Ripple: Accelerating Programmable Bootstraps for FHE with Wavelet Approximations

Charles Gouert[1(✉)], Mehmet Ugurbil[2], Dimitris Mouris[2], Miguel de Vega[2], and Nektarios G. Tsoutsos[1]

[1] University of Delaware, Newark, USA
{cgouert,tsoutsos}@udel.edu
[2] Nillion, Zug, Switzerland
{memo,dimitris,miguelv}@nillion.com

Abstract. Homomorphic encryption can address key privacy challenges in cloud-based outsourcing by enabling potentially untrusted servers to perform meaningful computation directly on encrypted data. While most homomorphic encryption schemes offer addition and multiplication over ciphertexts natively, any non-linear functions must be implemented as costly polynomial approximations due to this restricted computational model. Nevertheless, the CGGI cryptosystem is capable of performing arbitrary univariate functions over ciphertexts in the form of lookup tables through the use of programmable bootstrapping. While promising, this procedure can quickly become costly when high degrees of precision are required. To address this challenge, we propose Ripple: a framework that introduces different approximation methodologies based on discrete wavelet transforms (DWT) to decrease the number of entries in homomorphic lookup tables while maintaining high accuracy. Our empirical evaluations demonstrate significant error reduction compared to plain quantization methods across multiple non-linear functions. Notably, Ripple improves runtime performance for realistic applications, such as logistic regression and Euclidean distance.

Keywords: Cryptography · Homomorphic Encryption · Lookup Tables

1 Introduction

Cloud computing enables corporations to leverage powerful computational resources, while avoiding the cost and upkeep associated with maintaining local computing infrastructure. Along with numerous benefits, this gives rise to privacy concerns over outsourced data as cloud service providers can possibly view

C. Gouert and M. Ugurbil—The first two authors have equal contribution and appear in alphabetical order.

C. Gouert and N.G. Tsoutsos would like to acknowledge the support of the National Science Foundation (Award #2239334).

N. Mouha and N. Nikiforakis (Eds.): ISC 2024, LNCS 15257, pp. 273–293, 2025.
https://doi.org/10.1007/978-3-031-75757-0_14

data stored on their servers, and malicious actors have increasingly targeted cloud servers as they can be treasure troves of proprietary information from multiple clients [22,31]. While encryption techniques such as AES can be used to protect data confidentiality, the encrypted data must be static and the cloud cannot apply meaningful processing on ciphertexts (aside from storage). Thus, if a client wants to modify their data, they will have to download the ciphertexts, decrypt them, perform a computation to update the plaintext data, re-encrypt them, and re-upload the result to the cloud.

Fully Homomorphic encryption (FHE) is a powerful technique that helps address privacy concerns in cloud computing by allowing computation on encrypted data [13]. With FHE, a client can encrypt sensitive data, upload the corresponding ciphertexts to the cloud, have the cloud apply an arbitrary algorithm such as image classification, and then receive a valid encryption of the results, which can only be decrypted with the client's secret key. In this way, the cloud learns nothing about the contents of the input data, intermediate results, or the output of the encrypted computation.

Nevertheless, for many FHE schemes such as BGV [3] and CKKS [5], evaluating non-linear functions directly remains impossible, as only addition and multiplication operations can be executed over ciphertexts. Conversely, CGGI [7] allows encrypted lookup tables (LUTs), which allow non-linear functions to be computed exactly. Unfortunately, this operation is quite costly and becomes significantly more expensive as the plaintext modulus increases. Likewise, both the time and memory required to generate the lookup tables scale exponentially with the input's precision, with LUT sizes of 32 bits and larger becoming impractical.[1]

One solution to this problem is to simply quantize the LUT inputs (i.e., reduce the bit width) to lower the number of entries in the lookup table, resulting in faster evaluation and LUT generation times. The reduced precision caused by quantization, however, can negatively impact a wide variety of applications that require high precision. For instance, quantization in deep neural networks can result in non-negligible accuracy loss [20] and thus lead to incorrect classifications. Particularly, errors occurring due to quantization in early layers will propagate to subsequent layers, resulting in an avalanche effect, where the errors are compounded in each layer. Indeed, this effect is not limited to privacy-preserving inference and can happen in any application.

In this work, we propose the *Ripple* framework that offers new efficient techniques for encrypted LUT evaluation. Ripple provides all of the same benefits of quantization in terms of latency reduction while minimizing the accuracy loss resulting from reduced precision and bit widths. Our approach leverages the *discrete wavelet transform* (DWT) [28] to approximate non-linear functions and generate significantly smaller lookup tables while maintaining high accuracy. Moreover, Ripple employs multiple DWT families and introduces bespoke FHE-friendly protocols tailored to each family to maximize accuracy and minimize

[1] We empirically observed even 32-bit encrypted LUTs with the state-of-the-art TFHE-rs [32] FHE library require approximately 515 GB of RAM and 65 min. For reference, 30-bit LUTs took almost 15 min requiring over 120 GB..

latency. For instance, we find that some DWT families benefit from multiple LUTs while others need only a single LUT evaluation. We apply Ripple to a variety of non-linear functions that are widely used across several domains from machine learning [4,11] to statistics [26], as well as multiple realistic applications for homomorphic encryption [17,29], such as logistic regression inference and edge detection. Our contributions can be summarized as follows:

- We introduce Ripple to construct smaller encrypted LUTs with wavelet techniques without sacrificing accuracy.
- We propose multiple protocols for evaluating wavelet-encoded LUTs in the encrypted domain.
- We implement a suite of commonly adopted non-linear functions and optimize with Ripple, along with a set of benchmarks from various domains.

2 Preliminaries

2.1 Fully Homomorphic Encryption Overview

The key characteristic of all FHE schemes is *malleability*, where ciphertexts can be manipulated to change the underlying plaintext data predictably. All FHE schemes can be roughly divided into two categories depending on the primary computational domain: arithmetic-based schemes and Boolean-based schemes.

We focus on the latter class, where instead of encrypting integers or floating point numbers, Boolean schemes encrypt individual bits (or low-precision integers in certain cases). The addition and multiplication primitives are replaced by encrypted gate operations, such as `AND`, `OR`, and `NOT` gates. In practice, most logic gates are implemented as a series of linear operations between ciphertext polynomials followed by a *functional bootstrap*, which serves to scale the output to the expected value. Because Boolean schemes can support all standard logic gates, they are capable of executing arbitrary algorithms. Contrary to arithmetic-based schemes, this class of cryptosystems is capable of evaluating non-linear functions directly through the use of Boolean circuits.

Additionally, unlike arithmetic-based schemes, Boolean schemes do not need to utilize polynomial approximations. Indeed, non-linear operations can be implemented exactly as a Boolean circuit. As an example, the ReLU activation function is directly mapped to a multi-bit comparator circuit followed by a multiplexer. However, this may require a larger number of Boolean gates and the majority of gate types require at least one bootstrapping operation, resulting in relatively high latency for large circuits. Prior frameworks, such as the Google Transpiler [15], ArctyrEX [19], and HELM [16], exploit the inherent circuit-level parallelism to reduce the latency of circuit evaluation. Still, the performance of these frameworks is limited by the critical path (or greatest depth) of the homomorphic circuit. Additionally, all three approaches rely on logic and/or high-level synthesis methods to convert input programs to optimized Boolean circuits, which results in high pre-processing cost for non-trivial applications.

Alternatively, with proper parameter selection, Boolean schemes can encrypt low-precision integers. In the case of the CGGI cryptosystem [6], this allows for ciphertext addition and multiplication with a public constant, but not multiplication between two ciphertexts. Notably, it still retains the functional bootstrap, which can be utilized to evaluate $N{:}N$ lookup tables. This approach combines some of the key strengths of both arithmetic-based schemes and the Boolean mode of operation in the form of natively supported multi-bit arithmetic and a mechanism for exactly evaluating non-linear functions.

The primary challenge is the restriction on the size of the plaintext space, but this can be overcome by representing high-precision plaintexts as vectors of ciphertexts, where each encryption encodes a low-precision chunk of the original message. This very methodology is employed in the TFHE-rs library [32] in two different ways: a Chinese remainder theorem (CRT) method and a radix method. The former involves generating multiple residues by reducing an input message by a series of co-prime bases and encrypting each residue as a separate ciphertext. Upon decryption, the residues are combined to form the final higher-precision result. The second method involves decomposing the input data into a series of digits, each of which is decrypted individually. Ripple utilizes the latter approach as it provides a convenient way to truncate ciphertexts, which is an integral operation in the DWT protocols explained in the following section. Truncating the digits of a ciphertext array encoded with the radix decomposition can be done with negligible latency overhead as no FHE operations are required.

Programmable Bootstrapping (PBS). A crucial feature of the DM and CGGI FHE cryptosystems is the *functional bootstrap*, which takes advantage of the programmability of the bootstrapping algorithm employed in these schemes. A polynomial with crafted coefficients that encodes the set of desired output messages is rotated by an encrypted value and the first encrypted coefficient corresponding to the constant term of the polynomial is extracted. These two procedures, *blind rotation* and *extraction*, form the core bootstrapping steps.

By encoding chosen lookup table (LUT) entries in the coefficients of the polynomial to be rotated, one can evaluate a LUT T over a ciphertext. Essentially, this can be done by rotating the LUT polynomial by an encrypted amount (corresponding to the input ciphertext) and extracting the entry corresponding to the constant term. The result is a valid encryption that encodes the mapping from a LUT input to a desired LUT output. Thus, it allows computing arbitrary univariate functions by evaluating a function in the plaintext domain across all possible inputs and encoding them in the polynomial utilized during bootstrapping. This generalized bootstrapping technique is called *programmable bootstrapping (PBS)* [9,24]. Although the LUT needs to be relatively small to maintain efficient cryptographic parameter sets, it has two main advantages. First, it can encode any arbitrary univariate function, and second, it leads to a significant performance boost as it replaces expensive operations that otherwise would require multiple additions and multiplications.

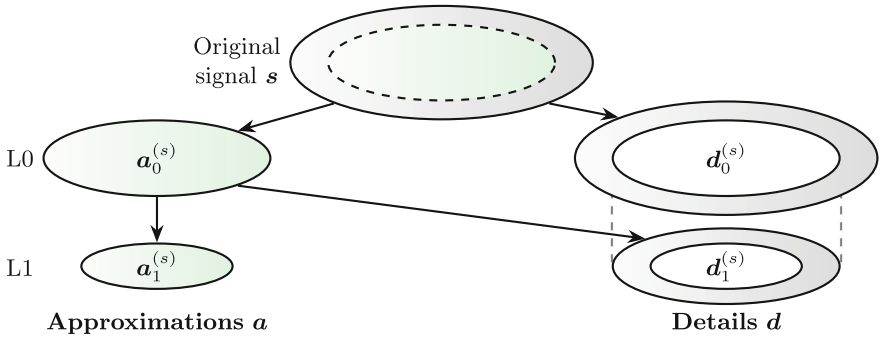


Fig. 1. Two DWT iterations. The signal $\boldsymbol{s}$ can be represented by the approximation $\boldsymbol{a}^{(s)}$ (green) and the details $\boldsymbol{d}^{(s)}$ (gray). (Color figure online)

2.2 The Discrete Wavelet Transform (DWT)

A Discrete Wavelet Transformation (DWT) is a process of splitting a discretely sampled signal into two parts: the approximation and the detail coefficients [28]. The former is half of the size of the original signal but encompasses the most interesting parts of it, while the latter contains information about the error incurred by the approximation coefficients. When combining both the detail coefficients and the approximation, one can reconstruct the original signal. Figure 1 illustrates the original signal on top and two applications of the DWT. In the first level (L0), we deconstruct the signal on top to approximation ($\boldsymbol{a}_0^{(s)}$) and detail coefficients ($\boldsymbol{d}_0^{(s)}$). Then, we can repeat the same process by treating the L0 approximation as a new signal and thus get new approximation ($\boldsymbol{a}_1^{(s)}$) and detail coefficients ($\boldsymbol{d}_1^{(s)}$) at level 1, and so on. In Fig. 1, notice that given the L1 approximation and detail coefficients along with the L0 detail coefficients, it is sufficient for recovering the original signal.

There exist multiple wavelet families such as the Daubechy (Db) and the Biorthogonal wavelets. A special case of Db wavelets is the Db-1 – or *Haar* – wavelet. The core idea in all families is a matrix multiplication with a constant matrix M, where M differs based on the family. Haar uses an orthogonal matrix to obtain the approximation coefficients, which are calculated by averaging every two consecutive points of the original signal. Biorthogonal, on the other hand, relies on two different matrices, where the transpose of one matrix is the inverse of the other. We delve more into the details of both wavelet families in Sect. 3.

Haar DWT. Haar applies a linear transformation to the input and generates the approximation and detail coefficients. Starting with a signal $\boldsymbol{s} = [s_0, \ldots, s_{2N-1}]$ of length $2N$, the Haar DWT generates the approximation coefficients $\boldsymbol{a} = [a_0, \ldots, a_{N-1}]$ and the detail coefficients $\boldsymbol{d} = [d_0, \ldots, d_{N-1}]$, each with N entries. More specifically, the approximations are generated as $a_k = (s_{2k} + s_{2k+1})/2$, while the details are generated by $d_k = (s_{2k} - s_{2k+1})/2$ for $k \in [0, \ldots, N)$.

It is easy to see that when the Haar DWT is applied to a one-hot vector, it results in yet another one-hot vector, albeit scaled. The index of the non-zero value in the new vector is in fact the old index divided by 2, which is akin to removing the last bit of the index. Therefore, we can simply truncate the value to get rid of the least significant bits and end up with the value that we want to do the lookup at. In effect, we have manually applied the Haar DWT transform to the one-hot vector using only truncation.

Biorthogonal DWT. The linear transformation in Haar can be seen as a matrix multiplication with some public orthogonal matrix.[2] The Biorthogonal DWT on the other hand, requires two different matrices, where the transpose of one matrix is the inverse of the other, i.e., $M_1^T = M_2^{-1}$. In this case, M_1 is used for the decomposition of the signal, while M_2 is used for the reconstruction.

In Haar, we observe that the points in the approximation coefficient are averaged from the input signal S and the details are complementary in order to be able to recover S. Conversely, in Biorthogonal wavelets, the approximations and details are computed with weighted averages [28]. Since the Biorthogonal wavelets have more non-zero filters, when the transform is applied to a one-hot vector, the result is not a one-hot vector. Even so, we can manually calculate the resulting vector as it is a weighted average of the two consecutive values starting at the most significant bits of the original index. The weights depend on the least significant bits of the original index; hence, by splitting the original index into the MSBs and LSBs, we can calculate the transformed vector.

3 The Ripple Framework

As mentioned in Sect. 2.1, a key feature of the CGGI cryptosystem is the ability to evaluate a lookup table with the programmable bootstrapping mechanism. Complex non-linear functions can now be encoded as LUTs and evaluated homomorphically, eliminating the need to perform expensive polynomial approximations. Unfortunately, as the size of the LUT grows, this technique becomes prohibitively expensive (recall footnote 1), and thus many non-linearities are impossible to evaluate in applications that require high precision.

We address this challenge by utilizing the DWT to reduce the size of LUTs without sacrificing correctness. Ripple is the first framework to explore wavelet approximations for FHE as a way to accelerate programmable bootstrapping. Our key observation is that if we apply the DWT to signals that represent smooth functions (e.g., logarithm, square root, sigmoid, etc.), then the detail coefficients are relatively small compared to the approximation coefficients. This means that our approximation is sufficient to represent the original signal and we can completely disregard the detail coefficients while maintaining a minimal error relative to the original function. Utilizing this observation, we can zero out the details in Fig. 1, and by just applying the DWT a single time, we can

[2] An orthogonal matrix M has the property that $MM^T = I$, where I is the identity matrix. A matrix M is orthogonal if its transpose (M^T) is equal to its inverse (M^{-1}).

halve the size of the LUT. This signal might still be quite big, so we can repeat the same process and half the LUT size even further. Of course, as this is an approximation, the smaller the LUT size, the higher the error we might have. With Ripple, however, these errors are marginal as we show in Sect. 5.

Ripple's Key Observation. The core idea behind Ripple relies on the orthogonality or biorthogonality of the DWT transform applied to the inner product. Consider the fact that the inner product between two vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ is equal to the product between the transpose of the first vector and the second, i.e., $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \boldsymbol{v}^T \cdot \boldsymbol{u}$. As described in Sect. 2.2, the DWT of both $\boldsymbol{v}$ and $\boldsymbol{u}$ is a multiplication with a matrix M, which we can view as $\mathsf{DWT}(\boldsymbol{v}) = M \cdot \boldsymbol{v}$. This results in a vector $\left[\frac{\boldsymbol{a}^{(v)}}{\boldsymbol{d}^{(v)}}\right]$, where $\boldsymbol{a}^{(v)}$ and $\boldsymbol{d}^{(v)}$ represent the approximation and detail coefficients of $\boldsymbol{v}$, respectively. Similarly for $\mathsf{DWT}(\boldsymbol{u}) = M \cdot \boldsymbol{u}$ we get $\boldsymbol{a}^{(u)}$ and $\boldsymbol{d}^{(u)}$. Observe that using orthogonality,

$$\begin{aligned} \langle \mathsf{DWT}(\boldsymbol{v}), \mathsf{DWT}(\boldsymbol{u}) \rangle &= \langle M \cdot \boldsymbol{v}, M \cdot \boldsymbol{u} \rangle = (M \cdot \boldsymbol{v})^T \cdot M \cdot \boldsymbol{u} \\ &= \boldsymbol{v}^T \cdot M^T \cdot M \cdot \boldsymbol{u} = \boldsymbol{v}^T \cdot I \cdot \boldsymbol{u} = \boldsymbol{v}^T \cdot \boldsymbol{u} = \langle \boldsymbol{v}, \boldsymbol{u} \rangle. \end{aligned} \tag{1}$$

We remark that the inner product of two vectors is equal to the sum of the inner product of the approximation coefficients and the inner product of the detail coefficients of the DWT transforms of the vectors, as follows:

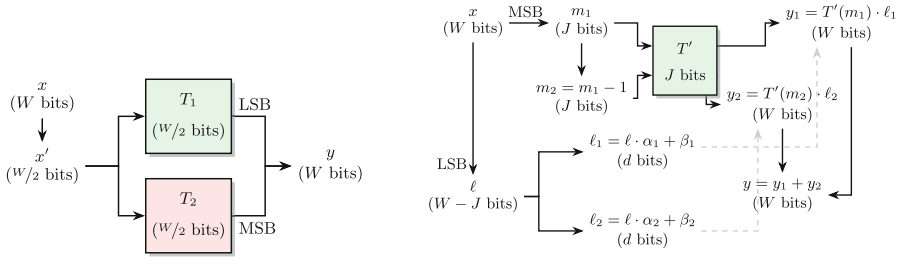
$$(M \cdot \boldsymbol{v})^T \cdot M \cdot \boldsymbol{u} = \left[\frac{\boldsymbol{a}^{(v)}}{\boldsymbol{d}^{(v)}}\right]^T \cdot \left[\frac{\boldsymbol{a}^{(u)}}{\boldsymbol{d}^{(u)}}\right] = \boldsymbol{a}^{(v)T} \cdot \boldsymbol{a}^{(u)} + \boldsymbol{d}^{(v)T} \cdot \boldsymbol{d}^{(u)}.$$

In Ripple, we represent $\langle \boldsymbol{v}, \boldsymbol{u} \rangle$ as $\langle \boldsymbol{a}^{(v)}, \boldsymbol{a}^{(u)} \rangle + \langle \boldsymbol{d}^{(v)}, \boldsymbol{d}^{(u)} \rangle$ which is approximately equal to the inner product of their respective approximation coefficient vectors $\boldsymbol{a}^{(v)}$ and $\boldsymbol{a}^{(u)}$. By dropping the detail coefficients, we get the approximation of the original inner product via the inner product of the approximation coefficients, which is key to our proposed lookup methodology. Thus, $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \langle \mathsf{DWT}(\boldsymbol{v}), \mathsf{DWT}(\boldsymbol{u}) \rangle \approx \langle \boldsymbol{a}^{(v)}, \boldsymbol{a}^{(u)} \rangle$.

This works nicely for orthogonal DWTs, but for the Biorthogonal DWT, we need extra considerations. Instead of applying the same transformation to both $\boldsymbol{v}$ and $\boldsymbol{u}$, we have to apply the decomposition matrix to one, while applying the reconstruction matrix to the other. Then, by biorthogonality of these matrices, the same observation holds and $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \langle M_1 \cdot \boldsymbol{v}, M_2 \cdot \boldsymbol{u} \rangle \approx \langle \boldsymbol{a}^{(v)}, \boldsymbol{a}^{(u)} \rangle$.

Applying Our Observation to the Encrypted Domain. In Eq. (1), we can view $\boldsymbol{v}$ as a one-hot vector where the non-zero value is at the index of the ciphertext and $\boldsymbol{u}$ as a public LUT T' that approximates the full-precision LUT T. Then, the inner product $\langle \boldsymbol{v}, \boldsymbol{u} \rangle$ will yield the lookup value in table T' at the non-zero index of $\boldsymbol{v}$, which is the lookup value at the ciphertext. In particular, $\langle \boldsymbol{a}^{(v)}, \boldsymbol{a}^{(u)} \rangle$ will be an approximation of this lookup. Notably, the approximation of the LUT is easy to calculate since it is in plaintext, while we need a way to efficiently calculate the approximation vector of the one-hot vector that represents the ciphertext. Fortunately, it turns out that this approximation vector can be calculated by a weighted sum of lookups.

Of course, the aforementioned technique is not practical in FHE, but programmable bootstrapping (PBS) can be leveraged for this purpose. Starting with a ciphertext x' (which is a truncated version of the full-precision x) we can evaluate the LUT T' on x' homomorphically and obtain ciphertext $y' = T'(x')$. The novelty of Ripple lies in that x' has fewer bits than x and T' has fewer entries than T, while y' is a close approximation to $y = T(x)$. Ripple focuses on the two most popular DWT families: Haar and Biorthogonal, which are both viable and constitute a tradeoff between accuracy and latency (as discussed in Sect. 5). In the following subsections, we describe how Ripple formulates and evaluates encrypted lookup tables with each of the two DWT families.



(a) Haar DWT. (b) Biorthogonal DWT.

Fig. 2. DWT LUT approximating $y = T(x)$. In (a), the approximation computes $y = T_1(x') \,\|\, T_2(x')$, where $x' = (x \gg {}^{W}/_{2})$. In (b), we compute $y = T'(m_1) \cdot \ell_1 + T'(m_2) \cdot \ell_2$, where $m_1 = (x \gg (W - J))$, $m_2 = m_1 - 1$, $\ell = x \mod 2^{W-J}$, $\ell_1 = \ell \cdot \alpha_1 + \beta_1$, and $\ell_2 = \ell \cdot \alpha_2 + \beta_2$ with $\alpha_1, \alpha_2, \beta_1, \beta_2$ being public scalars.

3.1 PBS with Haar DWT

Let W be the bit width of our input radix-decomposed ciphertext vector, where each ciphertext encodes a plaintext digit (as explained in Sect. 2). Naturally, we can represent the function that we want to approximate as a LUT T that maps k-bit inputs to k-bit outputs. For simplicity, we can assume that $k = W$. Creating a LUT of 2^W entries, however, is not always feasible. For instance, $W = 64$ requires generating and storing 2^{64} LUT entries, which is completely impractical. A straightforward option is to truncate our input ciphertexts by J bits and limit T to only have 2^{W-J} entries and operate over $W - J$ bit inputs and outputs. Notably, this truncation simply involves deleting encrypted digits and corresponds exactly to truncation in the plaintext domain. As it turns out, this approximation incurs high errors and is not sufficient for most applications.

Ripple takes a different approach: First, we apply the Haar DWT over the public LUT T iteratively J times by dropping the detail coefficients to end up with an approximation T' of our original function. The new table T' now operates over $W - J$ bits. Observe, however, that our input ciphertext vector is still W-bits long. To index T', Ripple truncates the vector to encode $W - J$ bits so it can be used during PBS to index the LUT. As long as $W - J$ bits is a

multiple of the size of our radix digits, the truncation is free. Conveniently, this is the index needed for the approximation vector $\boldsymbol{a}^{(v)}$. This results in an output ciphertext vector also encoding $W - J$ bits.

As we started with an W-bit input ciphertext vector, after approximating the function we need to end up with a W-bit output as well. To do so, we repeat the same process twice by building two LUTs (T_1 and T_2); one for the LSBs and one for the MSBs. In both cases, the truncated $W - J$ bit input ciphertext vector is used to index the LUT. This is illustrated in Fig. 2a, where $J = W/2$. Note that each of T_1 and T_2 has $W/2$ bits and thus takes approximately half the time to be evaluated compared to a W-bit LUT (which is not even possible to practically create for large W). Additionally, both tables can be evaluated in parallel and finally, the two outputs can be concatenated to get the final encrypted result encoding W bits of data.

We remark that, in certain functions, we only need to evaluate the LSB table and we can avoid evaluating T_2 altogether. For instance, any function where the output can fit in less than half the bit width, such as the square root or Sigmoid activation function, only needs a single LUT evaluation.

3.2 PBS with Biorthogonal DWT

The Haar DWT is quite efficient since its approximation vector is again one-hot and hence only has a single approximation that needs to be looked up; however, this benefit also comes at a limitation. Namely, all the points that have the same most significant bits evaluate to the same value. The Biorthogonal DWT overcomes this challenge by making use of the least significant bits of the input ciphertext, as well as the most significant bits. This comes at the cost of doing two lookups followed by two multiplications and an addition, but results in more accurate approximations. In the end, we get significantly better compression, in fact, the same compression used in JPEG2000 [27].

The Biorthogonal DWT follows a similar approach as the Haar DWT, yet incurs more operations as its linear transformation computes weighted averages. Contrary to Haar where we had to evaluate two separate LUTs, in Biorthogonal wavelets we need to evaluate a single LUT (T') across two different indices. We observe that this can also be done using two LUTs, where, as an optimization, the second table T_2 equals the first one (T_1) shifted by an index: $T_2(x) = T_1(x+1 \mod |T_1|)$, where $|T_1|$ refers to the size of T_1.

Figure 2b demonstrates how Ripple approximates an LUT with Biorthogonal wavelets. First, we extract the most significant bits (MSBs) from the input to be used for the lookup, and the least significant bits (LSBs), which will be used to combine the lookups via their weighted average. Starting with a W-bit input x, we split it into two parts of J and $W - J$ bits, which we call m_1 and ℓ, respectively. Here, J is equal to the depth of the DWT applied. The former (m_1) represents the J MSBs of x and it serves two purposes. First, from m_1, we create m_2 as $m_1 - 1$. Then, we use both m_1 and m_2 to index two consecutive entries of the DWT-encoded LUT and end up with ciphertexts $T'(m_1)$ and $T'(m_2)$. The latter (ℓ) represents the $W - J$ LSBs of x and is used to compute two linear

expressions with public constant values α_1, α_2, β_1, and β_2 as $\ell_1 = \ell \cdot \alpha_1 + \beta_1$, and $\ell_2 = \ell \cdot \alpha_2 + \beta_2$. These values come from the non-zero entries of the approximation coefficients vector $\boldsymbol{a}^{(v)}$: $\alpha_1 = -1$, $\alpha_2 = 1$, $\beta_1 = 2^J$, and $\beta_2 = 0$. Lastly, we multiply $T'(m_1)$ by ℓ_1 and $T'(m_2)$ by ℓ_2 computing the inner product in the DWT domain, so we go back to W bits and sum the two ciphertexts together to get our final output.

We are usually able to compress the LUT for the Biorthogonal DWT into J bits as these values do not have to cover the entire output range, unlike Haar which requires a W-bit LUT output. However, when this is not possible, we employ a similar method of evaluating multiple LUTs in parallel with the same J-bit inputs and combine the J-bit outputs to form a higher bit-width result.

4 Function-Centric Compression

In this Section, we investigate two optimizations to further reduce the LUT sizes of specific classes of non-linear functions; namely, we propose optimizations for symmetrical functions and functions where the complex non-linearity converges to some value outside a certain interval.

Symmetrical Functions. First, we exploit the symmetry of certain functions to further reduce the LUT size. This class includes common functions in machine learning, such as reciprocal, sigmoid (σ), hyperbolic tangent (tanh), and the error function (erf). More formally, these functions exhibit the following property: $\frac{1}{-x} = -\frac{1}{x}$, $\sigma(-x) = 1 - \sigma(x)$, $\mathsf{tanh}(-x) = -\mathsf{tanh}(x)$ and $\mathsf{erf}(-x) = -\mathsf{erf}(x)$. Therefore, if we know the sign of the input, we can evaluate the function strictly in the positive domain and then use this intermediate result to calculate the actual value by taking into account the sign of the input.

To apply this technique, the function must exhibit symmetry around zero, but we note that any symmetric function can be shifted to exhibit this required symmetry. For example, sigmoid becomes symmetric around zero after it is moved down on the y-axis by 0.5: $\sigma(-x) - 0.5 = -(\sigma(x) - 0.5)$. We stress that this technique easily generalizes to any symmetric function and reduces the size of the LUT by $2\times$ as only half the domain needs to be evaluated.

In general, we call a function $f(x)$ symmetric around the symmetry point $(x_{\mathtt{sym}}, y_{\mathtt{sym}})$ if $f(x_{\mathtt{sym}} - x) - y_{\mathtt{sym}} = c_{\mathtt{sym}} \cdot (f(x_{\mathtt{sym}} + x) - y_{\mathtt{sym}})$, for symmetry constant $c_{\mathtt{sym}}$. The symmetry constant defines the symmetry relationship; for instance if $c_{\mathtt{sym}} = 1$ then the function is reflected along the y-axis (like $f(x) = x^2$) while if $c_{\mathtt{sym}} = -1$ then the function is reflected along both the x and y axes (like $f(x) = \mathsf{tanh}(x)$). It is easy to see that, given $f(x)$ at some value $x_{\mathtt{sym}} + \delta$, we can compute f at $x_{\mathtt{sym}} - \delta$ simply by noting:

$$\begin{aligned} f(x_{\mathtt{sym}} - \delta) &= f(x_{\mathtt{sym}} - \delta) - y_{\mathtt{sym}} + y_{\mathtt{sym}} = c_{\mathtt{sym}} \cdot (f(x_{\mathtt{sym}} + \delta) - y_{\mathtt{sym}}) + y_{\mathtt{sym}} \\ &= c_{\mathtt{sym}} \cdot f(x_{\mathtt{sym}} + \delta) + (1 - c_{\mathtt{sym}}) \cdot y_{\mathtt{sym}}. \end{aligned}$$

Figure 3 demonstrates a function with $x_{\mathtt{sym}}, y_{\mathtt{sym}}$ and $c_{\mathtt{sym}} = -1$. Notice that for sigmoid, we have $x_{\mathtt{sym}} = 0, y_{\mathtt{sym}} = 0.5$ and $c_{\mathtt{sym}} = -1$, hence we get $\sigma(-x) =$

$-\sigma(x) + 1$. Further, if we set $g(x) = f(x_{\texttt{sym}} + x)$ and use the absolute value $|x - x_{\texttt{sym}}|$ and the condition $(x - x_{\texttt{sym}} < 0)$:

$$f(x) = g(x - x_{\texttt{sym}}) = \begin{cases} g(x - x_{\texttt{sym}}) & x - x_{\texttt{sym}} \geq 0, \\ c_{\texttt{sym}} \cdot g(x_{\texttt{sym}} - x) + (1 - c_{\texttt{sym}}) \cdot y_{\texttt{sym}} & x - x_{\texttt{sym}} < 0. \end{cases}$$

Note that this equation holds for all positive or negative values x. Thus, we build our LUT based on $g(x)$; we can use this equation to evaluate f at any value x.

Convergent Functions. We also optimize the class of functions that can be approximated by polynomials outside a region. We refer to these functions as piecewise functions and we specifically study the case where we have three intervals: two that we approximate with polynomials and an interval defined by a non-polynomial function that we evaluate with a LUT. Note that this technique can be generalized to any number of intervals.

A special case of this includes functions that are nearly constant outside a certain bounded interval. Coincidentally, sigmoid, tanh, and erf are also prominent examples of this. For instance, sigmoid is nearly 1 above a threshold of 8 and nearly 0 below -8 with less than 0.0003 of maximum difference. For tanh, the maximum difference is 0.0003 outside $[-4, 4]$, while for erf the difference is $1.54 \cdot 10^{-8}$ outside the same interval as tanh.

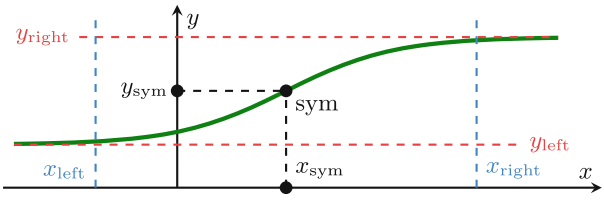


Fig. 3. Sigmoid like function where the symmetry point is x_{sym} and y_{sym}. x_{left} and x_{right} are the convergence bounds for y_{right} and y_{left}, respectively.

Given a function $f(x)$ and an interval $[x_{\texttt{left}}, x_{\texttt{right}}]$, we say this function is convergent if there exists a piecewise polynomial $p(x)$ such that for $x \notin [x_{\texttt{left}}, x_{\texttt{right}}]$, $|f(x) - p(x)| < \epsilon$ for a small constant ϵ. To approximate such convergent functions, we can check whether the input is inside the interval and select the lookup value $T(x)$ or the polynomial value accordingly as:

$$f(x) \approx (x \in [x_{\texttt{left}}, x_{\texttt{right}}]) \cdot \texttt{T}(x) + p(x).$$

Convergent Symmetrical Functions. Bringing together optimizations for symmetrical and convergent functions, we devise a function evaluation protocol in Algorithm 1 (note that only $f(x) - p(x)$ needs to be symmetric). We start by extracting the most significant bit, which is the sign bit of `ct` (called `ltz` for "less than zero" since this bit is 0 for positive values and 1 for negative). Then we

Algorithm 1. LUT Evaluation for Symmetric Convergent Functions

Public Inputs: `func` ▷ Function to be evaluated
`xsym, ysym` ▷ X and Y coordinates of the symmetry point
`csym` ▷ Symmetry coefficient
`threshold` ▷ LUT boundary
Private Input: `ct` ▷ Encrypted input value

1: **procedure** EVALUATE(`func`, `xsym`, `ysym`, `csym`, `ct`)
2: `ct` ← `ct` - `xsym` ▷ Shift to symmetry point
3: `ltz` ← EXTRACTSIGN(`ct`) ▷ Get the sign bit
4: `sign` ← 1 - 2 · `ltz` ▷ Calculate the sign as -1 or 1
5: `abs` ← `sign` · `ct` ▷ Compute the absolute value
6: `eval` ← PBS(`abs`) ▷ The LUT evaluation on `abs`
7: `sym` ← $(1 + \texttt{ltz} \cdot (\texttt{csym} - 1)) \cdot \texttt{eval} + \texttt{ltz} \cdot (1 - \texttt{csym}) \cdot \texttt{ysym}$
8: `check` ← CMP(`abs`, `threshold`) ▷ Is abs in LUT domain?
9: `poly` ← POLY(`abs`, `ltz`, `check`) ▷ Polynomial evaluation
10: **return** $\texttt{check} \cdot \texttt{sym} + \texttt{poly}$

calculate the sign as -1 (if `ltz` is 1) or 1 (if `ltz` is 0) and the absolute value (`abs` in line 5 in Algorithm 1) of the `ct` by multiplying it by the `sign`. Next, we use the LUT method to look up the function value at `abs`. Note that this could be the desired DWT LUT and involves truncation and multiplication operations. We bring it all together using the symmetry of the function (in line 7 in Algorithm 1). Next, we check if the `abs` is in the LUT domain. Notice that since the function is symmetric, we only have to check one end of $[x_{\texttt{left}}, x_{\texttt{right}}]$, which we call $\texttt{threshold} = x_{\texttt{right}} - x_{\texttt{sym}}$. Then we evaluate the polynomial approximation at `abs`. Finally, we merge the results using the convergence property.

For example, to evaluate sigmoid, we define $\text{PBS}(\texttt{abs}) := \sigma(\texttt{abs})$ for $\texttt{abs} \in [0, 8]$, $\text{CMP}(\texttt{abs}, \texttt{threshold}) := \texttt{abs} < \texttt{threshold}$, and $\text{POLY}(\texttt{abs}, \texttt{ltz}, \texttt{check}) := (1 - \texttt{check}) \cdot (1 - \texttt{ltz})$. Then we call EVALUATE with $x_{\texttt{sym}} = 0, y_{\texttt{sym}} = 0.5$ and $c_{\texttt{sym}} = -1$. On the other hand, for a non-convergent symmetrical function like reciprocal $f(x) = \frac{1}{x}$, we set $\text{CMP}(\dots) = 1$ and $\text{POLY}(\dots) = 0$.

5 Experimental Evaluation

We implemented Ripple using the state-of-the-art TFHE-rs [32] library. In our experiments, we compare Ripple against HELM [16], Romeo [18], and Google Transpiler [15], as well as baseline implementations in TFHE-rs that use LUTs configured for the full bit width of each application. For Ripple, we implemented three variants: (a) a quantized version that is similar to the baseline, but, in this case, we truncate before applying the LUTs in a similar way to preparing inputs for a Haar DWT lookup, (b) a Haar DWT variant, and (c) a Biorthogonal DWT variant. It is expected that the quantized version will outperform both the Haar and Biorthogonal DWT-encoded LUTs in terms of latency, with the last two incurring significantly fewer errors.

Table 1. Overview of runtime improvement (in seconds) and mean absolute error (MAE) over non-linear operations for Ripple using quantization, Haar, and Biorthogonal wavelets. Our baseline uses LUT sizes (T) to be equal to the word size (W), while the Ripple quantized version, as well as the Haar and Biorthogonal DWT variants, use approximations with T bits; k represents the precision.

Op.	W	k	T	Base	Quant.		Haar		Biortho.	
					Time	MAE	Time	MAE	Time	MAE
$\sqrt{x}$				9.5	2.4	2.79e-3	2.3	1.37e-3	3.1	1.92e-5
$1/x$				9.5	2.4	2.45e-4	2.4	1.26e-4	3.1	8.62e-6
$1/\sqrt{x}$				18.1	4.8	2.19e-4	2.4	1.14e-4	3.1	8.24e-6
$\log x$	24	16	12	9.5	2.4	1.73e-3	2.3	8.61e-4	3.1	7.42e-6
$\sigma(x)^*$				19.0	4.8	1.97e-2	2.3	6.37e-5	3.3	6.73e-6
$\mathsf{erf}(x)^*$				9.4	2.4	3.96e-3	2.4	1.28e-4	3.3	8.89e-6
$\mathsf{tanh}(x)^*$				9.5	2.4	3.96e-3	2.4	1.28e-4	3.3	9.41e-6
$\sqrt{x}$				N/A	3.2	6.93e-4	3.2	3.45e-4	4.2	1.35e-6
$1/x$				N/A	3.2	1.57e-5	3.2	8.03e-6	4.2	5.37e-7
$1/\sqrt{x}$				N/A	6.3	1.49e-5	3.2	7.70e-6	4.1	5.07e-7
$\log x$	32	20	16	N/A	3.2	3.99e-4	3.2	1.47e-4	4.1	1.31e-4
$\sigma(x)^*$				N/A	6.3	7.68e-6	3.2	3.97e-6	4.3	4.20e-7
$\mathsf{erf}(x)^*$				N/A	3.1	1.54e-5	3.2	7.95e-6	4.3	5.31e-7
$\mathsf{tanh}(x)^*$				N/A	3.2	1.54e-5	3.2	7.91e-6	4.3	5.19e-7

We perform a series of experiments varying from simple non-linear functions to more elaborate applications (depicted in Appendix A). We utilized a `c5.12xlarge` AWS EC2 instance with 48 virtual cores running Ubuntu 22.04. For all TFHE-rs modes, we used a parameter set corresponding to approximately 128 bits of security [2]. Specifically, we utilized parameters that allow ciphertexts to hold two data bits and two bits of carry used for intermediate computations. All radix ciphertexts are constructed as vectors of ciphertexts constructed with these parameters. Also, in our experiments, the LUT generations are pre-computed by the server.

Ripple Approximations. In Table 1 we compare the Ripple approximations with both the Haar and Biorthogonal wavelet families against a baseline TFHE-rs implementation (using a non-DWT LUT with the full bit-width) and also a quantized (i.e., approximation) version that truncates half of the bit-width of the inputs. Specifically, we consider non-linear functions including reciprocal, square root, and the sigmoid activation function. Our comparisons emphasize the time to evaluate each non-linear function in seconds as well as the mean absolute error (MAE), which is the sum of absolute errors divided by the number of samples. We perform three blocks of experiments for the same functions with different combinations of word sizes (W), precision values (k), and LUT sizes (T). We experimentally observed that setting T to $\mathsf{W}/2$ yields the fastest runtime while

minimizing the MAE. We set the precision k between the T and W because otherwise, quantization misses all the information in the fractional bits.

In Table 1, we observe that the runtime of all quantization, Haar DWT, and Biorthogonal DWT is faster than baseline, and increasingly so as the word size increases. This is intuitive as the LUTs are smaller, and therefore there are fewer PBS operations across the radix ciphertext. In the case of the Biorthogonal DWT, the extra computation associated with the multi-bit additions and multiplications is offset by the parallelism inherent in the algorithm. We observe that, for word size 24, the speedup is around to $3-4\times$, and for word size 32, the baseline is not possible to evaluate on our experimental server, as the RAM required to generate the encrypted LUTs exceeds 500 gigabytes for 32-bit tables. The Haar DWT is usually the same speed as quantization as both involve reducing the bit-width of the input by half and evaluating an LUT, but it is more accurate by halving the error observed. On the other hand, the Biorthogonal DWT is slower than the other two due to the overhead of multiplications, however, it is an order of magnitude more accurate.

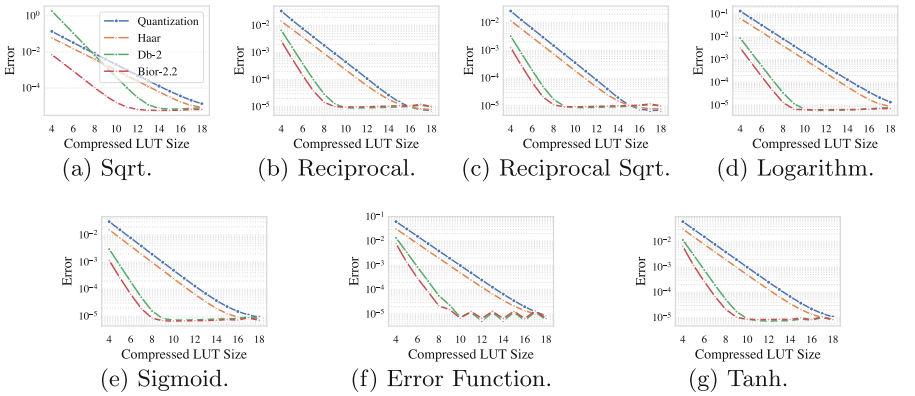


Fig. 4. Approximation errors for multiple non-linear functions for varying compressed LUT sizes with Ripple.

DWT Compression. To better understand the accuracy difference between quantization and various DWT methods (i.e., Haar and Biorthogonal), we plot the average approximation errors for various compressed LUT sizes in Fig. 4. We also evaluate a third DWT variant, Db-2 (which is the second wavelet after Haar in the Debauchies family of wavelets), to illustrate the benefits of Haar and Biorthogonal. We start with a word size of 20 and a precision of 16 and proceed to compress the tables to varying sizes from 4 to 18 bits (the x-axes). Interestingly, the errors vary based on the approximated non-linear function.

From the trend, it is clear that the Haar DWT is twice as accurate as quantization, while Db-2 and Biorthogonal DWTs are (in most cases) orders of magnitude more accurate. The difference in accuracy is most drastic around half the

word size, which is 10 in this case. As the LUT size increases, the error of quantization and Haar DWT decreases linearly to the LUT size, while the respective errors of Db-2 and Biorthogonal DWTs decay exponentially to the LUT size until we reach a LUT size of 10–14 bits. The accuracy of the Biorthogonal DWT seems almost equivalent to Db-2 DWT and better by a factor of approximately 2 after some compression threshold. This suggests that Biorthogonal DWT gives better accuracy for the functions under evaluation.

Biorthogonal's superior accuracy, coupled with the fact that Db-2 is more expensive to evaluate due to an increased number of LUT evaluations, leads us to discard it altogether. On the other hand, since Biorthogonal is more expensive than Haar, we can evaluate the trade-off based on our application. If we need higher accuracy, we can go with Biorthogonal DWT, while if speed is of the essence, we can choose Haar DWT.

We observe the trends across the board, however, there are some differences between functions. For square root (Fig. 4a), Db-2 starts performing very poorly as the compressed LUT size decreases past 14. Reciprocal (Fig. 4b) and reciprocal square root (Fig. 4c) follow this trend, except for sizes 15–18 we have that Haar and quantization are more accurate than Biorthogonal and Db-2. Logarithm (Fig. 4d), sigmoid (σ, Fig. 4e), and tanh (Fig. 4g) follow the same trend with slight boundary effects (i.e., divergent behavior around the boundaries) for compressed LUT sizes over 16. Error function (Fig. 4f) experiences boundary effects when DWT is applied an odd number of times, hence we get a zigzag pattern for the error for compressed LUT sizes between 10 and 20 bits. Other functions experience this boundary effect for compressed LUTs with larger bit sizes, resulting in a paradoxically higher error when the compressed LUT size is above 16. The boundary effects can be attributed to DWT filters wrapping around the columns of the DWT matrix.

Table 2. Overview of bounded symmetrical function optimizations. Baseline: we compress a LUT with size equal to the word size (W) to LUT size (T); Optimization: we compress LUT with size (T+3) to LUT size (T). We report the maximum absolute error.

Op.	W	k	T	Haar			Bior.		
				Baseline	Opt.	Diff.	Baseline	Opt.	Diff.
$\sigma(x)$	24	16	12	7.82e-3	2.57e-4	30×	3.60e-5	1.65e-5	2.2×
$\mathsf{erf}(x)$				3.52e-2	1.11e-3	32×	3.34e-4	3.43e-5	9.7×
$\mathsf{tanh}(x)$				3.12e-2	9.82e-4	32×	2.67e-4	3.01e-5	8.9×
$\sigma(x)$	32	20	16	7.81e-3	1.61e-5	485×	3.16e-5	1.04e-6	30×
$\mathsf{erf}(x)$				3.52e-2	6.93e-5	508×	3.16e-4	2.15e-6	147×
$\mathsf{tanh}(x)$				3.12e-2	6.15e-5	507×	2.51e-4	1.90e-6	132×

Function-Centric Compression. In Table 2 we compare Ripple's optimizations for convergent symmetrical functions with both Haar and Biorthogonal wavelet families, against a baseline without the optimizations. We start with a

word size equal to 24 bits (or 32 bits) and using the optimizations, we bring this down to 19 bits (or 23 bits), respectively, by reducing the domain of the integer part of the LUT from $[-256, 256]$ to $[0, 8]$ (or from $[-2048, 2048]$ to $[0, 8]$). This means that we now have 19 (or 23 bits) to represent a significantly smaller domain. We compare the maximum absolute error observed using the baseline and optimizations. Specifically, we observe a large decrease in the error for Haar DWT on the order of $30\times$ for 24 bits and $500\times$ for 32 bits, while the decrease for Biorthogonal is about $10\times$ for 24 bits and $100\times$ for 32 bits, respectively. The improvement in Biorthogonal is lower due to its already remarkable compression capabilities. Given the great increase in accuracy, applying these optimizations is beneficial when the computational overhead is acceptable.

6 Related Works

Ducas and Micciancio [12] first proposed the idea of using LUTs to evaluate arbitrary binary gates in FHE, while Chillotti et al. [8] extended this to evaluate arbitrary function evaluation as a tree of leveled multiplexers. Adoption was very limited, however, as it required expressing programs as deterministic automata and needed the control inputs of the multiplexers to be fresh ciphertexts (i.e., could not perform computation with them before to multiplexer). The programmable bootstrapping technique (PBS) introduced in [9] allows for efficient and general-purpose LUT evaluation. HELM [16] built on this technique and introduced a framework for automated conversion from Verilog hardware description language (HDL) to encrypted circuits. HELM employs three modes of operation, one that operates over binary gates, one that operates over integers and utilizes secure LUT evaluations, and a mixed mode that operates over binary circuits and "bridges" to integers to securely evaluate a LUT and then "bridges" back to the binary domain. However, HELM is only compatible with low-precision LUTs as bridging from an integer to bits requires multiple N to 1 LUTs. Conversely, Ripple has high precision and requires smaller LUTs to encode the same amount of information with negligible errors.

Romeo [18] and Google Transpiler [15] follow a similar approach as HELM, in that of relying on an HDL and logic synthesis. The latter, provides two different front-ends, one based on Yosys [30] and another one based on Google XLS [14]. Both works, however, rely on Boolean circuits and neither of them supports LUTs, resulting in costly operations for evaluating non-linear functions.

In a different line of work, Chung et al. [10] evaluate LUTs under FHE with the BGV, BFV, and CKKS cryptosystems. Their idea is to transform LUTs into low-degree multivariate polynomials and utilize the packing (or batching) capabilities of the aforementioned cryptosystems to simultaneously evaluate multiple LUTs. They demonstrate their approach by evaluating AES on an A-100 GPU in over 9.5 min (while batching 2048 ciphertexts). Although the amortized cost per ciphertext is under a second, the latency of this approach is far from practical. On the other hand, Ripple shows a way to accelerate LUT evaluation without sacrificing correctness and is orthogonal to the underlying cryptosystem. As a matter of fact, our techniques introduced in Ripple can be extended for BGV, BFV, and CKKS as well and further accelerate LUT evaluation.

7 Concluding Remarks

In this work, we introduce the Ripple framework that leverages different approximation techniques based on discrete wavelet transform families to reduce the number of homomorphic LUT entries in PBS. Previous works focused either on polynomial approximations or on full-size LUTs; these techniques incur high-performance overheads when high precision is required while the former also introduced non-negligible errors. Ripple, on the other hand, maintains high accuracy while it reduces the LUT sizes. Our empirical evaluations have shown significant error reduction compared to plain quantization methods across various non-linear functions, varying from square root and reciprocal computations to more elaborate sigmoid and hyperbolic tangent functions. A key benefit of Ripple is that it improves performance for several realistic benchmarks without sacrificing accuracy, compared to equivalent applications that utilize the full bit widths and incur slower LUT evaluation runtimes.

A Evaluation of Applications

Euclidean Distance. This application constitutes a formula for computing the distance between two n-dimensional points $\boldsymbol{u}$ and $\boldsymbol{v}$ in the Euclidean space. It has a plethora of applications from statistics and cluster analysis [4] to facial recognition [25] and has drawn the interest of recent FHE works [17,29]. The Euclidean distance can be computed by $d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$; however, computing non-linearities (e.g., the square root) in the encrypted domain is not a trivial task. Thus, many prior FHE works resort to computing the squared Euclidean distance and return it to the user, who needs to compute the final square root in the clear.

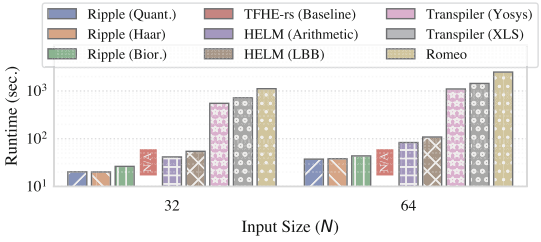


Fig. 5. Runtime comparisons for Euclidean distance between Ripple's three variants (Quantization, Haar DWT, and Biorthogonal DWT), TFHE-rs (baseline), HELM, Google Transpiler, and Romeo for vectors of 32 and 64 elements. Note that HELM, Transpiler, and Romeo only implement the squared Euclidean distance (i.e., without the square root computation). We use a word size W of 32 bits for all frameworks. Lastly, for 32 and 64 bits, TFHE-rs is not applicable (N/A) as the resources required for the LUT are impractical (see footnote 1).

For this benchmark, shown in Fig. 5, we use $\mathsf{W} = 32$ bits and use vector lengths 32 and 64 to demonstrate scalability. All Ripple variants perform the full Euclidean distance computation, while the related works compute the squared Euclidean distance and neglect the final square root calculation. We note that the TFHE-rs baseline is unable to evaluate the Euclidean distance with the required wordsize due to the astronomical cost of building 32-bit encrypted LUTs. For the Google Transpiler, we utilize both logic synthesis backends (i.e., Google XLS and Yosys), which optimize the circuit in different ways. For HELM, we utilize both LUT circuit modes (i.e., many-to-many LUTs for the arithmetic mode and a circuit of 2:1 LUTs for "lossless bidirectional bridging" or LBB). Overall, all three Ripple configurations outperform the related works in terms of latency while still taking into account the square root operation. However, as we observed in Table 1 the Haar and Biorthogonal approaches achieve significantly better approximations than the quantization variant. Notably, Haar also exhibits very competitive latencies across all non-linear functions and benchmarks.

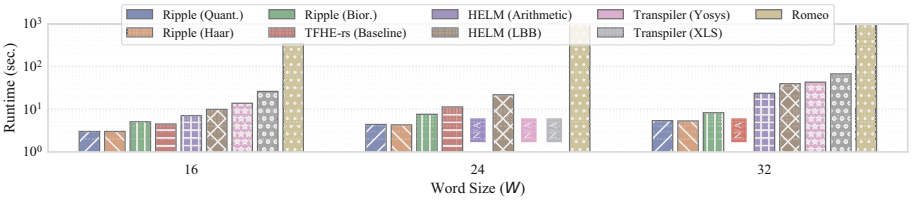


Fig. 6. Runtime comparisons for the logistic regression application for 4 attributes for word sizes of 16, 24, and 32 bits. For 24 bits, the arithmetic mode of HELM as well as both modes of the Google Transpiler are not applicable (N/A) as they rely on native word sizes. Lastly, for 32 bits, the TFHE-rs baseline is also N/A as the resources required for the LUT are impractical (see footnote 1).

Logistic Regression. Logistic Regression (LR) is a widely studied application in FHE from genome-wide association studies [23] to more generic applications [17,29] such as natural language processing [1]. This construction is well-suited to binary classification problems and is akin to a single-layer neural network with a sigmoid activation. In Ripple, we use DWT-encoded LUTs to directly compute the sigmoid activation function. The client decrypts the result, which represents the probability that the encrypted input belongs to the first class.

We utilize the Palmer penguin dataset [21], where each input consists of eight attributes that correspond to the physical characteristics of penguins (e.g., bill length, flipper length, etc.). Since logistic regression is particularly well-suited for binary classification, we remove entries in the dataset corresponding to the Chinstrap species. Figure 6 showcases our LR inference benchmark for four attributes. While our chosen dataset is composed of entries with eight attributes, we truncate it to match the dimensions used in related works. We observe that Ripple is

significantly faster than related works and also outperforms the TFHE-rs baseline using the full-bit width. The only exception is $\mathsf{W} = 16$ bits, where the baseline outperforms the Biorthogonal DWT; however, for 24 bits, the Biorthogonal DWT exhibits lower latency than the baseline.

To achieve high accuracy with our chosen dataset we utilize all eight attributes with a wordsize of 24 bits. For this binary classification, all modes achieve 100% accuracy; the baseline latency is 13.3 s per inference, while the Biorthogonal DWT variant classifies in 7.8 s. Lastly, the quantized variant that truncates half of the bits of the LUT input exhibits a latency of 6.2 s, while the Haar DWT slightly outperforms this with a latency of approximately 6 s.

References

1. Alexander Genkin, D.D.L., Madigan, D.: Large-scale Bayesian logistic regression for text categorization. Technometrics **49**(3), 291–304 (2007). https://doi.org/10.1198/004017007000000245
2. Becker, A., Ducas, L., Gama, N., Laarhoven, T.: New directions in nearest neighbor searching with applications to lattice sieving. In: Krauthgamer, R. (ed.) 27th SODA, pp. 10–24. ACM-SIAM (2016). https://doi.org/10.1137/1.9781611974331.ch2
3. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. In: Goldwasser, S. (ed.) ITCS 2012, pp. 309–325. ACM (2012). https://doi.org/10.1145/2090236.2090262
4. Carter, R.L., Morris, R., Blashfield, R.K.: On the partitioning of squared euclidean distance and its applications in cluster analysis. Psychometrika **54**(1), 9–23 (1989)
5. Cheon, J.H., Kim, A., Kim, M., Song, Y.: Homomorphic encryption for arithmetic of approximate numbers. In: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017. LNCS, vol. 10624, pp. 409–437. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70694-8_15
6. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: Faster fully homomorphic encryption: bootstrapping in less than 0.1 seconds. In: Cheon, J.H., Takagi, T. (eds.) ASIACRYPT 2016. LNCS, vol. 10031, pp. 3–33. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53887-6_1
7. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: Faster packed homomorphic operations and efficient circuit bootstrapping for TFHE. In: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017. LNCS, vol. 10624, pp. 377–408. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70694-8_14
8. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: TFHE: fast fully homomorphic encryption over the torus. J. Cryptol. **33**(1), 34–91 (2020). https://doi.org/10.1007/s00145-019-09319-x
9. Chillotti, I., Joye, M., Paillier, P.: Programmable bootstrapping enables efficient homomorphic inference of deep neural networks. In: Dolev, S., Margalit, O., Pinkas, B., Schwarzmann, A. (eds.) CSCML 2021. LNCS, vol. 12716, pp. 1–19. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78086-9_1
10. Chung, H., Kim, H., Kim, Y.S., Lee, Y.: Amortized Large Look-up Table Evaluation with Multivariate Polynomials for Homomorphic Encryption. Cryptology ePrint Archive, Paper 2024/274 (2024). https://eprint.iacr.org/2024/274

11. Dubey, S.R., Singh, S.K., Chaudhuri, B.B.: Activation functions in deep learning: a comprehensive survey and benchmark. Neurocomput. **503**(C), 92–108 (2022). https://doi.org/10.1016/j.neucom.2022.06.111
12. Ducas, L., Micciancio, D.: FHEW: bootstrapping homomorphic encryption in less than a second. In: Oswald, E., Fischlin, M. (eds.) EUROCRYPT 2015. LNCS, vol. 9056, pp. 617–640. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46800-5_24
13. Gentry, C.: A fully homomorphic encryption scheme. Ph.D. thesis, Stanford University (2009)
14. Google Research: Google XLS (2020). https://google.github.io/xls
15. Gorantala, S., et al.: A General Purpose Transpiler for Fully Homomorphic Encryption. Cryptology ePrint Archive, Report 2021/811 (2021). https://eprint.iacr.org/2021/811
16. Gouert, C., Mouris, D., Tsoutsos, N.G.: HELM: Navigating Homomorphic Encryption through Gates and Lookup Tables. Cryptology ePrint Archive, Paper 2023/1382 (2023). https://eprint.iacr.org/2023/1382
17. Gouert, C., Mouris, D., Tsoutsos, N.G.: SoK: new insights into fully homomorphic encryption libraries via standardized benchmarks. PoPETs **2023**(3), 154–172 (2023). https://doi.org/10.56553/popets-2023-0075
18. Gouert, C., Tsoutsos, N.G.: Romeo: conversion and evaluation of HDL designs in the encrypted domain. In: Proceedings of the 57th ACM/EDAC/IEEE Design Automation Conference. DAC 2020, Virtual Event, USA. IEEE Press (2020)
19. Gouert, C., et al.: Accelerated encrypted execution of general-purpose applications. Cryptology ePrint Archive, Report 2023/641 (2023)
20. Hashemi, S., Anthony, N., Tann, H., Bahar, R.I., Reda, S.: Understanding the impact of precision quantization on the accuracy and energy of neural networks. In: Proceedings of the Conference on Design, Automation & Test in Europe, DATE 2017, pp. 1478–1483. European Design and Automation Association, Leuven, BEL (2017)
21. Horst, A., Hill, A., Gorman, K.: Palmer archipelago penguins data in the palmerpenguins R package - an alternative to Anderson's irises. R J. **14**(1) (2022)
22. Juliadotter, N.V., Choo, K.K.R.: Cloud attack and risk assessment taxonomy. IEEE Cloud Comput. **2**(1), 14–20 (2015)
23. Kim, M., Song, Y., Li, B., Micciancio, D.: Semi-parallel logistic regression for GWAS on encrypted data. BMC Med. Genomics **13**(Suppl 7), 99 (2020)
24. Liu, Z., Micciancio, D., Polyakov, Y.: Large-precision homomorphic sign evaluation using FHEW/TFHE bootstrapping. In: Agrawal, S., Lin, D. (eds.) ASIACRYPT 2022, Part II. LNCS, vol. 13792, pp. 130–160. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-22966-4_5
25. Malkauthekar, M.D.: Analysis of euclidean distance and manhattan distance measure in face recognition. In: Third International Conference on Computational Intelligence and Information Technology (CIIT 2013), pp. 503–507. IET, Mumbai (2013). https://doi.org/10.1049/cp.2013.2636
26. Ribeiro, M.I.: Gaussian probability density functions: properties and error characterization. Institute for Systems and Robotics, Lisboa, Portugal (2004)
27. Unser, M., Blu, T.: Mathematical properties of the JPEG2000 wavelet filters. IEEE Trans. Image Process. **12**(9), 1080–1090 (2003)
28. Van Fleet, P.: Discrete Wavelet Transformations: An Elementary Approach with Applications. Wiley, Hoboken (2019). https://books.google.pt/books?id=jGAaxQEACAAJ

29. Viand, A., Jattke, P., Hithnawi, A.: SoK: fully homomorphic encryption compilers. In: 2021 IEEE Symposium on Security and Privacy, pp. 1092–1108. IEEE Computer Society Press (2021). https://doi.org/10.1109/SP40001.2021.00068
30. Wolf, C.: Yosys Open SYnthesis Suite (2016). https://yosyshq.net
31. Yang, P., Xiong, N., Ren, J.: Data security and privacy protection for cloud storage: a survey. IEEE Access **8**, 131723–131740 (2020)
32. Zama: TFHE-rs: A Pure Rust Implementation of the TFHE Scheme for Boolean and Integer Arithmetics Over Encrypted Data (2022). https://github.com/zama-ai/tfhe-rs