Published in S. Dolev et al., Eds., *Cyber Security Cryptography and Machine Learning (CSCML 2021)*, vol. 12716 of *Lecture Notes in Computer Science*, pp. 1-19, Springer, 2021.

Programmable Bootstrapping Enables Efficient Homomorphic Inference of Deep Neural Networks

Ilaria Chillotti, Marc Joye, and Pascal Paillier

Zama, France

Abstract. In many cases, machine learning and privacy are perceived to be at odds. Privacy concerns are especially relevant when the involved data are sensitive. This paper deals with the privacy-preserving inference of deep neural networks.
We report on first experiments with a new library implementing a variant of the TFHE fully homomorphic encryption scheme. The underlying key technology is the programmable bootstrapping. It enables the homomorphic evaluation of any function of a ciphertext, with a controlled level of noise. Our results indicate for the first time that deep neural networks are now within the reach of fully homomorphic encryption. Importantly, in contrast to prior works, our framework does not necessitate re-training the model.

Keywords: Fully homomorphic encryption · Programmable bootstrapping · Data privacy · Machine learning · Deep neural networks.

1 Introduction

Machine learning algorithms are extremely useful in many areas but the type of data that they deal with is often sensitive. Typical examples include algorithms for the detection of certain genetic diseases from DNA samples or the ones used for face recognition or email classification, to name a few. The processed data contain private information about users and could be used in many ways, from target advertising to blackmail or even threat in some cases. This is why it is essential to protect the data being used in machine learning applications. Privacy requirements are also pushed by recent regulations companies dealing with user's data must comply with, like the GDPR (General Data Protection Regulation) [14] in Europe or the CCPA (California Consumer Privacy Act) [8] in the US.

Fully homomorphic encryption. Cryptographic techniques are methods of choice when it comes to the protection of data. But traditional encryption algorithms merely protect data while it is in transit or at rest. Indeed, one limitation and structural property of traditional encryption schemes is that data first needs to be decrypted prior to being processed. As discussed earlier, this is not suited for machine learning applications. With traditional encryption schemes, the privacy

control lies in the hands of the recipient of the encrypted data. A fundamentally different approach is to rely on *fully homomorphic encryption* (FHE), first posed as a challenge in 1978 [26] and only solved in 2009 in a breakthrough result by Gentry [15]. In contrast to traditional encryption schemes, fully homomorphic encryption schemes allow the recipient to directly operate on encrypted data.

Controlling the noise. At the core of Gentry's result resides the technique of *bootstrapping*. All known instantiations of fully homomorphic encryption schemes produce noisy ciphertexts. Running homomorphic operations on these ciphertexts in turn increases the noise level in the resulting ciphertext. At some point, the noise present in a ciphertext may become too large and the ciphertext is no longer decryptable. A homomorphic encryption scheme supporting a predetermined noise threshold is termed *leveled*. Bootstrapping is a generic technique that allows refreshing ciphertexts. It therefore enables one to turn leveled homomorphic encryption schemes into *fully* homomorphic encryption schemes, and so to make them evaluate any possible function on ciphertexts. The key idea behind bootstrapping is to homomorphically evaluate the decryption circuit.

The works that followed Gentry's publication were aimed at proposing new schemes or at improving the bootstrapping in order to make FHE more efficient in practice. The most famous constructions are DGHV [11], BGV [5], GSW [16], and their variants. While the constructions that were successively proposed made the bootstrapping more practical, it still constituted the bottleneck (each bootstrapping taking a few minutes). A much faster bootstrapping, based on a GSW-type scheme, was later devised by Ducas and Micciancio [13], reducing the bootstrapping time to a sub second. Their technique was further improved and refined, which led to the development of the TFHE scheme [10].

Our techniques and contributions. This paper builds on the state-of-the-art TFHE scheme and extends the TFHE techniques to homomorphically evaluate deep neural networks. TFHE can operate in two modes: leveled and bootstrapped. The *leveled mode* supports linear combinations and a predetermined number of (external) products. The operations evaluated in this mode make the noise always grow. The leveled mode can be used to evaluate small-depth circuits. As for the *bootstrapped mode*, it enables a fine control of the noise by reducing it to a given level whenever it exceeds a certain threshold. Further, as will be shown, the bootstrapped mode is programmable and therefore enables the evaluation of more complex functions. For problems involving circuits of large depth, only the bootstrapped mode is applicable. Deep neural networks belong to that case.

Earlier works attempted to evaluate neural networks using fully homomorphic encryption. Cryptonets [12] was the first initiative towards this goal. They were able to perform a homomorphic inference over 5 layers against the MNIST dataset [21]. In order to limit the noise growth, the standard activation function was replaced with the square function. A number of subsequent works have adopted a similar approach and improved it in various directions. Among them, it is worth mentioning the results of the iDASH competition [17]. The winning solutions of the homomorphic encryption track of the last editions (namely, [19,2]

for 2018 and [18] for 2019) have all in common to rely on *leveled* homomorphic encryption.

Leveled solutions are however inherently limited in the type of tasks they can perform. In particular, in the case of neural networks, they can only accommodate networks with a moderate number of layers. In this paper, we want to emphasize that depth is *not* necessarily an issue and that deep neural networks can actually be evaluated homomorphically. For a specialized type of networks, this was already pointed out in a paper by Bourse *et al.* [4]; specifically, for discretized neural networks whose signals are restricted to the set $\{-1, 1\}$ and where the activation function is the sign function. Central to their scalable construction is an adaptation of the TFHE scheme so as to enable the evaluation of the sign function during a bootstrapping step. In a recent work, Boura *et al.* [3] investigated the applicability of fully homomorphic encryption for *classical* deep neural networks. They simulated the effect of noise propagation by adding a noise value drawn from a normal distribution to intermediate values, while evaluating the model *in the clear*. These experiments were carried out with models making use of the standard ReLU activation function but also with models making use of FHE-friendly variants thereof. Similar experiments were run by replacing max-pooling layers with FHE-friendly average-pooling layers. As a conclusion of their study, the authors of [3] recommend to favor FHE-friendly operations as they appear to be usually more resilient to noise perturbations.

This paper departs from previous works. We do not seek to design new operations or to modify the topology in order to make a given neural network more amenable to an FHE-based implementation. On the contrary, we stick to the original neural network model. This presents the tremendous advantage of not requiring to re-train a new model. As the operations and topology are unchanged, the already trained model can be used as is. The efforts to train a neural network should not be overlooked. This is a costly and time-consuming operation. Furthermore, in many cases, producing a new model is not even possible as this implies having access to the training dataset, which may demand the prior approval of the data owners or of some regulatory authorities.

In our approach, the evaluation of complex functions is achieved thanks to a combination of programmable bootstrapping techniques and leveled operations. Being *programmable* means that any function (including non-linear functions) of an input ciphertext can be obtained as the output of the bootstrapping. Interestingly, the resulting ciphertext features a controlled level of noise. The process can therefore be iterated over and over. So, in the case of machine learning applications, the depth of neural networks can be arbitrarily large. Our techniques are efficient and directly operate on words of a chosen size.

2 Preliminaries

2.1 Torus and Torus Polynomials

The letter 'T' in TFHE refers to the real torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, that is, the set of real numbers modulo 1. Any two elements of $\mathbb{T}$ can be added modulo 1: $(\mathbb{T}, +)$ forms

an abelian group. But $\mathbb{T}$ is not a ring as the internal product $\times$ of torus elements is not defined. The *external* product $\cdot$ between integers and torus elements is however well defined. Given $k \in \mathbb{Z}$ and $t \in \mathbb{T}$, the element $k \cdot t \in \mathbb{T}$ is defined as $k \cdot t = t + \cdots + t$ (k times) if $k \geq 0$ and $k \cdot t = (-k) \cdot (-t)$ if $k < 0$. Mathematically, $\mathbb{T}$ is endowed with a $\mathbb{Z}$-*module* structure.

Polynomials can as well be defined over the torus. As will become apparent, they allow for cryptographic operations otherwise not feasible. Let $\Phi \coloneqq \Phi(X)$ denote the M-th cyclotomic polynomial and let N denote its degree. For performance reasons, M is chosen as a power of 2, in which case it turns out that $N = M/2$ and $\Phi(X) = X^N + 1$. Consider the polynomial rings $\mathbb{R}_N[X] \coloneqq \mathbb{R}[X]/(X^N + 1)$ and $\mathbb{Z}_N[X] \coloneqq \mathbb{Z}[X]/(X^N + 1)$. This defines the $\mathbb{Z}_N[X]$-module $\mathbb{T}_N[X] \coloneqq \mathbb{R}_N[X]/\mathbb{Z}_N[X] = \mathbb{T}[X]/(X^N + 1)$. Elements of $\mathbb{T}_N[X]$ can therefore be seen as polynomials modulo $X^N + 1$ with coefficients in $\mathbb{T}$. Being a $\mathbb{Z}_N[X]$-module, elements in $\mathbb{T}_N[X]$ can be added together and (externally) multiplied by polynomials of $\mathbb{Z}_N[X]$.

Vectors are viewed as row matrices and are denoted with bold letters. Elements in $\mathbb{Z}$ or $\mathbb{T}$ are denoted with roman letters while polynomials are denoted with calligraphic letters. $\mathbb{B}$ is the integer subset $\{0, 1\}$ and, for N a power of 2, $\mathbb{B}_N[X]$ is the subset of polynomials in $\mathbb{Z}_N[X]$ with coefficients in $\mathbb{B}$. For a vector $\boldsymbol{z} \in (\mathbb{Z}/q\mathbb{Z})^n$, its norm $\|\boldsymbol{z}\|$ is defined as the shortest norm among the equivalent classes of $\boldsymbol{z} \in (\mathbb{Z}/q\mathbb{Z})^n$ in $\mathbb{Z}^n$. A polynomial in $\mathbb{Z}_N[X]$ can be identified with a vector in $\mathbb{Z}^N$: to a polynomial $\mathfrak{p} = p_0 + p_1\, X + \cdots + p_{N-1}\, X^{N-1}$ is associated the vector $(p_0, p_1, \ldots, p_{N-1})$. The norm of a polynomial is defined as the norm of its associated vector.

2.2 Probability Distributions

Two probability distributions will be used: the uniform distribution and the normal (a.k.a. Gaussian) distribution. They are respectively denoted by $\mathcal{U}$ and $\mathcal{N}$.

When the uniform distribution $\mathcal{U}$ is defined over an interval $[a, b]$, it is written $\mathcal{U}([a, b])$. Discrete intervals are indicated by double brackets; e.g., $[\![a, b]\!]$. The normal distribution is parametrized by its mean μ and its variance σ^2 and is written as $\mathcal{N}(\mu, \sigma^2)$. A normal distribution over the real numbers induces a discretized normal distribution over $\mathbb{Z}$: to a real value $X \in \mathbb{R}$ corresponds an integer value $Z = \lfloor X\, q \rceil$.

If $\mathfrak{D}$ is a distribution over a space S then $s \leftarrow \mathfrak{D}(S)$ indicates that s is chosen at random in S according to $\mathfrak{D}$; $s \xleftarrow{\$} S$ is a shorthand for $s \leftarrow \mathcal{U}(S)$.

3 Discretized TFHE

The LWE assumption over the torus (cf. Definition 1, Appendix A) essentially says that a torus element $r \in \mathbb{T}$ constructed as $r = \sum_{j=1}^{n} s_j \cdot a_j + e$ cannot be distinguished from a *random* torus element $r \in \mathbb{T}$, even if the torus vector $(a_1, \ldots, a_n)$ is given. Torus element $r = \sum_{j=1}^{n} s_j \cdot a_j + e$ can therefore be used as a random mask to conceal a "plaintext message" $\mu \in \mathbb{T}$ so as to form a

ciphertext $\boldsymbol{c} = (a_1, \ldots, a_n, r + \mu) \in \mathbb{T}^{n+1}$, where $s = (s_1, \ldots, s_n) \in \mathbb{B}^n$ plays the role of the private encryption key. The noise e is sampled from a normal error distribution $\chi = \mathcal{N}(0, \sigma^2)$. In the same way, the GLWE assumption (cf. Definition 2, Appendix A) gives rise to an encryption mechanism for polynomials of $\mathbb{T}_N[X]$: the encryption of $\mu \in \mathbb{T}_N[X]$ under key $\mathfrak{s} = (\mathfrak{s}_1, \ldots, \mathfrak{s}_k) \in \mathbb{B}_N[X]^k$ being given by $\mathfrak{c} = (\mathfrak{a}_1, \ldots, \mathfrak{a}_k, \mathfrak{r} + \mu) \in \mathbb{T}_N[X]^{k+1}$ with $\mathfrak{r} = \sum_{j=1}^{k} \mathfrak{s}_j \cdot \mathfrak{a}_j + \mathfrak{e}$. The noise $\mathfrak{e}$ is sampled from a normal error distribution $\chi = \mathcal{N}(0, \sigma^2)$ over $\mathbb{R}_N[X]$; namely, over polynomials of $\mathbb{R}_N[X]$ with coefficients drawn in $\mathcal{N}(0, \sigma^2)$.

As already pointed out in [10], LWE-based ciphertexts can be viewed as a special instance of GLWE-based ciphertexts for $(k, N) = (n, 1)$. Indeed, when $N = 1$, it turns out that $\mathbb{R}_N[X] = \mathbb{R}$, $\mathbb{Z}_N[X] = \mathbb{Z}$, and $\mathbb{T}_N[X] = \mathbb{T}$. Hence, to keep the presentation as general as possible, we will stick to the GLWE setting; LWE-based encryption being a particular case.

In the most generic setting, ciphertexts $\mathfrak{c} = (\mathfrak{a}_1, \ldots, \mathfrak{a}_k, \mathfrak{r} + \mu)$ are vectors of polynomials over $\mathbb{T}_N[X]$. These polynomials can in turn be regarded as vectors over $\mathbb{T}$. In a practical implementation, torus elements are represented with a finite precision (typically, 32 or 64 bits). Let Ω denote the bit-precision—for example, $\Omega = 32$ if the ciphertext components are represented with a precision of 32 bits. In this case, torus elements are restricted to elements of the form $\sum_{i=1}^{\Omega} t_i \cdot 2^{-i} \pmod 1$ with $t_i \in \{0, 1\}$. Essentially, the effect of working with a finite precision boils down to replacing $\mathbb{T}$ with the submodule

$$\hat{\mathbb{T}} \coloneqq q^{-1}\mathbb{Z}/\mathbb{Z} \subset \mathbb{T} \quad \text{where } q = 2^{\Omega}$$

and doing computations in $\hat{\mathbb{T}}_N[X] \coloneqq \hat{\mathbb{T}}[X]/(X^N + 1)$. Viewing $\frac{1}{q}$ as an element in $\hat{\mathbb{T}}_N[X]$, any polynomial $\mu \in \hat{\mathbb{T}}_N[X]$ can be written as

$$\mu = \bar{\mu} \cdot \tfrac{1}{q} \quad \text{for some } \bar{\mu} \in \hat{\mathbb{Z}}_N[X]$$

where $\hat{\mathbb{Z}}_N[X] \coloneqq (\mathbb{Z}/q\mathbb{Z})[X]/(X^N + 1)$.

3.1 Encoding/Decoding Messages

The input messages, prior to encryption, can be in any format. The role of the encoding/decoding process is to make them compatible with the encryption scheme.

It is useful to introduce some terminology and notation. An element $\bar{\mu} \in \hat{\mathbb{Z}}_N[X]$ entering the encryption algorithm is referred to as the *plaintext*. It matches a *cleartext* m in a certain finite message space $\mathcal{M}$. The correspondence between cleartexts and plaintexts is given by a message encoding function Encode; the reverse operation is the decoding function Decode. It is required that, for any cleartext $m \in \mathcal{M}$, the relation $\mathrm{Decode}(\mathrm{Encode}(m)) = m$ holds.

Let $\bar{\mu} = \bar{\mu}_0 + \bar{\mu}_1 X + \cdots + \bar{\mu}_{N-1} X^{N-1} \in \hat{\mathbb{Z}}_N[X]$ be a plaintext. Because ciphertexts are noisy and the noise is added to the right (i.e., less significant position), only the upper bits of $\bar{\mu}_i$ are used to encode cleartext messages. In

order to successfully complete, certain homomorphic operations demand some of the leading bits of $\bar{\mu}_i$ to be provisioned and set to 0. In the most general case, we let $\varpi \geq 0$ denote the number of these bits (called padding bits) and $\omega \geq 1$ the number of bits that are actually used to represent input cleartexts so that $\varpi + \omega \leq \Omega$. We define $p = 2^{\varpi+\omega}$. Parameter ω is referred to as the message bit-precision and parameter p as the message modulus. The coefficients $\bar{\mu}_i$ of plaintext polynomial $\bar{\mu} \in \hat{\mathbb{Z}}_N[X]$ are therefore of the form $\bar{\mu}_i = 2^{\Omega-(\varpi+\omega)}(\bar{\nu}_i \bmod 2^{\omega})$ for some $\bar{\nu}_i$, as shown in Fig. 1.

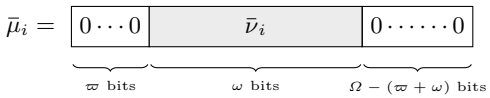


Fig. 1: Plaintext representation.

For an arbitrary element $\bar{x} = \bar{x}_H \frac{q}{p} \pm \bar{x}_L \in \mathbb{Z}/q\mathbb{Z}$ with $0 \leq \bar{x}_L \leq \frac{q}{2p}$, we define the function Upper that returns the value of $\bar{x}_H \frac{q}{p}$. It is specified as $\mathrm{Upper}\colon \mathbb{Z}/q\mathbb{Z} \to \mathbb{Z}/q\mathbb{Z}, \bar{x} \mapsto \mathrm{Upper}(\bar{x})$ with

$$\mathrm{Upper}_{q,p}(\bar{x}) = \frac{q}{p}\left\lfloor \frac{p\ \mathrm{lift}(\bar{x})}{q} \right\rceil \pmod q$$

where the function lift lifts elements of $\mathbb{Z}/q\mathbb{Z}$ to $\mathbb{Z}$. The function Upper naturally extends to polynomials of $\hat{\mathbb{Z}}_N[X]$ by applying it coefficient-wise.

In particular, for $q = 2^{\Omega}$ and $p = 2^{\varpi+\omega}$, the function reads as $\mathrm{Upper}_{q,p}(\bar{x}) = 2^{\Omega-(\varpi+\omega)}\left\lfloor \frac{2^{\varpi+\omega}\ \mathrm{lift}(\bar{x})}{2^{\Omega}} \right\rceil \pmod q$. Note that if $\bar{x} = \bar{x}_H\, 2^{\Omega-(\varpi+\omega)} \pm \bar{x}_L \in \mathbb{Z}/q\mathbb{Z}$ with $0 \leq \bar{x}_L \leq 2^{\Omega-(\varpi+\omega)-1}$ then $\mathrm{Upper}_{q,p}(\bar{x}) = \bar{x}_H\, 2^{\Omega-(\varpi+\omega)}$.

3.2 Description

We are now ready to present the implementation of TFHE encryption with a finite representation precision. We write $\overline{\mathrm{GLWE}}$ the corresponding encryption algorithm in the general case. The $\overline{\mathrm{LWE}}$ encryption algorithm coincides with the particular case $(k, N) = (n, 1)$.

KeyGen(1^{λ}) On input security parameter λ, define a pair of integers (k, N) with $k \geq 1$ and N a power of 2. Define also a normal error distribution $\chi = \mathcal{N}(0, \sigma^2)$ over $\mathbb{R}_N[X]$. Sample uniformly at random a vector $\mathfrak{z} = (\mathfrak{z}_1, \ldots, \mathfrak{z}_k) \xleftarrow{\$} \mathbb{B}_N[X]^k$. The plaintext space is $\mathcal{P}_N[X] = (\frac{q}{p}\mathbb{Z}/q\mathbb{Z})[X]/(X^N + 1) \subset \hat{\mathbb{Z}}_N[X] = (\mathbb{Z}/q\mathbb{Z})[X]/(X^N + 1)$ where $q = 2^{\Omega}$ and the message modulus is $p = 2^{\varpi+\omega}$, with $\Omega \geq \varpi + \omega$.
The public parameters are $\mathsf{pp} = \{k, N, \sigma, p, q\}$ and the private key is $\mathsf{sk} = \mathfrak{z}$.

$\mathsf{Encrypt}_{\mathsf{sk}}(\bar{\mu})$ The encryption of a plaintext $\bar{\mu} \in \mathcal{P}_N[X]$ is given by

$$\bar{\mathbf{c}} \leftarrow \overline{\mathrm{GLWE}}_{\mathfrak{z}}(\bar{\mu}) \coloneqq (\bar{a}_1, \ldots, \bar{a}_k, \bar{b}) \in \hat{\mathbb{Z}}_N[X]^{k+1}$$

with

$$\begin{cases} \bar{\mu}^* = \bar{\mu} + \bar{e} \pmod{(q, X^N+1)} \\ \bar{b} = \sum_{j=1}^{k} \mathfrak{z}_j \, \bar{a}_j + \bar{\mu}^* \pmod{(q, X^N+1)} \end{cases}$$

for a random polynomial vector $(\bar{a}_1, \ldots, \bar{a}_k) \xleftarrow{\$} \hat{\mathbb{Z}}_N[X]^k$ and a discrete noise $\bar{e} = \lfloor e\, q \rceil$ for some $e \leftarrow \mathbb{R}_N[X]$ whose coefficients are sampled in $\mathcal{N}(0, \sigma^2)$.

$\mathsf{Decrypt}_{\mathsf{sk}}(\bar{\mathbf{c}})$ To decrypt $\bar{\mathbf{c}} = (\bar{a}_1, \ldots, \bar{a}_k, \bar{b})$, use private key $\mathfrak{z} = (\mathfrak{z}_1, \ldots, \mathfrak{z}_k)$, compute in $\hat{\mathbb{Z}}_N[X]$

$$\bar{\mu}^* = \bar{b} - \sum_{j=1}^{k} \mathfrak{z}_j \, \bar{a}_j \pmod{(q, X^N+1)},$$

and output $\mathrm{Upper}_{q,p}(\bar{\mu}^*)$.

Correctness. Let $p = 2^{\varpi+\omega}$, $q = 2^{\Omega}$ and $\bar{\mu} = \bar{\mu}_0 + \cdots + \bar{\mu}_{N-1} X^{N-1}$ with $\bar{\mu}_i = \frac{q}{p} (\bar{\nu}_i \bmod 2^{\omega})$ for some $\bar{\nu}_i$; see Section 3.1. It can be verified that if $\bar{\mathbf{c}} \leftarrow \mathsf{Encrypt}_{\mathfrak{z}}(\bar{\mu})$ then $\mathsf{Decrypt}_{\mathfrak{z}}(\bar{\mathbf{c}}) = \bar{\mu}$, provided that $\|\bar{e}\|_\infty < \frac{q}{2p} = 2^{\Omega-(-\varpi+\omega)-1}$. Note that the same holds true even if some of the ϖ leading bits of the $\bar{\mu}_i$'s are non-zero.

In certain applications, it is acceptable that the decryption algorithm does not recover the *exact* initial plaintext but a close approximation thereof. In this case, the requirement becomes $\mathsf{Decrypt}_{\mathfrak{z}}(\bar{\mathbf{c}}) \approx \bar{\mu}$ and the condition on the bound of $\|\bar{e}\|_\infty$ can be relaxed.

3.3 Leveled Operations

FHE enables directly performing operations on ciphertexts. Depending on the type of operation, the resulting noise level increases more or less.

Addition. Clearly, $\overline{\mathrm{GLWE}}$ ciphertexts are homomorphic with respect to the addition. Let $\bar{\mathbf{c}}_1 \leftarrow \overline{\mathrm{GLWE}}_{\mathfrak{z}}(\bar{\mu}_1)$ and $\bar{\mathbf{c}}_2 \leftarrow \overline{\mathrm{GLWE}}_{\mathfrak{z}}(\bar{\mu}_2)$ with $\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2 \in \hat{\mathbb{Z}}_N[X]^{k+1}$ be the respective encryptions of plaintexts $\bar{\mu}_1, \bar{\mu}_2 \in \mathcal{P}_N[X]$; i.e.,

$$\bar{\mathbf{c}}_i = \left(\bar{a}_1^{(i)}, \ldots, \bar{a}_k^{(i)}, \bar{b}_i\right) \qquad (i \in \{1, 2\})$$

with $\bar{b}_i = \sum_{j=1}^{k} \mathfrak{z}_j \, \bar{a}_j^{(i)} + \bar{\mu}_i + \bar{e}_i$. Then $\bar{\mathbf{c}}_3 = \bar{\mathbf{c}}_1 + \bar{\mathbf{c}}_2 = \left(\bar{a}_1^{(1)} + \bar{a}_1^{(2)}, \ldots, \bar{a}_k^{(1)} + \bar{a}_k^{(2)}, \bar{b}_1 + \bar{b}_2\right)$ is a $\overline{\mathrm{GLWE}}$ encryption of $(\bar{\mu}_1 + \bar{\mu}_2) \in \mathcal{P}_N[X]$, provided that the resulting noise $\bar{e}_3 = \bar{e}_1 + \bar{e}_2$ keeps small.

Scalar multiplication. By extension, $\overline{\mathrm{GLWE}}$ ciphertexts are homomorphic with respect to the multiplication by a constant. Let $K \in \mathbb{Z}_{\geq 0}$ and $\bar{\mathbf{c}} \leftarrow \overline{\mathrm{GLWE}}_{\mathfrak{z}}(\bar{\mu}) = (\bar{a}_1, \ldots, \bar{a}_k, \bar{b})$ with $\bar{b} = \sum_{j=1}^{k} \mathfrak{z}_j \, \bar{a}_j + \bar{\mu} + \bar{e}$. Then, for $K \geq 0$, $K \cdot \bar{\mathbf{c}} = \bar{\mathbf{c}} + \cdots + \bar{\mathbf{c}}$ (K times) is an encryption of $K \cdot \bar{\mu} \in \mathcal{P}_N[X]$, provided that $K \cdot \bar{e}$ keeps small. If $K < 0$ then $K \cdot \bar{\mathbf{c}} = (-K) \cdot (-\bar{\mathbf{c}})$. More generally, if $\mathcal{K} \in \mathbb{Z}_N[X]$ then $\mathcal{K} \cdot \bar{\mathbf{c}}$ is an encryption of $\mathcal{K} \cdot \bar{\mu} \in \mathcal{P}_N[X]$, provided that $\mathcal{K} \cdot \bar{e}$ keeps small.

External product. $\overline{\text{GLWE}}$ ciphertexts do not support a native internal multiplication which, in practice, means that two $\overline{\text{GLWE}}$ ciphertexts cannot be directly multiplied. In order to perform a multiplication, a clever matrix-based approach put forward in the GSW construction [16] can be used. By analogy, we write $\overline{\text{GGSW}}$ the corresponding encryption algorithm in the general case; the particular case $(k, N) = (n, 1)$ is denoted by $\overline{\text{GSW}}$.

With the previous $\overline{\text{GLWE}}$ notation, let parameters $B = 2^\beta$ and ℓ with $\beta, \ell \geq 1$ and such that $\ell\beta \leq \Omega$. Define also the vector $\boldsymbol{g} = (2^{\Omega-\beta}, 2^{\Omega-2\beta}, \ldots, 2^{\Omega-\ell\beta})$. The $\overline{\text{GGSW}}$ encryption of a plaintext $m \in \mathbb{Z}_N[X]$ with respect to a $\overline{\text{GLWE}}$ encryption key $\mathfrak{z} \in \mathbb{B}_N[X]^k$ is defined as

$$\overline{\mathbf{C}} \leftarrow \overline{\text{GGSW}}_{\mathfrak{z}}(m) \coloneqq \overline{\mathbf{Z}} + m \cdot \mathbf{G}^\intercal \in \hat{\mathbb{Z}}_N[X]^{(k+1)\ell \times (k+1)}$$

where

$$\overline{\mathbf{Z}} \leftarrow \left.\begin{pmatrix} \overline{\text{GLWE}}_{\mathfrak{z}}(0) \\ \vdots \\ \overline{\text{GLWE}}_{\mathfrak{z}}(0) \end{pmatrix}\right\} (k+1)\ell \text{ rows}$$

is a matrix containing on each row a fresh $\overline{\text{GLWE}}$ encryption of 0 and where

$$\mathbf{G}^\intercal = \mathbf{I}_{k+1} \otimes \boldsymbol{g}^\intercal = \text{diag}(\underbrace{\boldsymbol{g}^\intercal, \ldots, \boldsymbol{g}^\intercal}_{k+1}) \in \mathbb{Z}_N[X]^{(k+1)\ell \times (k+1)}$$

is the so-called *gadget matrix* [23], with $\mathbf{I}_{k+1}$ the identity matrix of size $k+1$. It is worth noting that any element d in $\mathbb{Z}/q\mathbb{Z}$, viewed as an integer in $[\![-\frac{q}{2}, \frac{q}{2}[\![$, can always be approximated by a signed-digit radix-B expansion of size ℓ as

$$d \approx q \sum_{i=1}^{\ell} d_i \, B^{-i} = \sum_{i=1}^{\ell} d_i \, 2^{\Omega - i\beta} = \boldsymbol{g}^{-1}(d)\, \boldsymbol{g}^\intercal$$

where $\boldsymbol{g}^{-1}(d) \coloneqq (d_1, \ldots, d_\ell) \in \mathbb{Z}^\ell$ with digits $d_i \in [\![-\frac{B}{2}, \frac{B}{2}[\![$ and with an approximation error that is bounded by $\left|\boldsymbol{g}^{-1}(d)\, \boldsymbol{g}^\intercal - d\right| \leq q/(2B^\ell) = 2^{\Omega-\beta\ell-1}$.

By extension, for a polynomial $\mathfrak{p} = p_0 + \cdots + p_{N-1} X^{N-1} \in \hat{\mathbb{Z}}_N[X]$ whose coefficients are viewed as integers in $[\![-\frac{q}{2}, \frac{q}{2}[\![$, the decomposition $\boldsymbol{g}^{-1}(\mathfrak{p}) \in \mathbb{Z}_N[X]^\ell$ is defined as $\boldsymbol{g}^{-1}(\mathfrak{p}) = \sum_{j=0}^{N-1} \boldsymbol{g}^{-1}(p_j)\, X^j$. Clearly, it holds that $\left\|\boldsymbol{g}^{-1}(\mathfrak{p})\, \boldsymbol{g}^\intercal - \mathfrak{p}\right\|_\infty \leq 2^{\Omega-\beta\ell-1}$. Finally, for a vector of $k+1$ polynomials, $\boldsymbol{\mathfrak{p}} = (\mathfrak{p}_1, \ldots, \mathfrak{p}_{k+1}) \in \hat{\mathbb{Z}}_N[X]^{k+1}$, the decomposition $\mathbf{G}^{-1}(\boldsymbol{\mathfrak{p}}) \in \mathbb{Z}_N[X]^{(k+1)\ell}$ is defined as $\mathbf{G}^{-1}(\boldsymbol{\mathfrak{p}}) = (\boldsymbol{g}^{-1}(\mathfrak{p}_1), \ldots, \boldsymbol{g}^{-1}(\mathfrak{p}_{k+1}))$ and $\left\|\mathbf{G}^{-1}(\boldsymbol{\mathfrak{p}})\, \mathbf{G}^\intercal - \boldsymbol{\mathfrak{p}}\right\|_\infty \leq 2^{\Omega-\beta\ell-1}$.

Interestingly, the gadget decomposition of $\overline{\text{GLWE}}$ ciphertexts gives rise to an *external product* [10] with $\overline{\text{GGSW}}$ ciphertexts. Specifically, for plaintexts $m_1 \in \mathbb{Z}_N[X]$ and $\mu_2 \in \mathcal{P}_N[X]$, if $\overline{\mathbf{C}_1} \leftarrow \overline{\text{GGSW}}_{\mathfrak{z}}(m_1)$ and $\bar{\mathbf{c}}_2 \leftarrow \overline{\text{GLWE}}(\mu_2)$ then their external product, denoted by $\boxdot$, is given by

$$\bar{\mathbf{c}}_3 = \overline{\mathbf{C}_1} \boxdot \bar{\mathbf{c}}_2 \coloneqq \mathbf{G}^{-1}(\bar{\mathbf{c}}_2)\, \overline{\mathbf{C}_1} \quad .$$

A little algebra shows that

$$\bar{\mathbf{c}}_3 = \mathbf{G}^{-1}(\bar{\mathbf{c}}_2)\,(\overline{\mathfrak{Z}} + m_1 \cdot \mathbf{G}^{\intercal}) = \underbrace{\mathbf{G}^{-1}(\bar{\mathbf{c}}_2)\,\overline{\mathfrak{Z}}}_{=\overline{\mathrm{GLWE}}_{\mathfrak{z}}(0)} + \underbrace{m_1 \cdot \mathbf{G}^{-1}(\bar{\mathbf{c}}_2)\,\mathbf{G}^{\intercal}}_{\approx m_1\,\bar{\mathbf{c}}_2}$$

is a $\overline{\mathrm{GLWE}}$ encryption of $m_1\,\mu_2 \in \mathcal{P}_N[X]$, provided that the resulting noise (including the approximation error) keeps small.

The CMux gate. Starting from the external product, a new leveled operation can be defined: the 'controlled' multiplexer or CMux [10, § 3.4]. A CMux acts as a selector according to a bit—but over encrypted data. It takes as input two $\overline{\mathrm{GLWE}}$ ciphertexts $\bar{\mathbf{c}}_0$ and $\bar{\mathbf{c}}_1$, respectively encrypting plaintexts μ_0 and $\mu_1 \in \mathcal{P}_N[X]$, and a $\overline{\mathrm{GGSW}}$ ciphertext $\overline{\mathcal{C}}$ encrypting a bit b. The result is a $\overline{\mathrm{GLWE}}$ ciphertext $\bar{\mathbf{c}}'$ encrypting μ_b, provided that the resulting noise keeps small. The CMux gate is given by:

$$\bar{\mathbf{c}}' \leftarrow \mathrm{CMux}(\overline{\mathcal{C}}, \bar{\mathbf{c}}_0, \bar{\mathbf{c}}_1) \coloneqq \overline{\mathcal{C}} \boxdot (\bar{\mathbf{c}}_1 - \bar{\mathbf{c}}_0) + \bar{\mathbf{c}}_0 \ .$$

It plays a central role in the performance of homomorphic computations and, especially, inside the bootstrapping.

4 Programmable Bootstrapping

The *programmable bootstrapping* is an extension of the bootstrapping technique that allows resetting the noise to a fixed level while—at the same time—evaluating a function on the input ciphertext.

In this section, we first explain in detail how to perform the regular bootstrapping. We then proceed with the programmable bootstrapping and show how to evaluate any function expressed as a look-up table. When f is the identity function, that coincides with a regular bootstrapping.

4.1 Blind Rotation

As aforementioned, Gentry's bootstrapping boils down to homomorphically decrypt a ciphertext using a homomorphic encryption of its own decryption key, with the goal of reducing the noise the ciphertext contains.

Intuition. Consider an $\overline{\mathrm{LWE}}$ ciphertext $\bar{\boldsymbol{c}} \leftarrow \overline{\mathrm{LWE}}_{\boldsymbol{s}}(\bar{\mu}) = (\bar{a}_1, \ldots, \bar{a}_n, \bar{b}) \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$ where $\bar{a}_j \xleftarrow{\$} \mathbb{Z}/q\mathbb{Z}$ and $\bar{b} = \sum_{j=1}^{n} s_j\,\bar{a}_j + \bar{\mu}^*$ with $\bar{\mu}^* = \bar{\mu} + \bar{e}$ for some discrete noise $\bar{e} = \lfloor e\,q \rceil$ with $e \leftarrow \mathcal{N}(0, \sigma^2)$. Ciphertext $\bar{\boldsymbol{c}}$ is an encryption of plaintext $\bar{\mu} \in \mathcal{P} = \frac{q}{p}\mathbb{Z}/q\mathbb{Z}$ under the secret key $\boldsymbol{s} = (s_1, \ldots, s_n) \in \mathbb{B}^n$. It can be decrypted using $\boldsymbol{s}$ in two steps as $\bar{\mu}^* \leftarrow \bar{b} - \sum_{j=1}^{n} s_j\,\bar{a}_j$ and $\bar{\mu} \leftarrow \mathrm{Upper}_{q,p}(\bar{\mu}^*)$.

In order to bootstrap, one way to look at the decryption (without the rounding) is to see that

$$-\bar{\mu}^* = -\bar{b} + \textstyle\sum_{j=1}^{n} s_j\,\bar{a}_j \pmod q$$

and to put this value at the exponent of X to get the monomial $X^{-\bar{\mu}^*}$. Note that there are q possible values for $\bar{\mu}^*$. The rough idea (more technical details are given later) is then to build a polynomial—that we call *test polynomial*—such that each one of its coefficients encodes the noise-free value corresponding to $\bar{\mu}^*$ (namely, $\bar{\mu} = \mathrm{Upper}_{q,p}(\bar{\mu}^*)$), for all the possible $\bar{\mu}^*$'s. Specifically, if we suppose for a moment that the test polynomial is the degree-q polynomial $\bar{\mathfrak{v}} = \bar{v}_0 + \bar{v}_1\, X + \cdots + \bar{v}_{q-1}\, X^{q-1}$ then its i^{th} coefficient is set to $\bar{v}_i = \mathrm{Upper}_{q,p}(i \bmod q)$. By rotating the test polynomial of $\bar{\mu}^*$ positions, the value of $\bar{\mu}$ moves to the constant coefficient position. This is illustrated in Fig. 2. It then remains to extract it.

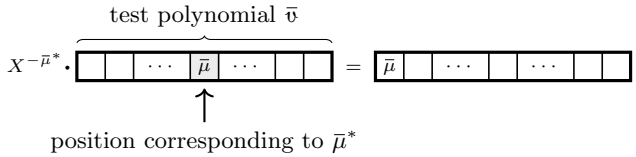


Fig. 2: Rotating the test polynomial.

Of course, this rotation is done homomorphically (hence the name blind rotation) and since $X^{-\bar{\mu}^*} \cdot \bar{\mathfrak{v}}$ is a polynomial, this is where $\overline{\mathrm{GLWE}}$ encryption comes into play.

Polynomials used in $\overline{\mathrm{GLWE}}$ are defined modulo $X^N + 1$. This means that X as a multiplicative element of $\mathbb{Z}_N[X]$ is of order $2N$. However, as appearing in the $\overline{\mathrm{LWE}}$ encryption, $\bar{\mu}^*$ is defined modulo q. It therefore needs to be rescaled modulo $2N$. As a consequence, instead of using the relation $-\bar{\mu}^* = -\bar{b} + \sum_{j=1}^{n} s_j\, \bar{a}_j \pmod q$, we have to rely on the approximation

$$-\tilde{\mu}^* = -\tilde{b} + \textstyle\sum_{j=1}^{n} s_j\, \tilde{a}_j \pmod{2N}\,,$$

where $\tilde{b} = \left\lfloor \frac{2N(\bar{b} \bmod q)}{q} \right\rceil$ and $\tilde{a}_j = \left\lfloor \frac{2N(\bar{a}_j \bmod q)}{q} \right\rceil$. This approximation may generate a small additional error that adds to the noise. We call this additional error *drift*. It depends on both the size n of the input $\overline{\mathrm{LWE}}$ and the ring size N used in $\overline{\mathrm{GLWE}}$ during the rotation. The impact of the drift on the result can be dealt with by a careful choice of the parameters. In particular, a smaller value for n or a larger value for N is expected to decrease the resulting drift.

Also, because the test polynomial $\bar{\mathfrak{v}}$ lies in $\hat{\mathbb{Z}}_N[X]$ and thus has N coefficients, at most N values for $\tilde{\mu}^*$ can be encoded. This is addressed by ensuring that the most significant bit of $\tilde{\mu}^*$ is set to 0; that is, parameter $\varpi \geq 1$ (see Section 3.1). In this case, $\tilde{\mu}^*$ can take at most N possible values and the test polynomial is formed as $\bar{\mathfrak{v}} = \bar{v}_0 + \cdots + \bar{v}_{N-1}\, X^{N-1}$ with

$$\bar{v}_i = \mathrm{Upper}_{q,p}\big(\tfrac{q}{2N}\, i \bmod q\big)\ .$$

Implementation. It remains to explain how to compute the product between $X^{-\tilde{\mu}^*}$ and the test polynomial $\bar{\mathfrak{v}}$ under $\overline{\mathrm{GLWE}}$ encryption. It turns out that such a computation can be computed as a succession of CMux gates [10, § 4.3].

An accumulator ACC is initialized with a $\overline{\mathrm{GLWE}}$ encryption of $X^{-\tilde{b}} \cdot \bar{\mathfrak{v}}$. It is then updated in a for-loop where, at iteration i (for $1 \leq i \leq n$) it is multiplied by $X^{s_i\, \tilde{a}_i}$. The multiplication is performed thanks to the CMux operation, $\mathrm{ACC} \leftarrow \mathrm{CMux}(\mathsf{bsk}[i], \mathrm{ACC}, X^{\tilde{a}_i} \cdot \mathrm{ACC})$. Here, $\mathsf{bsk}[i] \leftarrow \overline{\mathrm{GGSW}}_{\boldsymbol{\mathfrak{z}}'}(s_i)$ (for $i = 1, \ldots, n$), are the *bootstrapping keys*; i.e., a list of $\overline{\mathrm{GGSW}}$ encryptions of the elements of the secret key $\boldsymbol{s}$ under some encryption key $\boldsymbol{\mathfrak{z}}' \in \mathbb{B}_N[X]^k$. The output of the blind rotation is $\mathrm{ACC} \leftarrow \overline{\mathrm{GLWE}}_{\boldsymbol{\mathfrak{z}}'}(X^{-\tilde{\mu}^*} \cdot \bar{\mathfrak{v}})$. More details about this procedure are provided in Appendix B.1.

Sample extraction. The remaining step of the bootstrapping consists in extracting the constant coefficient of $\bar{\mathfrak{u}} \coloneqq X^{-\tilde{\mu}^*} \cdot \bar{\mathfrak{v}}$ as a $\overline{\mathrm{LWE}}$ ciphertext of $\bar{\mu}$. This is an easy operation—called *sample extraction*—which is performed by simply extracting some of the coefficients of the $\overline{\mathrm{GLWE}}$ ciphertext. See Appendix B.2.

Key switching. The loop is almost closed. With the above procedure, input ciphertext $\bar{\boldsymbol{c}} \leftarrow \overline{\mathrm{LWE}}_{\boldsymbol{s}}(\bar{\mu}) \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$ and resulting output ciphertext $\bar{\boldsymbol{c}}' \leftarrow \overline{\mathrm{LWE}}_{\boldsymbol{s}'}(\bar{\mu}) \in (\mathbb{Z}/q\mathbb{Z})^{kN+1}$ both encrypt plaintext $\bar{\mu}$ but make use of different keys and have a different format.

In order to convert $\bar{\boldsymbol{c}}'$ back to the original setting, an operation called *key switching* can be performed. The key-switching technique is classical in FHE. See Appendix B.3 for a detailed description.

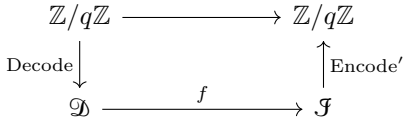
4.2 Look-up Table Evaluation

In the previous section, the blind rotation is used to perform a bootstrapping. Surprisingly, the same technique can be adapted so as to evaluate a function at the same time. The function is evaluated as a look-up table that is encoded in the test polynomial.

Specifically, suppose we intend to evaluate—over encrypted data—an arbitrary function f with domain $\mathfrak{D}$ and image $\mathfrak{F}$, $f\colon \mathfrak{D} \to \mathfrak{F}, x \mapsto y = f(x)$. We assume we are given the encoding functions $\mathrm{Encode}\colon \mathfrak{D} \to \mathbb{Z}/q\mathbb{Z}$ and $\mathrm{Encode}'\colon \mathfrak{F} \to \mathbb{Z}/q\mathbb{Z}$, and the matching decoding functions Decode and Decode′, as specified in Section 3.1.

We showed in the previous section that selecting for the test polynomial $\bar{\mathfrak{v}} = \bar{v}_0 + \cdots + \bar{v}_{N-1}\, X^{N-1}$ with $\bar{v}_i = \mathrm{Upper}_{q,p}\big(\frac{q}{2N}\, i \bmod q\big)$ transforms a ciphertext of $\bar{\mu}$ into another ciphertext of $\bar{\mu}$ with a lesser noise.

Consider now the following diagram



and suppose we define a look-up table as pairs $(i, T[i])$ for $0 \leq i \leq N-1$ with

$$T[i] = \text{Encode}' \circ f \circ \text{Decode} \circ \text{Upper}_{q,p}\left(\tfrac{q}{2N}\, i \bmod q\right) \ .$$

As the diagram suggests, we program the test polynomial as

$$\bar{\mathfrak{v}} = \bar{v}_0 + \cdots + \bar{v}_{N-1}\, X^{N-1} \quad \text{with } \bar{v}_i = T[i]\,,$$

the rest of the process described in Section 4.1 remains unchanged. Doing so, up to the drift, an input ciphertext of $\bar{\mu}$ (encoding some value $x \in \mathbb{D}$) will be transformed into a ciphertext of a value encoding $f(x)$. Furthermore, being the output of a bootstrapping, the resulting ciphertext enjoys a low level of noise. The whole process is what we call *programmable bootstrapping*.

Remark 1. As explained, the regular bootstrapping requires an encoding parameter $\varpi \geq 1$ (cf. Section 4.1). This condition can be lifted in the programmable bootstrapping when the entries of the look-up table are negacyclic; i.e., when $T[i+N] \equiv -T[i] \pmod{2N}$. In that case, $2N$ values are actually programmed.

5 Application to Neural Networks

All the tools that we need are now on the table. In this section, we apply them in order to evaluate homomorphically neural networks.

Neural networks (NN) were originally built in computer science by analogy to the human brain in order to solve complex problems that machines were not able to solve before. The neural networks can be trained and then used to classify objects, detect diseases, do face recognition, and so on.

The different layers in a neural network are typically aimed at successively extracting discriminating features or patterns from the input data. The number of layers and the type of operations that is performed in each layer depend on the task the neural network is trying to achieve.

We review below a number of layers that are commonly used to build neural networks. The list is non-exhaustive. Our techniques are generic and support all known types of layers. Each layer receives inputs from the previous layer, performs some computations, and produces outputs. The outputs then flow to the next layer as inputs. Two types of layers are distinguished when working over encrypted data:

- layers that can be evaluated homomorphically using leveled operations; and
- layers involving non-linear or more complex operations, in which case one or several programmable bootstrappings (PBS) are required.

We note that the first type of layers may also resort to bootstrap operations on some intermediate values whenever the noise exceeds a certain threshold.

5.1 Layers without PBS

Dense/linear layer. A (fully connected) dense layer computes the dot product between the inputs and a matrix of weights. A bias vector can be added. An activation function is then applied component-wise to produce the outputs. When there is no activation function, a dense layer is also called linear layer.

When evaluated homomorphically, the weights and the bias vector are provided in the clear. The evaluation of a dense layer (the activation excepted) thus consists of a series of multiplications by constants and additions, which are all leveled operations.

The activation functions are treated in the next section (see *activation layer*).

Convolution layer. A convolution layer convolves the input layer with a convolution kernel (a.k.a. filter) that is composed of a tensor of weights so as to produce a tensor of outputs. Biases can be added to the outputs. Moreover, an activation function can be applied to the outputs.

The filters are provided in the clear. Hence, as for the dense layer, the homomorphic evaluation of a convolution layer (without activation) consists of a series of multiplications by constants and additions.

Addition layer. An addition layer performs component-wise additions. Over encrypted data, those are leveled operations.

Flatten layer. A flatten layer reshapes its input into a lower-dimensional array so that it can for example be fed into a subsequent dense layer.

Over encrypted data, the flattening function simply consists in rearranging the input ciphertexts. No homomorphic operation is required.

Global average pooling layer. A global average pooling layer computes the average of the components of its inputs. Specifically, if n denotes the number of components and a_i denotes the value of component i, the global average pooling function computes $\left(\sum_{i=1}^{n} a_i\right)/n$.

In a homomorphic evaluation, the global average pooling can be reduced to the computation of the sum $\sum_{i=1}^{n} a_i$. The division by n is then performed in the next programmable bootstrapping—e.g., in a dense layer or a convolution layer—by dividing the weights by the same quantity. Hence, the sole homomorphic operation required to evaluate a global average pooling layer is the addition of ciphertexts, which is a leveled operation.

5.2 Layers with PBS

Activation layer: ReLU. An activation layer is used to inject non-linearity in the neural networks. It is crucial in the learning. There are many activation functions that can be used in an activation layer. One of the most popular activation functions is the Rectified Linear Unit (ReLU) function. Other commonly used activations include the sigmoid function or the hyperbolic tangent function.

As detailed in Section 4, the homomorphic evaluation of an activation function (as any function) can be performed via a programmable bootstrapping (PBS), with the outputs of the function encoded inside the test polynomial.

Max-pooling layer. A max-pooling layer extracts a fixed-size subset of components from the inputs and computes their maximum.

At first sight, as the max function is multivariate (i.e., it takes multiple arguments on input), it is unclear how it can be evaluated homomorphically. With two arguments, the max function can however be expressed using the [univariate] ReLU function, $\max(x, y) = y + \mathrm{ReLU}(x - y)$. Hence, from $\mathrm{Encrypt}(x)$ and $\mathrm{Encrypt}(y)$, their maximum can be evaluated as $\mathrm{Encrypt}(\max(x, y)) = \mathrm{Encrypt}(y) + \mathrm{Encrypt}(\mathrm{ReLU}(z))$ with $\mathrm{Encrypt}(z) = \mathrm{Encrypt}(x) - \mathrm{Encrypt}(y)$. This requires a couple of addition/subtraction of ciphertexts plus the homomorphic evaluation of a ReLU function, the cost of which is one PBS. When there are more than two arguments, the basic relation $\max(x_1, \ldots, x_{k-1}, x_k) = \max(y_k, x_k)$ with $y_k = \max(x_1, \ldots, x_{k-1})$ can be used. For a series of k components $(x_1, \ldots, x_k)$, the homomorphic evaluation of the max-pooling function thus amounts to $(k - 1)$ PBS.

6 Experimental Results & Benchmarks

We conducted a series of numerical experiments to assess the performance. We report below results against the MNIST dataset [21], which contains 28×28 images of handwritten digits. For testing purposes, we designed depth-20, 50, 100 neural networks, respectively noted NN-20, NN-50 and NN-100. These networks all include dense and convolution layers with activation functions; every hidden layer possesses at least 92 active neurons.

Parameter sets. The overall targeted security levels are 80 bits and 128 bits. The selected cryptographic parameters are defined by (k, N, σ) for $\overline{\mathrm{GLWE}}$ encryption and (n, σ) for $\overline{\mathrm{LWE}}$ encryption. The word-size is $\Omega = 64$ bits.

Table 1: Cryptographic parameters.

Security level		$\overline{\mathrm{GLWE}}$			$\overline{\mathrm{LWE}}$	
		k	N	σ	n	σ
80 bits	(I)	1	1024	2^{-40}	542	2^{-21}
	(II)	1	2048	2^{-60}	592	2^{-23}
128 bits	(III)	1	4096	2^{-62}	938	2^{-23}

The different parameter sets I, II & III meet at least the claimed security level and were validated using the `lwe-estimator` (https://bitbucket.org/

malb/lwe-estimator/) [1]. They can be used for the homomorphic inference of networks requiring a maximal precision of 8 bits up to 12 bits.

Performance analysis. Experiments were performed on three different types of machines, respectively referred to as PC, AWS, and AWS2:

- a personal computer with 2.6 GHz 6-Core Intel® Core™ i7 processor,
- a 3.00 GHz Intel® Xeon® Platinum 8275CL processor with 96 vCPUs hosted on AWS, and
- as above but with 8 NVIDIA® A100 Tensor Core GPUs.

Table 2: Performance comparison (computed from 1000 runs).

(a) Results in the clear.

	Run-time		Accuracy
	PC	AWS	
NN-20	0.17 ms	0.19 ms	97.5 %
NN-50	0.20 ms	0.30 ms	95.4 %
NN-100	0.33 ms	0.46 ms	95.2 %

(b) Results over encrypted data.

		Run-time			Accuracy
		PC	AWS	AWS2	
80-*bit security*:					
NN-20	(I)	12.49 s	2.85 s	0.69 s	97.2 %
	(II)	30.04 s	6.19 s	2.10 s	**97.5 %**
NN-50	(I)	26.71 s	5.90 s	1.73 s	93.4 %
	(II)	71.71 s	13.00 s	5.27 s	**95.1 %**
NN-100	(I)	46.61 s	11.18 s	3.46 s	87.3 %
	(II)	108.73 s	24.13 s	10.24 s	**91.1 %**
128-*bit security*:					
NN-20	(III)	115.52 s	21.17 s	7.53 s	**97.1 %**
NN-50	(III)	233.55 s	43.91 s	18.89 s	**94.7 %**
NN-100	(III)	481.61 s	81.47 s	37.65 s	**83.0 %**

For reference, Table 2a lists the run-time and accuracy for an unencrypted inference. They were measured using `ONNX Runtime` [24]. The run-time and accuracy over encrypted data for different settings are presented in Table 2b.

These clearly indicate the importance of the parameter choice and the different trade-offs that can be obtained. In particular, for a given security level, a larger value for parameter N increases the accuracy (at the expense of more processing). It is important to note that the given times correspond to the evaluation of a single inference run independently; in particular, the times are not amortized over a batch of inferences.

7 Conclusion

We presented a general framework for the evaluation of deep neural networks using fully homomorphic encryption. Our approach scales efficiently with the number of layers while providing good accuracy results. To do so we employ a versatile combination of encoding methods and of programmable bootstrapping techniques. To the best of our knowledge, our results set new records in the homomorphic inference of deep neural networks.

The practicality of our framework invites further works. First, it would be interesting to know the impact of the use of specialized hardware in our framework. We believe that several orders of magnitude in the processing times could be gained in that way. Another interesting work would be to investigate how to extend our techniques to the homomorphic training of neural networks or, more generally, to other intensive machine-learning tasks.

Availability. The library implementing our extended version of TFHE has been developed in Rust. It is available as an open-source project on GitHub at URL https://github.com/zama-ai/concrete.

Acknowledgments. We are grateful to our colleagues at Zama for their help and support in running the experiments.

References

1. Albrecht, M.R., Player, R., Scott, S.: On the concrete hardness of learning with errors. Journal of Mathematical Cryptology **9**(3), 169–203 (2015)
2. Blatt, M., Gusev, A., Polyakov, Y., Goldwasser, S.: Secure large-scale genome-wide association studies using homomorphic encryption. Cryptology ePrint Archive, Report 2020/563 (2020)
3. Boura, C., Gama, N., Georgieva, M., Jetchev, D.: Simulating homomorphic evaluation of deep learning predictions. In: Cyber Security Cryptography and Machine Learning (CSCML 2019). Lecture Notes in Computer Science, vol. 11527, pp. 212–230. Springer (2019)
4. Bourse, F., Minelli, M., Minihold, M., Paillier, P.: Fast homomorphic evaluation of deep discretized neural networks. In: Advances in Cryptology – CRYPTO 2018, Part III. Lecture Notes in Computer Science, vol. 10993, pp. 483–512. Springer (2018)

5. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. ACM Transactions Computation Theory **6**(3), 13:1–13:36 (2014), earlier version in ITCS 2012
6. Brakerski, Z., Langlois, A., Peikert, C., Regev, O., Stehlé, D.: Classical hardness of learning with errors. In: 45th Annual ACM Symposium on Theory of Computing. pp. 575–584. ACM Press (2013)
7. Brakerski, Z., Vaikuntanathan, V.: Efficient fully homomorphic encryption from (standard) LWE. SIAM Journal on Computing **43**(2), 831–871 (2014), earlier version in FOCS 2011
8. California Consumer Privacy Act (CCPA). https://www.oag.ca.gov/privacy/ccpa
9. Cheon, J.H., Stehlé, D.: Fully homomophic encryption over the integers revisited. In: Advances in Cryptology – EUROCRYPT 2015, Part I. Lecture Notes in Computer Science, vol. 9056, pp. 513–536. Springer (2015)
10. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: TFHE: Fast fully homomorphic encryption over the torus. Journal of Cryptology **33**(1), 34–91 (2020), earlier versions in ASIACRYPT 2016 and 2017
11. van Dijk, M., Gentry, C., Halevi, S., Vaikuntanathan, V.: Fully homomorphic encryption over the integers. In: Advances in Cryptology – EUROCRYPT 2010. Lecture Notes in Computer Science, vol. 6110, pp. 24–43. Springer (2010)
12. Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. In: 33rd International Conference on Machine Learning (ICML 2016). Proceedings of Machine Learning Research, vol. 48, pp. 201–210. PMLR (2016)
13. Ducas, L., Micciancio, D.: FHEW: Bootstrapping homomorphic encryption in less than a second. In: Advances in Cryptology – EUROCRYPT 2015, Part I. Lecture Notes in Computer Science, vol. 9056, pp. 617–640. Springer (2015)
14. The EU General Data Protection Regulation (GDPR). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN
15. Gentry, C.: Computing arbitrary functions of encrypted data. Communications of the ACM **53**(3), 97–105 (2010), earlier version in STOC 2009
16. Gentry, C., Sahai, A., Waters, B.: Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In: Advances in Cryptology – CRYPTO 2013, Part I. Lecture Notes in Computer Science, vol. 8042, pp. 75–92. Springer (2013)
17. iDASH secure genome analysis competition. http://www.humangenomeprivacy.org
18. Kim, M., Harmanci, A., Bossuat, J.P., Carpov, S., Cheon, J.H., Chillotti, I., Cho, W., Froelicher, D., Gama, N., Georgieva, M., Hong, S., Hubaux, J.P., Kim, D., Lauter, K., Ma, Y., Ohno-Machado, L., Sofia, H., Son, Y., Song, Y., Troncoso-Pastoriza, J., Jiang, X.: Ultra-fast homomorphic encryption models enable secure outsourcing of genotype imputation. bioXxiv (2020)
19. Kim, M., Song, Y., Li, B., Micciancio, D.: Semi-parallel logistic regression for GWAS on encrypted data. Cryptology ePrint Archive, Report 2019/294 (2019)
20. Langlois, A., Stehlé, D.: Worst-case to average-case reductions for module lattices. Designs, Codes and Cryptography **75**(3), 565–599 (2015)
21. LeCun, Y., Cortez, C., Burges, C.C.J.: The MNIST database of handwritten digits, available at http://yann.lecun.com/exdb/mnist/
22. Lyubashevsky, V., Peikert, C., Regev, O.: On ideal lattices and learning with errors over rings. Journal of the ACM **60**(6), 43:1–43:35 (2013), earlier version in EUROCRYPT 2010

23. Micciancio, D., Peikert, C.: Trapdoors for lattices: Simpler, tighter, faster, smaller. In: Advances in Cryptology – EUROCRYPT 2012. Lecture Notes in Computer Science, vol. 7237, pp. 700–718. Springer (2012)
24. ONNX Runtime: Optimize and accelerate machine learning inferencing and training. https://microsoft.github.io/onnxruntime/index.html
25. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. Journal of the ACM **56**(6), 34:1–34:40 (2009), earlier version in STOC 2005
26. Rivest, R.L., Adleman, L., Detouzos, M.L.: On data banks and privacy homomorphisms. In: Foundations of Secure Computation. pp. 165–179. Academic Press (1978)
27. Stehlé, D., Steinfeld, R., Tanaka, K., Xagawa, K.: Efficient public key encryption based on ideal lattices. In: Advances in Cryptology – ASIACRYPT 2009. Lecture Notes in Computer Science, vol. 5912, pp. 617–635. Springer (2009)

A Complexity Assumptions Over the Real Torus

In 2005, Regev [25] introduced the *learning with errors (LWE) problem*. Generalizations and extensions to ring structures were subsequently proposed in [27,22]. The security of TFHE relies on the hardness of torus-based problems [6,9]: the LWE assumption and the GLWE assumption [5,20] over the torus.

Definition 1 (LWE problem over the torus). *Let $n \in \mathbb{N}$ and let $\boldsymbol{s} = (s_1, \ldots, s_n) \xleftarrow{\$} \mathbb{B}^n$. Let also χ be an error distribution over $\mathbb{R}$. The* learning with errors *(LWE)* over the torus problem *is to distinguish the following distributions:*

- $\mathcal{D}_0 = \left\{(\boldsymbol{a}, r) \mid \boldsymbol{a} \xleftarrow{\$} \mathbb{T}^n, r \xleftarrow{\$} \mathbb{T}\right\}$;
- $\mathcal{D}_1 = \left\{(\boldsymbol{a}, r) \mid \boldsymbol{a} = (a_1, \ldots, a_n) \xleftarrow{\$} \mathbb{T}^n, r = \sum_{j=1}^{n} s_j \cdot a_j + e, e \leftarrow \chi\right\}$.

Definition 2 (GLWE problem over the torus). *Let $N, k \in \mathbb{N}$ with N a power of 2 and let $\mathfrak{s} = (\mathfrak{s}_1, \ldots, \mathfrak{s}_k) \xleftarrow{\$} \mathbb{B}_N[X]^k$. Let also χ be an error distribution over $\mathbb{R}_N[X]$. The* general learning with errors *(GLWE)* over the torus problem *is to distinguish the following distributions:*

- $\mathcal{D}_0 = \left\{(\mathfrak{a}, \mathfrak{r}) \mid \mathfrak{a} \xleftarrow{\$} \mathbb{T}_N[X]^k, \mathfrak{r} \xleftarrow{\$} \mathbb{T}_N[X]\right\}$;
- $\mathcal{D}_1 = \left\{(\mathfrak{a}, \mathfrak{r}) \mid \mathfrak{a} = (\mathfrak{a}_1, \ldots, \mathfrak{a}_k) \xleftarrow{\$} \mathbb{T}_N[X]^k, \mathfrak{r} = \sum_{j=1}^{k} \mathfrak{s}_j \cdot \mathfrak{a}_j + \mathfrak{e}, \mathfrak{e} \leftarrow \chi\right\}$.

The *decisional LWE assumption* (resp. the *decisional GLWE assumption*) asserts that solving the LWE problem (resp. GLWE problem) is infeasible for some security parameter λ, where $n := n(\lambda)$ and $\chi := \chi(\lambda)$ (resp. $N := N(\lambda)$, $k = k(\lambda)$, and $\chi := \chi(\lambda)$).

B Algorithms

We use the notations of Section 4. The input of the (programmable) bootstrapping is an $\overline{\text{LWE}}$ ciphertext $\bar{\boldsymbol{c}} \leftarrow \overline{\text{LWE}}_{\boldsymbol{s}}(\bar{\mu}) = (\bar{a}_1, \ldots, \bar{a}_n, \bar{b}) \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$ that encrypts a plaintext $\bar{\mu} \in \mathbb{Z}/q\mathbb{Z}$ under the secret key $\boldsymbol{s} = (s_1, \ldots, s_n) \in \mathbb{B}^n$.

B.1 Blind Rotation

The secret key bits s_j used to encrypt the input $\overline{\mathrm{LWE}}$ ciphertext cannot be revealed. They are instead provided as *bootstrapping keys*; i.e., encrypted under some encryption key $\mathfrak{z}' \in \mathbb{B}_N[X]^k$: $\mathsf{bsk}[j] \leftarrow \overline{\mathrm{GGSW}}_{\mathfrak{z}'}(s_j)$ for all $j = 1, \ldots, n$.

We then have:

Algorithm 1: Blind rotation.

$\mathrm{ACC} \leftarrow (0, \ldots, 0, X^{-\tilde{b}} \cdot \bar{\mathfrak{v}})$
for $i = 1$ **to** n **do**
  $\mathrm{ACC} \leftarrow \mathrm{CMux}(\mathsf{bsk}[i], \mathrm{ACC}, X^{\tilde{a}_i} \cdot \mathrm{ACC})$
end for
return ACC

At the end of the loop, ACC contains a $\overline{\mathrm{GLWE}}$ encryption of $X^{-\tilde{\mu}^*} \cdot \bar{\mathfrak{v}}$ under key $\mathfrak{z}'$.

B.2 Sample Extraction

The sample extraction algorithm extracts the constant coefficient $\bar{\mu}$ of polynomial $\bar{\mathfrak{u}} \in \hat{\mathbb{Z}}_N[X]$ in $\overline{\mathrm{GLWE}}$ ciphertext $\bar{\boldsymbol{c}}'$ as a $\overline{\mathrm{LWE}}$ ciphertext of $\bar{\mu}$. In more detail, let $\mathfrak{z}' = (\mathfrak{z}'_1, \ldots, \mathfrak{z}'_k) \in \mathbb{B}_N[X]^k$ with $\mathfrak{z}'_j = s'_{j,0} + \cdots + s'_{j,N-1} X^{N-1}$ for $1 \leq j \leq k$. Parsing $\bar{\boldsymbol{c}}' \leftarrow \overline{\mathrm{GLWE}}_{\mathfrak{z}'}(\bar{\mathfrak{u}}) \in \hat{\mathbb{Z}}_N[X]^k$ as $(\bar{\alpha}'_1, \ldots, \bar{\alpha}'_k, \bar{\mathfrak{b}}')$ with $\bar{\alpha}'_j = \bar{a}'_{j,0} + \cdots + \bar{a}'_{j,N-1} X^{N-1}$ for $1 \leq j \leq k$ and $\bar{\mathfrak{b}}' = \bar{b}'_0 + \cdots + \bar{b}'_{N-1} X^{N-1}$, it can be verified that $\bar{\boldsymbol{c}}' \coloneqq (\bar{a}'_{1,0}, -\bar{a}'_{1,N-1}, \ldots, -\bar{a}'_{1,1}, \ldots, \bar{a}'_{k,0}, -\bar{a}'_{k,N-1}, \ldots, -\bar{a}'_{k,1}, \bar{b}'_0) \in (\mathbb{Z}/q\mathbb{Z})^{kN+1}$ is a $\overline{\mathrm{LWE}}$ encryption of $\bar{\mu}$ under the key $\boldsymbol{s}' = (s'_1, \ldots, s'_{kN}) \in \mathbb{B}^{kN}$ where $s'_{l+1+(j-1)N} \coloneqq s'_{j,l}$ for $1 \leq j \leq k$ and $0 \leq l \leq N-1$.

B.3 Key Switching

The key switching technique can be used to switch encryption keys in different parameter sets [7, § 1.2]. Its implementation requires *key-switching keys*, i.e., $\overline{\mathrm{LWE}}$ encryptions of the key bits of $\boldsymbol{s}'$ with respect to the original key $\boldsymbol{s}$. Assume we are given the key-switching keys $\mathsf{ksk}[i,j] \leftarrow \overline{\mathrm{LWE}}_{\boldsymbol{s}}\big(s'_i \cdot \frac{q}{B_{\mathsf{KS}}^j}\big)$ for all $1 \leq i \leq kN$ and $1 \leq j \leq \ell_{\mathsf{KS}}$, for some parameters $B \coloneqq B_{\mathsf{KS}}$ and $\ell \coloneqq \ell_{\mathsf{KS}}$ defining a gadget decomposition (see Section 3.3). Adapting [10, § 4.1] teaches that, on input $\overline{\mathrm{LWE}}$ ciphertext $\boldsymbol{c}' \leftarrow \overline{\mathrm{LWE}}_{\boldsymbol{s}'}(\bar{\mu}) = (\bar{a}'_1, \ldots, \bar{a}'_{kN}, \bar{b}') \in (\mathbb{Z}/q\mathbb{Z})^{kN+1}$ under the key $\boldsymbol{s}' = (s_1, \ldots, s_{kN}) \in \mathbb{B}^{kN}$,

$$\bar{\boldsymbol{c}}'' \coloneqq (0, \ldots, 0, \bar{b}') - \sum_{i=1}^{kN} \sum_{j=1}^{\ell_{\mathsf{KS}}} \bar{a}'_{i,j}\, \mathsf{ksk}[i,j] \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$$

where $(\bar{a}'_{i,1}, \ldots, \bar{a}'_{i,\ell_{\mathsf{KS}}}) = \boldsymbol{g}^{-1}(\bar{a}'_i)$ is an $\overline{\mathrm{LWE}}$ encryption of $\bar{\mu}$ under key $\boldsymbol{s}$, provided that the resulting noise keeps small.