u, d)$  must be contained in the set  $\mathcal{H}_{U_k}$  which was used to create the proof. Here,  $k$  is the iteration where  $(u, d)$  was unlearned; the observation holds by assuming soundness of the SNARK and collision-resistance of the hash function. We can further infer that  $(u, d)$  must also be part of future sets  $\mathcal{H}_{U_i}$ ,  $k < i \leq \ell$  and by collision-resistance  $(u, d)$  must also be part of the underlying datasets  $U_i$ . Finally, we use the fact that the proof attests that the intersection of  $\mathcal{H}_{U_\ell}$  and  $\mathcal{H}_{D_\ell}$  is empty. This yields a contradiction and shows that  $(u, d)$  cannot be present in the last dataset  $D_\ell$ .  $\square$

## VI. EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of our instantiated protocol. First, we implement and compare the protocol’s main building blocks for three types of unlearning approaches captured by admissible functions. The techniques from machine unlearning literature we consider are retraining-based unlearning, amnesiac unlearning and optimization-based unlearning (cf. Section II-A). Second, we study the applicability to different ML models and datasets.

The goal of our experiments is to evaluate the feasibility of verifiable machine unlearning and understand how different unlearning techniques influence the proof creation and verification times. Our salient findings include:

- 1) The majority of the costs stem from the cost of performing verifiable training, which is at most  $5\times$  more expensive than verifiable unlearning. In scenarios where one trusts the training process, this results in immediate

savings. Note that we did not aim to optimize the verifiable computation component; this is orthogonal to the problem considered in this paper. We provide suggestions on how to achieve better performance in Section VII-A.

- 2) Unlearning techniques that rely on simple mechanisms such as adding/subtracting information from model parameters (*e.g.*, amnesiac unlearning) are intuitively cheaper to prove (in comparison to retraining-based approaches). However, hidden costs emerge in having to verify the integrity of the inputs needed for such methods.

All experiments are performed on a server running Ubuntu 22.04 with 256 GB RAM and two Intel Xeon Gold 5320 CPUs. Our code is available at <https://github.com/cleverhans-lab/verifiable-unlearning>.

### A. Cryptographic Primitives

**Proof System.** Our instantiation is generic and can be implemented with any secure SNARK (cf. Section II-B) *i.e.*, the SNARK needs to satisfy completeness, soundness, and knowledge soundness. In this work, we use Spartan [60] as it is efficient and, more importantly, transparent, *i.e.*, it does not require a trusted setup. Spartan comes in two variants, as a succinct non-interactive zero-knowledge (NIZK) proof system and as a SNARK. Similar to the work of Angel *et al.* [6], we use the NIZK variant, where verification time is linear in the size of the R1CS instance (see below). By using the SNARK variant, some verification cost can be offset to the server and a one-time pre-processing step for the user.

Depending on the application it might also be sensible to use a different proof system. One alternative would be Groth16 [34], which, for example, requires a trusted setup, but has the advantage of constant verification time and proof size.

Spartan is implemented on the `ristretto255` elliptic curve, a prime-order group abstraction atop `curve25519`. Following prior work on verifiable computation [6], [53], [75], we convert the computation of our circuits into *Rank-1 Constraint Systems* (R1CS) instances; *i.e.*, the statements in  $R_I$ ,  $R_T$  and  $R_U$  (cf. Figure 3) are represented as a constraint system over a *finite field*. More specifically, an R1CS instance is described by a tuple  $(\mathbb{F}, A, B, C, io, m)$ , where  $\mathbb{F}$  is the finite field,  $A, B, C \in \mathbb{F}^{m \times m}$  are matrices of size  $m \geq |io| + 1$  and  $io$  is the public input and output of the instance. R1CS is a generalization of arithmetic circuit satisfiability. We say an R1CS instance is satisfiable if there exists a witness  $w \in \mathbb{F}^{m-|io|-1}$  such that  $(A \cdot z) \circ (B \cdot z) = (C \cdot z)$  for  $z = (io, 1, w)$ , where  $\cdot$  is the matrix-vector product and  $\circ$  the Hadamard product. Since  $A, B, C$  are generally sparse matrices, a parameter  $n$  is sometimes specified, denoting the maximum number of non-zero entries in each matrix.

We describe our circuits using the ZoKrates programming language [21] and use the CirC compiler infrastructure [52] to facilitate the conversion to R1CS. The compiler ensures that the computation graph does not have loops, and a “flat” computation is performed. To represent data and other parameters in a finite field, we convert them into *fixed precision* real numbers.

TABLE I: *Run-Time of Protocol Functions.* We compare the running time between the protocols subtasks. We consider retraining-based unlearning, amnesiac unlearning, and optimization-based unlearning. We report the relative difference with retraining in gray.

	Retraining		Amnesiac		Optimization	
<i>Proof of Training</i>						
RICS	8,056,887	×1.00	8,130,535	×1.01	7,980,878	×0.99
II.Prove w/ $R_T$	4m 32s	×1.00	4m 32s	×1.00	4m 31s	×0.99
II.Vrfy w/ $R_T$	1m 36s	×1.00	1m 37s	×1.01	1m 35s	×0.99
<i>Proof of Unlearning</i>						
RICS	8,102,288	×1.00	616,005	×0.08	919,456	×0.11
II.Prove w/ $R_U$	4m 58s	×1.00	2m 18s	×0.46	0m 53s	×0.18
II.Vrfy w/ $R_U$	1m 48s	×1.00	0m 49s	×0.45	0m 20s	×0.19
<i>Proof of Non-Membership</i>						
ComputeChainPath	< 1s	×1.00	< 1s	×1.00	< 1s	×1.00
VerifyChainPath	< 1s	×1.00	< 1s	×1.00	< 1s	×1.00

RICS: #constraints

**Hash Function.** We only require collision-resistance for the hash function. It is beneficial to use an algebraic hash function where most operations can be directly done in the finite field. Bit-wise hash functions such as the SHA family of hash functions are much slower in that regard. We use Poseidon [32] as it is particularly designed for zero-knowledge proof systems. To be used with Spartan, we implement a version of Poseidon for the `ristretto255` curve. Similar to the proof system, our instantiation is generic and can work with any hash function. Other good options are *Pedersen Hash* [39, p.76] or *MIMC* [5].

### B. Protocol Instantiation

We implement the high-level functions of the instantiated protocol (from Section V) for retraining-based unlearning, amnesiac unlearning [33], and optimization-based unlearning [40], [72] as introduced in Section II-A. We consider the subtasks of proof of training, proof of unlearning, and proof of non-membership.

We start our evaluation by comparing and understanding the overheads of each subtask between the techniques. To this end, we consider a linear regression model and train this model for 3 epochs with SGD as a general purpose approach. We use a synthetic dataset  $D$  and set the batch size to 1. First, we compute a proof of training with the addition of 100 data points with 10 features each. We set  $|D_0| = 0$ ,  $|D_1^+| = 100$ , and  $|U_0| = 0$  accordingly. Subsequently, we compute the proof of unlearning and simulate the deletion of 10 data points and set  $|D_1| = 100$ ,  $|U_1| = 0$ , and  $|U_2^+| = 10$ . For optimization-based unlearning, we unlearn for 3 epochs.

The results from these experiments are presented in Table I. Across all techniques, compilation time of RICS instances ranges between 17s ( $C_U$  for optimization-based unlearning) and 48m 45s ( $C_U$  for retraining-based unlearning).

**Proof of Training.** We observe that the complexity of the training is comparable between unlearning approaches. The underlying RICS instances have between 7,980,878—8,130,535 constraints and proving time varies insignificantly between 4m 31s—4m 32s. Recall that in amnesiac unlearning,

TABLE II: *Proving Time vs. Model Capacity.* We compare the proving time of proof of training for different classes of models with increasing capacity.

Classifier	RICS	II.Prove	II.Vrfy
Linear Regression	8,056,887	4m 33s	1m 36s
Logistic Regression	9,048,909	5m 8s	1m 45s
Neural Network ( $N = 2$ )	21,867,010	9m 50s	3m 34s
Neural Network ( $N = 4$ )	42,030,731	24m 16s	6m 42s

RICS: #constraints

we also need to collect model updates that are later used for unlearning, which introduces negligible overhead compared to the training costs itself.

**Proof of Unlearning.** Runtime of generating and verifying the proof of unlearning shows more variance. Amnesiac unlearning is over  $2\times$  faster and optimization-based unlearning over  $5\times$  faster than retraining-based unlearning. This is despite the RICS instance of optimization-based unlearning being almost 50% larger compared to the amnesiac instance (919,456 vs. 616,005 constraints) but it is still more efficient to compute as it is 63% more sparse (*i.e.*, 7,660,455 vs. 12,248,390 entries are non-zero). The main difference is that amnesiac unlearning requires to maintain and verify a state from training (*i.e.*, the model updates) while optimization-based unlearning does not require a state.

**Proof System.** In general, we observe that verification is  $2\times$ – $3\times$  faster than proof generation. This is dependent on the choice of the proof system. For example, by using the SNARK variant of Spartan, we can offload some of the verification costs to the server and an additional pre-processing for the user. In this case, proving time increases to 33m 39s—34m 12s for the proof of training across all techniques and verification time reduces to  $\mu$ s, but the user needs to run a one-time pre-processing step which takes between 8m 6s—8m 12s.

**Proof of Non-Membership.** Finally, proof of non-membership is very efficient. The implementation is independent of the unlearning scheme used and both proving and verification requires  $< 1s$ .

### C. Circuit Complexity

The dominant component of the protocols’ run-time is the complexity of the circuit used to generate proofs of training and unlearning. This complexity depends mainly on (a) the unlearning technique, (b) the complexity of the model, and (c) the size of the dataset. In the following, we first consider model complexity and study different classes of models. Next, we look on the complexity of the dataset. In both cases, we focus on retraining-based unlearning as the baseline from Table I and, more specifically, on the training circuit  $C_T$ .

**Model Complexity.** To understand the effects of the choice of ML model, we follow related work [77], and consider linear regression, logistic regression and neural networks for classification. For the neural networks, we focus on models with one hidden layer and varying numbers of (hidden) neurons

TABLE III: *Scalability to Benchmark Datasets.* We compute the proof of training for different datasets from the PMLB benchmark suite [56].

Dataset	Size		RICS	II.Prove	II.Vrfy
Creditscore	100	6	3,986,308	2m 22s	0m 47s
Patient	88	8	4,579,718	2m 28s	0m 53s
Cy Young	92	10	5,903,988	3m 16s	1m 9s
Corral	160	6	6,347,236	3m 43s	1m 15s
Lawsuit	264	4	7,190,981	4m 7s	1m 27s
Breast cancer	286	9	16,514,048	9m 25s	3m 18s
Monk3	554	6	21,841,281	13m 36s	4m 32s

Size: #data points  $\times$  #features  
 RICS: #constraints

$N \in \{2, 4\}$ . For activation, we use the sigmoid function and approximate it with a third-order polynomial as done in [42], [43]. Again, we train each model with SGD for 3 epochs on a synthetic dataset consisting of 100 training points with 10 features each.

The results are summarized in Table II. We observe that the circuit size increases together with the complexity of the model. For instance, the number of RICS constraints increases by  $1.12 \times$  to 9,048,909 constraints when going from linear to logistic regression. This is intuitive: in logistic regression, we additionally need to evaluate the sigmoid activation which induces this overhead. In a similar vein, moving from logistic regression to neural networks increases the circuit further to 21,867,010 ( $N = 2$ ) and 42,030,731 ( $N = 4$ ) constraints respectively.

**Benchmark Datasets.** To understand the impact of the dataset and the practical applicability of the protocol, we now turn to benchmark datasets. We choose several datasets from the PMLB benchmark suite [56] (as considered in related work [6] on verifiable computation of numerical optimization problems) and train a linear regression model for all datasets. To make results comparable, we train all models for 3 epochs with a learning rate of 0.1. As commonly done, we split the data into 80:20 train test split. Models achieve a test accuracy between 73% and 92%.

Results are presented in Table III. For all models, we observe a linear dependence between run-time and dataset size. Generating a proof for the smallest dataset with 600 total features (*i.e.*, total points  $\times$  features) requires 2m 22s and for the largest dataset with 3,324 total features requires 13m 36s.

## VII. DISCUSSION

In this section, we discuss potential improvements to our work.

### A. Scalability

Our experiments with the instantiated system show that the run-time of the protocol is dominated by generating and verifying the proof of training and unlearning. We base our construction on *Verified Computation* (VC) and, consequently, inherit its limitations, *e.g.*, in terms of scalability. This can also be observed for other VC-based approaches in the ML setting [64], [38], [42], [43]. Any future advances in VC will

lead to run-time improvements for our approach. Nevertheless, we discuss how one can improve performance with the primitives available today.

**SNARK-friendly Techniques.** It is known that certain computations are more amendable to efficient SNARK verification than others. A classic example of this is the development of SNARK-friendly hash functions [32], [5], [39]. Similarly, there exist ML paradigms that are also more amenable to verification. For example, inference using quantized models [41], [23] or lookup tables for expensive computations [41] reduce costs. Furthermore, when there exists a unique ML model (*i.e.*, a global optimum for the underlying optimization problem), proving and verification complexity can be improved even further [6]. In our experiments, we observed that online computation of model updates in optimization-based unlearning is faster than verifying model updates in amnesiac unlearning as the verification of input values involves expensive calculation of hash values. We envision future work to focus on developing SNARK-friendly unlearning techniques combining above observations.

**Offloading Computation.** Orthogonal to the employed unlearning technique and ML model, one can offload expensive proof generation steps to the user (*e.g.*, the evaluation of a non-linear activation function). We can split the proving and verification processes such that the server creates a proof for certain types of computations and shares partial results with the (honest) user who performs (and thus verifies) expensive computations themselves.

**Application-specific Relaxations.** Finally, depending on the application, it might be possible to avoid the expensive generation of the proof of training. Consider, for instance, an application where data is collected only once and data will *only* be removed at a later point in time (*e.g.*, biomedical user studies or other human-involved data collection processes). In this case, proof of training only needs to be performed once and—if users further trust the initial training phase—it might be sufficient to only prove unlearning.

### B. Alternative Instantiations

**External Trust.** Our instantiation in Section V avoids having a trusted third party and instead relies only on cryptographic protocols to guarantee security. For efficiency purposes and to remove the burden from the user, one can introduce a trusted auditor who verifies on behalf of a user (as we discuss towards the end of Section IV-A). This can be achieved by either having a dedicated trusted third party (*e.g.*, one that does not have a motivation to collude with the server such as another cloud provider), or distributed auditors where trust is established from multiple independent verifications.

**Trusted Hardware.** If TEEs (*e.g.*, Intel SGX [49]) are available, then one can run training and unlearning procedures within it and return a digest signed by a TEE provider to the user. When using a TEE, one needs to consider common

concerns such as trusting a hardware vendor, availability of said vendor for signing the digest, limited memory [35], their applicability to ML-related tasks that involve GPU computation [69], and side-channels [51], [71]. Some of these issues were addressed in the independent and concurrent work of Weng *et al.* [74] (cf. Section VIII).

**Minimizing Redundancy.** In our instantiation, a user who has requested unlearning is required to verify future updates to ensure that their data point has not been re-added. If we combine VC with an additional proof of secure data erasure, we can give similar guarantees while not requiring the user to verify all updates. However, secure erasure is a non-trivial problem in itself and was considered in *e.g.*, [55]. Formalizing deletion compliance from a server’s perspective [27] can also be seen as complementary problem.

### C. Privacy

Formalizing privacy for unlearning protocols is an interesting direction for future work and requires to establish an additional security definition. Although it is out-of-scope for our work, we want to highlight that our instantiation does not require the users to know the datasets or model. In fact, they only see hash commitments and the SNARK proofs. If the hash function satisfies pseudo-randomness or is modeled as a random oracle, then hash values do not leak any information about the underlying data points as long as the input space is large enough. Additionally, if the SNARK satisfies the zero-knowledge property [31] (which most SNARKs including Spartan do), then the proof also does not leak information about the witness. However, we require users to know whether training or unlearning happened because they need to know which verification procedure to run. Privacy in the context of model inference has been studied more extensively, *e.g.*, Gao *et al.* [25] define security notions for deletion hiding and reconstruction. An overview for different formalizations of inference privacy is also given in [57].

## VIII. RELATED WORK

Our approach for verifiable machine unlearning naturally touches different areas of security and ML research. In the following, we examine related concepts and methods.

**Verifying Unlearning.** Prior work [26], [66] aims at verifying unlearning by embedding backdoors [36] in models (using data whose unlearning is to be verified) and verifying backdoor removal on unlearning. However, such approaches are probabilistic with no theoretical guarantees of when they work, unlike our cryptography-informed approach which produces verifiable proofs.

The work of Guo *et al.* [37] provides end-users with a certificate that the new model is influenced by the specific data in a quantifiably low manner. While this certificate conceptually bounds the influence of a data point from an algorithmic perspective, it provides no guarantee that the entity executing the algorithm (*i.e.*, server) did so correctly. In our

work, we aim to capture exactly this and provide cryptographic guarantees of correctness of execution.

Concurrently to our work, Weng *et al.* [74] propose an unlearning framework based on TEE. They model unlearning in two phases: a setup phase, where the user sends data which is used to train an ML model, and a deletion phase, where a new model is trained without the data point that the user requests to delete. Their protocol uses unlearning based on SISA [15] and can be captured by our framework as well. In contrast to our instantiation that is based only on cryptographic primitives, their approach relies on trusted hardware (*i.e.*, the correctness and integrity of the SGX enclave) as well as cryptographic assumptions (*i.e.*, EUF-CMA security of the signature scheme used by the enclave and collision-resistance of the hash function).

**Proving Model Inference.** There exist various approaches to proving inference using SNARKs [46], [48], [73], [23], [41] which complements our protocol in that regard. Another approach would be to use trusted execution environments to do so as suggested in [74].

**Verifiable Computation.** We use verifiable computation for proof of training and proof of unlearning. There has been a series of works demonstrating a remarkable progress in making these schemes (and those related to verification of data used for computation) practical [60], [63], [16], [62], [70], [61], [45], [10], [11], [13], [8], [12], [9], [28], [54], [24]. To verify the computation of training an ML model, Zhao *et al.* [77] also propose verification using a SNARK. However, their primary objective is to design a scheme to ensure that the payments made to servers are correct. In our work, however, we aim to design a scheme to verify the correctness of data deletion when training ML models. Otti [6] is a compiler that is aimed at designing efficient arithmetic circuits for problems that involve optimization (such as those commonly found in ML). DIZK [75] is a distributed system capable of distributing the compute required for proof creation.

## IX. CONCLUSION

The problem of unlearning has gained significant interest in terms of definitions and algorithms for updating model parameters. However, regardless of the definition or the algorithm the server uses to update the model, the user has no way to verify that the server indeed executed the unlearning procedure. In this paper, we define unlearning as a security problem and propose a framework to capture the guarantees verifiable unlearning needs to provide. We propose the first verifiable unlearning procedure based on cryptographic primitives instantiated using SNARKs and hash chains. Our implementation shows the feasibility of our approach on several benchmark datasets and machine learning models. Future work includes determining which unlearning techniques are most suitable for efficient verifiable computation, while at the same time devising methods specifically for verifying machine learning code.

## ACKNOWLEDGEMENTS

Thorsten Eisenhofer and Doreen Riepel were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC 2092 CASA - 390781972. Nicolas Papernot would like to acknowledge his sponsors, who support his research with financial and in-kind contributions, including Apple, CIFAR through the Canada CIFAR AI Chair program, DARPA through the GARD program, Intel, NSERC through the Discovery grant, Meta, and Ontario through the Early Researcher Award program. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. We would like to thank members of the CleverHans Lab for their feedback. We would also like to thank Sebastian Angel, Jess Woods, and Eleftherios Ioannidis for input related to the SNARK compilers.

## REFERENCES

- [1] Certificate Transparency. Official website.
- [2] General Data Protection Regulation (GDPR). Official Legal Text, 2016.
- [3] California Consumer Privacy Act (CCPA). Official Legal Text, 2018.
- [4] Personal Information Protection and Electronic Documents Act (PIPEDA). Official Legal Text, 2019.
- [5] Martin R. Albrecht, Lorenzo Grassi, Christian Rechberger, Arnab Roy, and Tyge Tiessen. MiMC: Efficient Encryption and Cryptographic Hashing with Minimal Multiplicative Complexity. In *International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT)*, 2016.
- [6] Sebastian Angel, Andrew J. Blumberg, Eleftherios Ioannidis, and Jess Woods. Efficient Representation of Numerical Optimization Problems for SNARKs. In *USENIX Security Symposium*, 2022.
- [7] Thomas Baumhauer, Pascal Schöttle, and Matthias Zeppelzauer. Machine Unlearning: Linear Filtration for Logit-based Classifiers. *Machine Learning*, 2022.
- [8] Eli Ben-Sasson, Iddo Bentov, Yinon Horesh, and Michael Riabzev. Scalable Zero Knowledge with No Trusted Setup. In *Annual International Cryptology Conference (CRYPTO)*, 2019.
- [9] Eli Ben-Sasson, Alessandro Chiesa, Daniel Genkin, and Eran Tromer. On the Concrete Efficiency of Probabilistically-Checkable Proofs. In *Symposium on Theory of Computing Conference (STOC)*, 2013.
- [10] Eli Ben-Sasson, Alessandro Chiesa, Daniel Genkin, Eran Tromer, and Madars Virza. SNARKs for C: Verifying Program Executions Succinctly and in Zero Knowledge. In *Annual International Cryptology Conference (CRYPTO)*, 2013.
- [11] Eli Ben-Sasson, Alessandro Chiesa, Eran Tromer, and Madars Virza. Succinct Non-Interactive Zero Knowledge for a von Neumann Architecture. In *USENIX Security Symposium*, 2014.
- [12] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Short PCPs Verifiable in Polylogarithmic Time. In *IEEE Conference on Computational Complexity (CCC)*, 2005.
- [13] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of Proximity, Shorter PCPs, and Applications to Coding. *SIAM Journal on Computing*, 2006.
- [14] Nir Bitansky, Ran Canetti, Alessandro Chiesa, and Eran Tromer. From Extractable Collision Resistance to Succinct Non-Interactive Arguments of Knowledge, and Back Again. In *Innovations in Theoretical Computer Science (ITCS)*, 2012.
- [15] Lucas Bourtole, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2021.
- [16] Benjamin Braun, Ariel J. Feldman, Zuo Cheng Ren, Srinath T. V. Setty, Andrew J. Blumberg, and Michael Walfish. Verifying Computations with State. In *ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, 2013.
- [17] Gavin Brown, Mark Bun, Vitaly Feldman, Adam D. Smith, and Kunal Talwar. When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning? In *ACM SIGACT Symposium on Theory of Computing (STOC)*, 2021.
- [18] Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2015.
- [19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *USENIX Security Symposium*, 2019.
- [20] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph Unlearning. In *ACM Conference on Computer and Communications Security (CCS)*, 2021.
- [21] Jacob Eberhardt and Stefan Tai. ZoKrates - Scalable Privacy-Preserving Off-Chain Computations. In *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2018.
- [22] Vitaly Feldman. Does Learning Require Memorization? A Short Tale About a Long Tail. In *ACM SIGACT Symposium on Theory of Computing (STOC)*, 2020.
- [23] Boyuan Feng, Lianke Qin, Zhenfei Zhang, Yufei Ding, and Shumo Chu. ZEN: Efficient Zero-Knowledge Proofs for Neural Networks. *Cryptology ePrint Archive*, 2021.
- [24] Dario Fiore, Cédric Fournet, Esha Ghosh, Markulf Kohlweiss, Olga Ohrimenko, and Bryan Parno. Hash First, Argue Later: Adaptive Verifiable Computations on Outsourced Data. In *ACM Conference on Computer and Communications Security (CCS)*, 2016.
- [25] Ji Gao, Sanjam Garg, Mohammad Mahmood, and Prashant Nalini Vasudevan. Deletion Inference, Reconstruction, and Compliance in Machine (Un)learning. In *Privacy Enhancing Technologies Symposium (PETS)*, 2022.
- [26] Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. VeriFi: Towards Verifiable Federated Unlearning. *Computing Research Repository (CoRR)*, 2022.
- [27] Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing Data Deletion in the Context of the Right to Be Forgotten. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2020.
- [28] Rosario Gennaro, Craig Gentry, and Bryan Parno. Non-interactive Verifiable Computing: Outsourcing Computation to Untrusted Workers. In *Annual International Cryptology Conference (CRYPTO)*, 2010.
- [29] Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning (ICML)*, 2019.
- [30] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [31] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The Knowledge Complexity of Interactive Proof Systems. *SIAM Journal on Computing (SICOMP)*, 1989.
- [32] Lorenzo Grassi, Dmitry Khovratovich, Christian Rechberger, Arnab Roy, and Markus Schofnegger. Poseidon: A New Hash Function for Zero-Knowledge Proof Systems. In *USENIX Security Symposium*, 2021.
- [33] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac Machine Learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- [34] Jens Groth. On the Size of Pairing-Based Non-interactive Arguments. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2016.
- [35] Karan Grover, Shruti Tople, Shweta Shinde, Ranjita Bhagwan, and Ramachandran Ramjee. Privado: Practical and Secure DNN Inference with Enclaves. *Computing Research Repository (CoRR)*, 2018.
- [36] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *Computing Research Repository (CoRR)*, 2017.
- [37] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified Data Removal from Machine Learning Models. In *International Conference on Machine Learning (ICML)*, 2020.
- [38] Inbar Helbitz and Shai Avidan. Reducing ReLU Count for Privacy-Preserving CNN Speedup. *Computing Research Repository (CoRR)*, 2021.
- [39] Daira Hopwood, Sean Bowe, Taylor Hornby, and Nathan Wilcox. Zcash. Protocol Specification (Version 2022.3.4), 2022.

- [40] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge Unlearning for Mitigating Privacy Risks in Language Models. *Computing Research Repository (CoRR)*, 2022.
- [41] Daniel Kang, Tatsunori Hashimoto, Ion Stoica, and Yi Sun. Scaling up Trustless DNN Inference with Zero-Knowledge Proofs. *Computing Research Repository (CoRR)*, 2022.
- [42] Andrey Kim, Yongsoo Song, Miran Kim, Keewoo Lee, and Jung Hee Cheon. Logistic Regression Model Training based on the Approximate Homomorphic Encryption. *Cryptology ePrint Archive*, 2018.
- [43] Miran Kim, Yongsoo Song, Shuang Wang, Yuhou Xia, and Xiaoqian Jiang. Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation. *JMIR Medical Informatics*, 2018.
- [44] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning (ICML)*.
- [45] Jonathan Lee, Kirill Nikitin, and Srinath Setty. Replicated State Machines without Replicated Execution. In *IEEE Symposium on Security and Privacy (S&P)*, 2020.
- [46] Seunghwan Lee, Hankyung Ko, Jihye Kim, and Hyunok Oh. vCNN: Verifiable Convolutional Neural Network. *Cryptology ePrint Archive*, 2020.
- [47] Klas Leino and Matt Fredrikson. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In *USENIX Security Symposium*, 2020.
- [48] Tianyi Liu, Xiang Xie, and Yupeng Zhang. ZkCNN: Zero Knowledge Proofs for Convolutional Neural Network Predictions and Accuracy. In *ACM Conference on Computer and Communications Security (CCS)*, 2021.
- [49] Frank McKeen, Ilya Alexandrovich, Alex Berenson, Carlos V. Rozas, Hisham Shafi, Vedvyas Shanbhogue, and Uday R. Savagaonkar. Innovative Instructions and Software Model for Isolated Execution. In *Workshop on Hardware and Architectural Support for Security and Privacy (HASP)*, 2013.
- [50] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning. In *Algorithmic Learning Theory (ALT)*, 2021.
- [51] Olga Ohrimenko, Felix Schuster, Cedric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. Oblivious Multi-Party Machine Learning on Trusted Processors. In *USENIX Security Symposium*, 2016.
- [52] Alex Ozdemir, Fraser Brown, and Riad S. Wahby. CirC: Compiler Infrastructure for Proof Systems, Software Verification, and more. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.
- [53] Alex Ozdemir, Riad S. Wahby, Barry Whitehat, and Dan Boneh. Scaling Verifiable Computation Using Efficient Set Accumulators. In *USENIX Security Symposium*, 2020.
- [54] Bryan Parno, Jon Howell, Craig Gentry, and Mariana Raykova. Pinocchio: Nearly Practical Verifiable Computation. In *IEEE Symposium on Security and Privacy (S&P)*, 2013.
- [55] Daniele Perito and Gene Tsudik. Secure Code Update for Embedded Devices via Proofs of Secure Erasure. In *European Symposium on Research in Computer Security (ESORICS)*, 2010.
- [56] Joseph D. Romano, Trang T. Le, William G. La Cava, John T. Gregg, Daniel J. Goldberg, Praneel Chakraborty, Natasha L. Ray, Daniel S. Himmelstein, Weixuan Fu, and Jason H. Moore. PMLB v1.0: An Open-Source Dataset Collection for Benchmarking Machine Learning Methods. *Bioinformatics*, 2022.
- [57] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella Béguelin. SoK: Let The Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning. *Computing Research Repository (CoRR)*, 2022.
- [58] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember What You Want to Forget: Algorithms for Machine Unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [59] Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations (ICLR)*, 2017.
- [60] Srinath Setty. Spartan: Efficient and General-Purpose zkSNARKs Without Trusted Setup. In *Annual International Cryptology Conference (CRYPTO)*, 2020.
- [61] Srinath Setty, Sebastian Angel, Trinabh Gupta, and Jonathan Lee. Proving the Correct Execution of Concurrent Services in Zero-Knowledge. In *Symposium on Operating Systems Design and Implementation (OSDI)*, 2018.
- [62] Srinath Setty, Victor Vu, Nikhil Panpalia, Benjamin Braun, Andrew J Blumberg, and Michael Walfish. Taking Proof-Based Verified Computation a Few Steps Closer to Practicality. In *USENIX Security Symposium*, 2012.
- [63] Srinath TV Setty, Richard McPherson, Andrew J Blumberg, and Michael Walfish. Making Argument Systems for Outsourced Computation Practical (Sometimes). In *Symposium on Network and Distributed System Security (NDSS)*, 2012.
- [64] Avital Shafra, Gil Segev, Shmuel Peleg, and Yedid Hoshen. Crypto-Oriented Neural Architecture Design. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2021.
- [65] Iliia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, and Ross J. Anderson. Manipulating SGD with Data Ordering Attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [66] David Marco Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Towards Probabilistic Verification of Machine Unlearning. *Computing Research Repository (CoRR)*, 2020.
- [67] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding Factors Influencing Machine Unlearning. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2022.
- [68] Anvith Thudi, Hengrui Jia, Iliia Shumailov, and Nicolas Papernot. On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning. In *USENIX Security Symposium*, 2022.
- [69] Stavros Volos, Kapil Vaswani, and Rodrigo Bruno. Graviton: Trusted Execution Environments on GPUs. In *Symposium on Operating Systems Design and Implementation, (OSDI)*, 2018.
- [70] Victor Vu, Srinath Setty, Andrew J Blumberg, and Michael Walfish. A Hybrid Architecture for Interactive Verifiable Computation. In *IEEE Symposium on Security and Privacy (S&P)*, 2013.
- [71] Wenhao Wang, Guoxing Chen, Xiaorui Pan, Yinqian Zhang, XiaoFeng Wang, Vincent Bindschaedler, Haixu Tang, and Carl A Gunter. Leaky Cauldron on the Dark Land: Understanding Memory Side-Channel Hazards in SGX. In *ACM Conference on Computer and Communications Security (CCS)*, 2017.
- [72] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine Unlearning of Features and Labels. In *Symposium on Network and Distributed System Security (NDSS)*, 2023.
- [73] Chenkai Weng, Kang Yang, Xiang Xie, Jonathan Katz, and Xiao Wang. Mystique: Efficient Conversions for Zero-Knowledge Proofs with Applications to Machine Learning. In *USENIX Security Symposium*, 2021.
- [74] Jia-Si Weng, Shenglong Yao, Yuefeng Du, Junjie Huang, Jian Weng, and Cong Wang. Proof of Unlearning: Definitions and Instantiation. *Computing Research Repository (CoRR)*, 2022.
- [75] Howard Wu, Wenting Zheng, Alessandro Chiesa, Raluca Ada Popa, and Ion Stoica. DIZK: A Distributed Zero Knowledge Proof System. In *USENIX Security Symposium*, 2018.
- [76] Yinjun Wu, Edgar Dobriban, and Susan B. Davidson. DeltaGrad: Rapid Retraining of Machine Learning Models. In *International Conference on Machine Learning (ICML)*, 2020.
- [77] Lingchen Zhao, Qian Wang, Cong Wang, Qi Li, Chao Shen, and Bo Feng. VeriML: Enabling Integrity Assurances and Fair Payments for Machine Learning as a Service. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 2021.

APPENDIX

A schematic overview of our complete instantiated protocol is given in Figure 4. Additional algorithms are described below.

```

HashData( $\mathcal{H}$ )
00  $\Psi := \text{Hash}(d_0)$ 
01 for  $h_d \in \mathcal{H}$ :
02    $\Psi := \text{Hash}(\Psi, h_d)$ 
03 return  $\Psi$ 

AppendHashData( $h, \mathcal{H}$ )
04  $\Psi := h$ 
05 for  $h_d \in \mathcal{H}$ :
06    $\Psi := \text{Hash}(\Psi, h_d)$ 
07 return  $\Psi$ 

HashDataRecord( $u, d = (x, y)$ )
08  $h_d := \text{Hash}(u)$ 
09 for  $x_j \in x$ :
10    $h_d := \text{Hash}(h_d, \text{Hash}(x_j))$ 
11  $h_d := \text{Hash}(h_d, \text{Hash}(y))$ 
12 return  $h_d$ 

ComputeChainPath( $u, d, \mathcal{H}_U$ )
13  $h_d := \text{HashDataRecord}(u, d)$ 
14  $idx_d := \mathcal{H}_U.\text{index}(h_d)$ 
15 if  $idx_d = \perp$ :
16   return  $\perp$ 
17 # get intermediate hash from chain below  $d$ 
18  $\Psi := \text{HashData}(\mathcal{H}_U : idx_d)$ 
19  $\pi_{u,d} := [\Psi]$ 
20 # add path from  $d$ 
21 for  $h_d \in \mathcal{H}_{U_{idx_d+1}}$ :
22    $\pi_{u,d}.\text{append}(h_d)$ 
23 return  $\pi_{u,d}$ 

VerifyChainPath( $u, d, h_U, \pi_{u,d}$ )
24 # recompute hash  $\Psi$  from path  $\pi_{u,d}$ 
25  $\Psi := \text{Hash}(\pi_{u,d}[0], \text{HashDataRecord}(u, d))$ 
26 for node in  $\pi_{u,d}[1: ]$ :
27    $\Psi := \text{Hash}(\Psi, \text{node})$ 
28 # verify final hash
29 return  $\llbracket \Psi = h_U \rrbracket$ 

HashModel( $m = [w_0, \dots, w_n]$ )
30  $h_m := \text{Hash}(m[0])$ 
31 for  $w_i \in m[1: ]$ :
32    $h_m := \text{Hash}(h_m, \text{Hash}(w_i))$ 
33 return  $h_m$ 

HashState( $st_f$ )
34  $h_{st_f} := \text{Hash}(st_f[0])$ 
35 for  $s_i \in h_{st_f}[1: ]$ :
36    $h_{st_f} := \text{Hash}(h_{st_f}, s_i)$ 
37 return  $h_{st_f}$ 

```

*Proof (of Theorem 1).* We need to prove the three properties in Definition 1 capturing the initialization, the proof of training and the proof of unlearning which includes the proof of non-membership.

Initialization. First, running `Init` yields the initialized state  $st_{f,0}$  and model  $m_0$  obtained by executing  $f_I$ . Using the hash function to commit to those two values and additionally the empty sets  $\mathcal{H}_{D_0}$  and  $\mathcal{H}_{U_0}$ , an instance  $(\phi_0, \omega_0) \in R_I$  can be derived and the SNARK proof  $\pi_0$  can be created using `II.Prove`. By correctness of the relation and completeness of the SNARK,  $\phi_0$  will be valid for  $com_0$  and  $\text{II.Vrfy}(R_I, pp_I, \pi_0, \phi_0) = 1$ .

Proof of Update. For the second property, recall the inputs and outputs of `ProveTraining` and `ProveUnlearning`. The state  $st_{S,i-1}$  contains the state  $st_{f,i-1}$ , the set  $\mathcal{H}_{D_{i-1}}$  of hashed data records, the set  $\mathcal{H}_{U_{i-1}}$  of hashed unlearned data records and the previous commitment  $com_{i-1}$ . In the proof of training, the new state  $st_{f,i}$  and the new model  $m_i$  are computed by running function  $f_T$  on  $st_{f,i-1}$  and the set  $D_i^+$  of data records to be added. The new sets  $\mathcal{H}_{D_i}$  and  $\mathcal{H}_{U_i}$  are computed from the previous ones and updated with  $D_i^+$ . The commitment is computed by hashing the four components.

Then the training instance  $(\phi_i, \omega_i) \in R_T$  can be derived and the SNARK proof  $\pi_i$  computed. The proof attests (cf. Figure 3) that (1)  $m_i$  was computed correctly since  $f_T$  was executed, (2) the training data does not contain unlearned record since the hashes of the new data records are not contained in  $\mathcal{H}_{U_i}$ , and (3) the set of unlearned data records has not changed since the commitments to the unlearned data records are the same. By correctness of the relation and perfect completeness of the SNARK, we have  $\text{II.Vrfy}(R_T, pp_T, \pi_i, \phi_i) = 1$ .

Note that there exists a special case where the server is unable to create a proof although the datasets are valid. This is the case whenever there exist two distinct data records  $(u, d) \in D_i^+$ ,  $(u', d') \in U_i$ , where  $U_i = \bigcup_{j \in [i]} U_j^+$  is the dataset implicitly contained in  $\mathcal{H}_{U_i}$ , such that  $\text{HashDataRecord}(u, d) = \text{HashDataRecord}(u', d')$ . However, we only require computational completeness and assume that the datasets are provided by a PPT adversary. Then this translates to finding a collision for the hash function which happens with negligible probability if the hash function is collision-resistance. Hence, `VerifyTraining` will output 1 with probability  $1 - \text{negl}(\lambda)$ .

Proof of Unlearning. Completeness for the proof of unlearning proceeds similar. The new state  $st_{f,i}$  and the new model  $m_i$  are computed by running function  $f_U$  on  $st_{f,i-1}$  and the set  $U_i^+$  of data records to be deleted. The new sets  $\mathcal{H}_{D_i}$  and  $\mathcal{H}_{U_i}$  are computed from the previous ones and updated by removing and appending  $U_i^+$ , respectively. The commitment is computed by hashing the four components.

The unlearning instance  $(\phi_i, \omega_i) \in R_U$  is derived and the SNARK proof  $\pi_i$  that is computed attests (cf. Figure 3) that (1)  $m_i$  was computed correctly since  $f_U$  was executed, (2) the training data does not contain unlearned record since we removed the records in  $U_i^+$  from  $\mathcal{H}_{D_i}$ , and (3) previous set of unlearned data records is a subset of the updated set since we added the records in  $U_i^+$  to  $\mathcal{H}_{U_i}$  to which we commit. By correctness of the relation and perfect completeness of the SNARK, we have  $\text{II.Vrfy}(R_U, pp_U, \pi_i, \phi_i) = 1$  and `VerifyUnlearn` will output 1 with probability 1.

Finally, consider the algorithm `ProveNonMembership`. If a data record  $(u, d)$  was unlearned in iteration  $i$ , then its hash is present in  $\mathcal{H}_{U_i}$ . The proof of non-membership  $\pi_{u,d}$  consists of the chain path to  $(u, d)$  in the chain of  $\mathcal{H}_{U_i}$ . Let  $com_i$  be the commitment for this iteration, then by correctness of the tree path algorithm, `VerifyNonMembership` will output 1 with probability 1.  $\square$

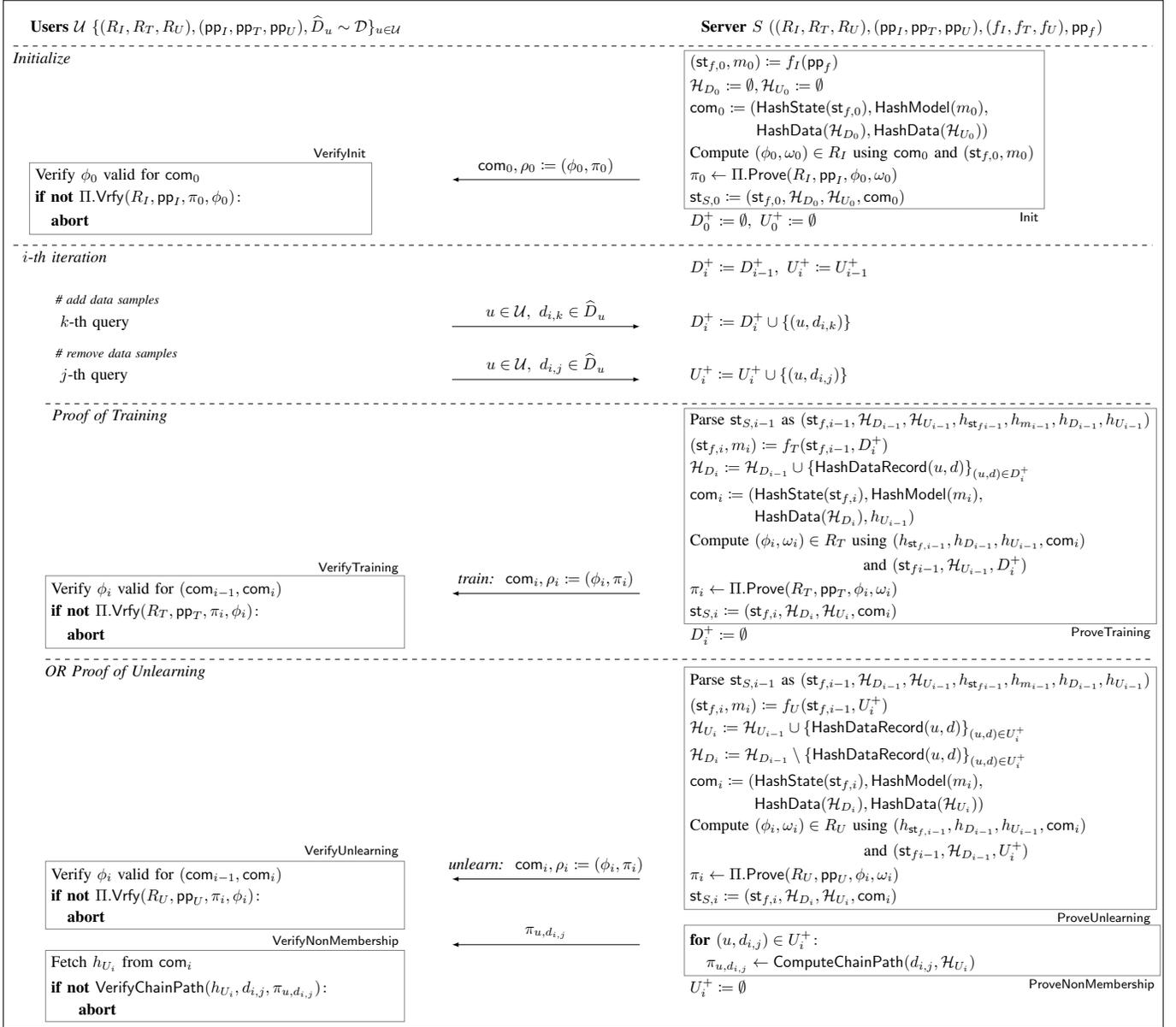


Fig. 4: **Instantiated Protocol  $\Phi_f$** . We instantiate the protocol for the triple of admissible functions  $f = (f_I, f_T, f_U)$  with two primitives: a SNARK  $\Pi$ , and a hash function Hash.

*Proof (of Theorem 2).* Let  $\mathcal{A}$  be an adversary against unlearning security (as defined in Figure 2) of our instantiation. We will first argue that for all  $\mathcal{A}$  there exists an extractor  $\mathcal{E}$  that outputs the underlying datasets  $D_i$ . This follows directly from the knowledge soundness of the SNARK for relations  $R_I, R_T$  and  $R_U$ . For this, look at the private inputs to the circuits in Figure 3 which translate to the witness. Initialization gives us that  $D_0 = \emptyset$ . The proof of training inputs  $D_i^+$  and the proof of unlearning inputs  $U_i^+$  such that we can extract  $D_i = D_{i-1} \cup D_i^+$  if  $\text{mode}_i = \text{train}$  and  $D_i = D_{i-1} \setminus U_i^+$  if  $\text{mode}_i = \text{unlearn}$ .

We will now prove the theorem by the sequence of games given in Figure 5 and analyze the probability that these games

will output 1.

**Game  $G_0$ .** Let  $G_0$  be the original game  $\text{GameUnlearn}$  and  $\mathcal{E}$  be the extractor. Recall that the adversary must output a sequence of tuples  $(k, (u, d), \pi_{u,d}, \{\text{mode}_i: \text{com}_i, \rho_i\}_{i \in [0:\ell]})$  for some  $\ell \in \mathbb{N}$ , where  $\text{com}_i = (h_{\text{st}_{f,i}}, h_{m_i}, h_{D_i}, h_{U_i})$  and  $\rho_i = (\phi_i, \pi_i)$  for  $i \in [0:\ell]$ . We iterate over the winning conditions and return 0 as soon as one of them is violated (cf. Figure 5). For book-keeping we also compute all sets of unlearned data points  $U_i$ , as well as the sets  $D_i^+, U_i^+$  from  $D_i$  as described for the extractor. Note that this is only a conceptual change at this point and we have

$$\Pr[G_0 \Rightarrow 1] = \Pr[\text{GameUnlearn}_{\mathcal{A}, \mathcal{E}, \Phi_f, \mathcal{D}}(1^\lambda) \Rightarrow 1].$$

```

G0-G2
00 ppI ← Π.Setup(1λ, RI)
01 ppT ← Π.Setup(1λ, RT)
02 ppU ← Π.Setup(1λ, RU)
03 (k, (u, d), πu,d, {modei: (hstf,i, hmi, hDi, hUi), (φi, πi)}i∈[0:ℓ];
   {Di}i∈[0:ℓ]) ← (A|E)(RI, RT, RU, ppI, ppT, ppU, fI, fT, fU, ppf)
04
05 # Pre-processing
06 U0 := ∅
07 for i ∈ [ℓ]
08   if modei = train:
09     Di+ := Di \ Di-1
10   if modei = unlearn:
11     Ui+ := Di-1 \ Di
12     Ui := Ui-1 ∪ Ui+
13
14 # Verify commitments
15 for i ∈ [0 : ℓ]:
16   HDi := {HashDataRecord(u, d)}(u,d)∈Di
17   h'Di := HashData(HDi)
18   if h'Di ≠ hDi:
19     return 0
20
21 # Verify initialization
22 Verify φ0 valid for (hstf,0, hm0, hD0, hU0)
23 if not Π.Vrfy(RI, ppI, π0, φ0):
24   return 0
25
26 # Re-compute initialization
27 (stf,0, m0) := fI(ppf) // G1-G2
28 if hstf,0 ≠ HashState(hstf,0) or hm0 ≠ HashModel(m0): // G1-G2
29   return 0 // G1-G2
30 # Verify proof of training
31 for i ∈ [ℓ] s.t. modei = train:
32   Verify φi valid for (hstf,i, hmi, hDi, hUi)
33   if not Π.Vrfy(RT, ppT, πi, φi):
34     return 0
35
36 # Verify proof of unlearning
37 for i ∈ [ℓ] s.t. modei = unlearn:
38   Verify φi valid for (hstf,i, hmi, hDi, hUi)
39   if not Π.Vrfy(RU, ppU, πi, φi):
40     return 0
41
42 # Re-compute state and model and compare to commitment
43 for i ∈ [ℓ]: // G1-G2
44   if modei = train: // G1-G2
45     (stf,i, mi) := fT(stf,i-1, Di+) // G1-G2
46   if modei = unlearn: // G1-G2
47     (stf,i, mi) := fT(stf,i-1, Ui+) // G1-G2
48   if hstf,i ≠ HashState(stf,i) or hmi ≠ HashModel(mi): // G1-G2
49     return 0 // G1-G2
50
51 # Verify proof of non-membership
52 if not VerifyChainPath(hUk, u, d, πu,d):
53   return 0
54
55 # Check membership of d in Ui
56 for i ∈ [k : ℓ]: // G2
57   if (u, d) ∉ Ui: // G2
58     return 0 // G2
59
60 # Adversary wins if point unlearned & re-added
61 if k < ℓ and (u, d) ∈ Uk+ and (u, d) ∈ Dℓ:
62   return 1
63 return 0

```

Fig. 5: Games  $G_0$ - $G_2$  for the proof of Theorem 2. We prove unlearning security for our instantiated protocol  $\Phi_f$  in Figure 4, where  $f = (f_I, f_T, f_U)$  and hyperparameter  $pp_f$  are fixed by the participating parties and determine relations  $R_I, R_T$  and  $R_U$ .

**Game  $G_1$ .** In  $G_1$ , we compute the state  $st_{f,i}$  and the model  $m_i$  for each iteration from the corresponding datasets by applying  $f_I, f_T$  and  $f_U$ . We then check whether the hashes of state and model correspond to  $h_{st_{f,i}}$  and  $h_{m_i}$  in the commitment. If this is not the case, the game outputs 0. We claim

$$|\Pr[G_1 \Rightarrow 1] - \Pr[G_0 \Rightarrow 1]| \leq \text{negl}(\lambda).$$

To prove the claim we argue in the following steps:

- First,  $\pi_i$  proves that the adversary knows a state  $st'_{f,i}$  and a dataset  $D_i^{+'}$  for each proof of training (or a dataset  $U_i^{+'}$  for each proof of unlearning) such that model  $m'_i$  was computed by applying function  $f_T$  (or function  $f_U$ ) to state  $st'_{f,i-1}$  and dataset  $D_i^{+'}$  (or dataset  $U_i^{+'}$ ). It also proves that the commitment aligns with the inputs. Since the functions are deterministic, we thus have  $h_{st_{f,i}} = \text{HashState}(st'_{f,i})$  and  $h_{m_i} = \text{HashData}(m'_i)$  as well as  $h_{D_i} = \text{HashData}(H'_{D_i})$ , where  $H'_{D_i}$  is the set of all hashed data records in  $D_i^{+'}$ .

By soundness of the SNARK, the adversary can only forge a proof for an invalid statement with negligible probability, so we can assume the proof was generated honestly with a witness. By knowledge soundness, the extractor is able to compute this witness such that  $D_i' = D_i$ .

- Second, we claim that then  $st'_{f,i} = st_{f,i}$  and  $m'_i = m_i$  are the actual state and model used for the next iteration. This is true unless the adversary finds a collision in the hash function such that  $\text{HashState}(st'_{f,i}) = \text{HashState}(st_{f,i}) = h_{st_{f,i}}$  or  $\text{HashModel}(m'_i) = \text{HashModel}(m_i) = h_{m_i}$ , which we assume to happen only with negligible probability.

**Game  $G_2$ .** In  $G_2$ , we check whether the data record  $(u, d)$  output by  $\mathcal{A}$  is contained in the underlying datasets  $U_i$  of the  $k$ -th and all subsequent iterations. We will show that

$$|\Pr[G_2 \Rightarrow 1] - \Pr[G_1 \Rightarrow 1]| \leq \text{negl}(\lambda).$$

For this we first look again at the SNARK proof  $\pi_i$  and the underlying circuits. If a proof of training is performed, the adversary must prove that  $h_{U_i} = h_{U_{i-1}}$ . This implies—assuming no hash collision occurs—that  $\mathcal{H}_{U_i} = \mathcal{H}_{U_{i-1}}$  and  $U_i = U_{i-1}$ . If a proof of unlearning is performed, the SNARK proof ensures that  $\mathcal{H}_{U_{i-1}} \subset \mathcal{H}_{U_i}$  and thus  $U_{i-1} \subset U_i$ , again using collision-resistance of the hash function. Thus, if  $(u, d) \in U_k$ , it must also be true that  $(u, d) \in U_{k+1}, \dots, (u, d) \in U_\ell$ . By soundness of the SNARK, the adversary cannot prove a false statement, so the above claims must hold.

We also know that  $(u, d) \in U_k$  since the proof of non-membership consists of the path from the hashed data record

$(u, d)$  to the hash  $h_{U_k}$  contained in the  $k$ -th commitment. Since the adversary can only win if the proof verifies successfully, we know that in this case the hash value of  $(u, d)$ , in the following denoted by  $h_{u,d} := \text{HashDataRecord}(u, d)$ , must be a node in the hash chain constructed from  $\mathcal{H}_{U_k}$ . Unless the adversary finds another data record  $(u', d')$  such that  $\text{HashDataRecord}(u', d')$  maps to the same hash value  $h_{u,d}$ —which happens with negligible probability—the record  $(u, d)$  must be contained in  $U_k$ .

Finally, we show that  $\Pr[G_2 \Rightarrow 1] \leq \text{negl}(\lambda)$ . For this, recall that  $\pi_i$  also attests that no unlearned data point is contained in the dataset, in particular that the intersection  $\mathcal{H}_{D_i} \cap \mathcal{H}_{U_i}$  is empty. Together with the fact that the commitments  $h_{D_i}$  and  $h_{U_i}$  are constructed from  $\mathcal{H}_{D_i}$  and  $\mathcal{H}_{U_i}$  (due to soundness of the SNARK), the hashed datasets must have been obtained from the corresponding dataset  $D_i$  and  $U_i$  (unless the adversary has found a collision in the hash function). Combining with previous results, this implies that  $D_i \cap U_i = \emptyset$  for all  $i \in [\ell]$ . As shown above, we know that  $(u, d) \in U_\ell$ . The final winning condition requires that  $(u, d) \in D_\ell$ . This cannot be the case since it would contradict the fact that the intersection of the two sets is empty, which proves the final claim.

Collecting the probabilities yields

$$\Pr[\text{GameUnlearn}_{\mathcal{A}, \mathcal{E}, \Phi_f, \mathcal{D}}(1^\lambda)] \leq \text{negl}(\lambda) ,$$

which concludes the proof of Theorem 2.  $\square$