Optimizing relinearization in circuits for homomorphic encryption

Hao Chen
Microsoft Research
haoche@microsoft.com

Abstract

Fully homomorphic encryption (FHE) allows an untrusted party to evaluate arithmetic circuits, i.e., perform additions and multiplications on encrypted data, without having the decryption key.

One of the most efficient class of FHE schemes include BGV/FV schemes, which are based on the hardness of the RLWE problem. They share some common features: ciphertext sizes grow after each homomorphic multiplication; multiplication is much more costly than addition, and the cost of homomorphic multiplication scales linearly with the input ciphertext sizes. Furthermore, there is a special relinearization operation that reduce the size of a ciphertext, and the cost of relinearization is on the same order of magnitude as homomorpic multiplication. This motivates us to define a discrete optimization problem, which is to decide where (and how much) in a given circuit to relinearize, in order to minimize the total computational cost.

In this paper, we formally define the relinearize problem. We prove that the problem is NP-hard. In addition, in the special case where each vertex has at most one outgoing edge, we give a polynomial-time algorithm.

1 Introduction

Fully homomorphic encryption (FHE) is an encryption technique which allows any untrusted party to evaluate functions on encrypted data without the decryption key. As a typical application, FHE allows a client to outsource computation to an untrusted cloud. It has generated interest in fields such as health and finance, due to the need to analyze sensitive data without having access to the data itself. Since Gentry introduced the first FHE scheme in 2009, there has been a line of work that proposed new FHE schemes with improved efficiency, among which two of the most widely used schemes are \[\text{BGV14}\] and its scale-invariant counterpart \[\text{FV12}\]. Implementations of these schemes include \[\text{HIL14}, \text{CLP}\], and \[\text{AMBG}+\text{16}\]. There has been numerous work that design applications based on these schemes. Some of them \(\text{GBDL}+\text{16}, \text{BCIV17}\) evaluate machine learning models on encrypted data. Others use FHE to design secure protocols such as private information retrieval \[\text{MBFK16}\] and private set intersection \[\text{CLR17}\].

Unfortunately, in these schemes homomorphic operations are still several-orders of magnitude slower than performing the same operation on plaintexts. Therefore, any optimization in the computation time has great interest. In order to use FHE to evaluate a function, one first needs to express the function as an arithmetic circuit. The circuit is represented as a direct acyclic graph with each vertex being either an input, an output, or an arithmetic operation such as multiplication and addition. In both schemes mentioned above, a fresh ciphertext is a pair of polynomials. When a homomorphic multiplication is performed, the length of the output ciphertext grows. More precisely, if we denote the length of a ciphertext \(c\) by \(l(c)\), then \(l(c_1 \otimes c_2) = l(c_1) + l(c_2) - 1\). The
length of the result of a homomorphic addition is the maximum length of the two operands, i.e.,
\[ l(c_1 \oplus c_2) = \max\{l(c_1), l(c_2)\}. \]

Roughly speaking, the computational cost to perform a homomorphic multiplication scales linearly with its input lengths. In both schemes, we can model the amount of work it takes to perform a homomorphic multiplication between two ciphertexts \( c_1 \) and \( c_2 \) by
\[ k_m(l(c_1) + l(c_2)), \]
where \( k_m \) is some scheme-dependent constant. In FHE, there is also a squaring operation, which takes as input an encryption of \( x \) and returns an encryption of \( x^2 \). It has the same cost\(^1\) and length effect as multiplication, but only takes one input.

Homomorphic additions, on the other hand, takes much less time to perform compared to multiplication. Hence in this work we will assume that additions are “free”. For the same reason, we will adopt the common notation from the FHE literature, and denote by depth of a circuit by the largest number of multiplication vertices contained in a path.

Note that it is undesirable to let the ciphertext sizes grow, since it will increase both the computational cost and the storage burden. To control the ciphertext sizes, both schemes support a special operation called Relinearization. Effectively, relinearizing a ciphertext means reducing its length, while keeping the underlying message the same. We can use this operation to reduce the length of a ciphertext to any integer between two and its original length. The cost of relinearization scales linearly with the reduction in ciphertext length. In other words, there exists a constant \( k_r \) such that reducing the ciphertext lengths by \( i \) takes \( i \cdot k_r \) units of work.

Suppose we are given an arithmetic circuit to perform on encrypted inputs. It is now an optimization problem to decide where and how much to relinearize, in order to minimize the total amount of work, consisting of multiplication cost and relinearization cost. Previous works employ the simple strategy of relinearizing after every multiplication/squaring. In this way, the multiplication costs are kept minimal. However, this strategy is not always optimal, as we will demonstrate in Section 2.1.

1.1 Roadmap

In Section 2, we will formally describe the problem and show why this simple strategy can be sub-optimal. In Section 3, we prove that the relinearize problem is NP-hard by reducing from the knapsack problem. Finally, in Section 4, we restrict to the special case where each vertex in the circuit has at most one outgoing edge, and give a polynomial time algorithm.

1.2 Related work

The work \([CDS15]\) is an effort to find a good circuit representation of a function, in order to minimize the total computation time.

Bootstrapping is an operation that refreshes the so-called noise in FHE ciphertexts. It is an essential yet expensive operation. The two papers \([LP13]\) and \([BLMZ17]\) aim at minimizing the total number of bootstrapping operations in a circuit, while keeping the noise from overflowing in order to ensure the final result is correct. In their work, the authors implicitly assume the relinearization is done after every multiplication. Similarly, we will make a simplifying assumption that the bootstrapping time is a constant, so that it does not factor into our optimization problem.\(^1\)

\(^1\)Actually, the cost of squaring \( x \) is a constant factor smaller than multiplying \( x \) with itself. For simplicity, we will assume that the costs are equal. This simplification does not invalidate the results.
It will be interesting to combine these works in order to achieve an overall optimization that targets both operations.

**Acknowledgement** The author thanks Rebecca Hoberg and Mohit Singh for helpful discussions in preparing this work. We modify the usual definition of the arithmetic circuits to include the squaring operations in FHE.

## 2 Problem Description

To formally describe our problem, we need to properly define circuits used in FHE applications.

**Definition 1.** A (squaring-enabled) arithmetic circuit is a directed acyclic graph $G = (V, E)$, where there are three kinds of vertices: input vertices have indegree 0 and outdegree 1; output vertices have indegree $\in \{1, 2\}$ and outdegree 0; add/multiply operation vertices have indegree 2 and outdegree 1; finally, square operation vertices have indegree and outdegree both equal to 1.

We will define the relinearize problem as an integer programming problem on arithmetic circuits. For every vertex $i$, we maintain an integer variable $l_{\text{new}}(i)$ (the final length of vertex $i$ during homomorphic evaluation of $G$), and an integer variable $x_i$, which indicates the amount of relinearization at $i$. We will denote the two parents of a vertex $i$ by $p_1(i)$ and $p_2(i)$. If $i$ is a squaring vertex, then we set $p_1(i) = p_2(i)$. We denote addition vertices by $\oplus$ and multiplication/square vertices by $\otimes$. To resolve ambiguity, we make the convention that if a $\otimes$ vertex has two distinct parents, then it is understood as a multiplication; otherwise it is a squaring.

Then the relinearize problem on $G$ is

$$
\text{minimize } k_r \sum_{i \in V} x_i + \sum_{i = \otimes} k_m (l_{\text{new}}(i) + x_i),
$$

s.t.

$$
l_{\text{new}}(i) = l_{\text{new}}(p_1(i)) + l_{\text{new}}(p_2(i)) - 1 - x_i
$$

for all $i$

$$
l_{\text{new}}(i) \geq l_{\text{new}}(p_1(i)) - x_i
$$

if $i = \oplus$

$$
l_{\text{new}}(i) \geq l_{\text{new}}(p_2(i)) - x_i
$$

if $i = \otimes$

$$
x_i, l_{\text{new}}(i) \in \mathbb{Z}_{\geq 0}
$$

for all $i$

### 2.1 An example

To demonstrate the non-triviality of the relinearize problem, we consider the following circuit:

```
  ⊗
  ⊗    ⊕
  ⊕    ⊕
  ⊕    ⊕
 ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕
```

3
First, we apply the simple strategy and relinearize at every multiplication vertex. Then the total cost is equal to $12k_m + 3k_r$. Alternatively, we can choose to only relinearize the vertex $u$. Then the multiplication cost increases to $14k_m$, while the relinearization cost is $k_r$, so the total cost is $14k_m + k_r$. Comparing this with the previous cost, we see that as long as $k_r > k_m$, the simple strategy is not optimal.

3 NP-hardness of the Relinearize Problem

We prove a polynomial reduction from the knapsack problem to the relinearize problem, which establishes that the latter problem is NP-hard. First we recall the definition of knapsack problem.

**Definition 2.** Given positive integers $v_1, \ldots, v_n$, $w_1, \ldots, w_n$ and $W$. The (0-1) knapsack problem is:

$$\text{maximize } \sum_{i=1}^{n} v_i x_i$$

subject to $x_i \in \{0, 1\}$ and $\sum w_i x_i \leq W$.

For our convenience, we make some modifications to the setting of the relinearize problem. We change the inputs lengths from two to one, and we modify the equation $l(c_1 \otimes c_2) = l(c_1) + l(c_2) - 1$ to $l(c_1 \otimes c_2) = l(c_1) + l(c_2)$. One can check that under this modification, the length of every vertex is smaller by one. Hence the modified problem is equivalent to the original problem.

To prepare for the main theorem, we make some convenient definitions.

**Definition 3.** A circuit is of type $L(k)$ if it consists of one input vertex, one output vertex, and multiplication/squaring vertices, such that if the first non-input vertex length is reduced from 2 to 1, then the length of the output vertex reduces by $k$.

Figure 1 is an example of $L(7)$.

![Figure 1: an example of $L(7)$](image)

**Lemma 1.** For all integers $k \geq 1$, there exists a circuit of type $L(k)$ which has at most $2\lceil \log k \rceil$ vertices. Moreover, the cost to evaluate this circuit is bounded above by $4k_m k \lceil \log(k) \rceil$. 
Proof. If $k$ is a power of 2, we can realize $L(k)$ by a circuit that does $\log(k) + 1$ consecutive squarings. The total cost of executing the circuit is $k_m \cdot (2 + 4 + \cdots + 2k) < 4k_m k$. In general, we can start by building the circuit $L(2^{\lceil \log(k) \rceil})$. Then for every nonzero bit in the binary representation of $k$, we need to add a multiplication vertex. Since there are at most $\log(k)$ bits, we know the number of vertices is at most $2\log(k)$.

As for the evaluation cost, note that each vertex in the circuit has length bounded above by $2k$, hence evaluating it has cost bounded by $2kk_m$. The claim follows because there are at most $2\lceil \log(k) \rceil$ vertices. \hfill $\square$

Next we describe some simple ways to construct new circuits from old ones.

**Definition 4.** (1) The addition/multiplication of two circuits. Take two circuits $G_1$ and $G_2$ with unique output vertices $v_1$ and $v_2$. Then $G_1 \oplus G_2$ (resp. $G_1 \oslash G_2$) is the circuit that is the union of $G_1$ and $G_2$, plus an extra addition (resp. multiplication) vertex that has $v_1$ and $v_2$ as parents. See Figure 2 for an example.

(2) The concatenation of two circuits. Let $G_1, G_2$ be two circuits such that the number of output vertices of $G_1$ is equal to the number of inputs of $G_2$. Then we simply “feed” the outputs of $G_1$ to inputs of $G_2$. We denote the resulting circuit by $G_1 \lhd G_2$. See Figure 3 for an example.

(3) The $K$-repeat of a circuit along a subset of vertices. Let $G$ be a circuit and let $S = \{s_1, \ldots, s_k\}$ be vertices of $G$. Let $K$ be a positive integer. Then we keep the vertices $s_i$ and all their ancestors, and copy the rest of the circuit $K$ times. The resulting circuit is denoted by $G_S^{(K)}$. See Figure 4 for an example.

(4) The gluing of two circuits along a subset of vertices. Let $G_1$ and $G_2$ be two circuits and $S_1, S_2$ be subsets of their vertices, such that the subgraph of $G_1$ consisting of ancestors of $S_1$ (including vertices in $S_1$) is isomorphic to the corresponding subgraph in $G_2$. Then the gluing of $G_1$ and $G_2$ along $S_1, S_2$ is the circuit that contains the common subgraph and the disjoint union of the rest of the two graphs. We denote the new circuit by $G_1 \star_{S_1} G_2$ when $S_2$ and the isomorphism is clear from context. See Figure 5 for an example. Note that (3) is a special case of (4).

Now we are ready to state our main theorem. Consider a knapsack problem with parameters $v_i (1 \leq i \leq n), w_i (1 \leq i \leq n)$ and $W$.

**Theorem 1.** There exists a circuit $G = G(v_i, w_i, W)$, and integers $k_m, k_r$ such that

(1) $G$ has $O(\text{polylog}(v_i, w_i, W)) \cdot \text{poly}(n)$ vertices.

(2) $k_m, k_r = O(\text{poly}(v_i, w_i, W; n))$. 


Figure 3: Example of $G_1 \bowtie G_2$

Figure 4: Example of $G_S^{(K)}$ for $K = 2$ and $S = \{s_1,s_2\}$

Figure 5: Example of $G_1 *_{s_1} G_2$
Proof. By Lemma 1, the total cost of evaluating a circuit of type \(s\) could have nonzero relinearization are the evaluation is similar for \(L\) of vertices in \(h\) hence relinearizing any single vertex in this circuit has benefit bounded by 4 and \(k\) same reason, the benefit of relinearizing any vertex in any of the Suppose \(Lemma 2.\)

Suppose \(Lemma 2.\) Suppose the claim is false. Then there exists \(i\) such that \(l_i = 2\). We relinearize the vertex \(s_i\), which reduces the length of the final output in each copy of \(L(w_i)\) by \(w_i\), and the length of the output vertex of 

\[(L(w_1) \Box L(w_2)) \cdots \Box L(w_n))\]

is reduced by \(w_i\). Since \(\sum l_iw_i > W\), the length of the input vertex in each \(L(T)\) is reduced by at least one, and the cost reduction from each \(L(T)\) is at least \(k_mT\). Hence the benefit we collect from relinearizing \(s_i\) is at least \(k_mKT\), whereas the cost is \(k_r\). Since we assumed \(k_mKT > k_r\), we know relinearizing the vertex \(s_i\) reduces the total cost. This is a contradiction, since we started with an optimal solution. \(\square\)

Now we can starting proving Theorem \(\square\)
Proof. (of Theorem 1) Let $M = W + \sum_i w_i + \sum_i v_i$. We take $T = [5M \log M], k_r = 25[M \log M \log(M \log M)], K = 6[\log(M \log M)]$ and $k_m = 1$. It is easy to see that $K, T, k_r$ are of size polynomial in $W, w_i, v_i$.

One can verify that $k_m KT > k_r$ and $k_r > 4k_m(T \log T + W' \log W')$. Thus, by Lemma 3 we have $\sum_i w_i \leq W'$ if $l_i$ are the new length of $s_i$ in any optimal solution to the relinearization problem on $G^0$. This means we have the correct constraint. However, the costs are wrong: the total cost of evaluating the circuit $G^0$ is given by

$$K(\sum_i r_i l_i) + \sum_i k_r(2 - l_i) + C,$$

where as we proved in Lemma 1 $r_i \leq 4w_i \log(w_i)$. Here $C$ is the cost of evaluating all the $L(T)$ circuits plus all the $L(W')$ circuits. The fact that $C$ is a constant follows from Lemma 3.

Note that the coefficient before $l_i$ is equal to $Kr_i - k_r$, and we want to modify this coefficient to $-v_i$. First, note that

$$Kr_i - k_r \leq K4w_i \log w_i - k_r \leq (24 - 25)[M \log M \log(M \log M)] \leq -M \leq -v_i, \forall i.$$ 

Let $\lambda_i = k_r - Kr_i - v_i \in \mathbb{Z}_{\geq 0}$. We claim that there exists a circuit $L'(\lambda_i)$ of such that relinearizing its first non-input vertex reduces the total multiplication cost by $\lambda_i$. We omit the details of construction of $L'$ since it is similar to that of $L$. In particular, the $L'(\lambda_i)$ can be constructed with at most $2 \log(\lambda_i)$ vertices. We then let

$$G^i = G^0 \ast_{s_i} L'(\lambda_1), \ldots, G^i = G^{i-1} \ast_{s_i} L'(\lambda_i), \ldots, G^n = G^{n-1} \ast_{s_n} L'(\lambda_n)$$

and set $G = G^n$. Since $\lambda_i < k_r$, one can see that in any optimal solution of the relinearize problem on $G$, the vertices in $L'(\lambda_i)$ have zero relinearization. Thus, the relinearize problem on $G$ is equivalent to

$$\min \sum_{i=1}^n -v_i l_i + C', \text{ s.t. } l_i \in \{1, 2\} \text{ and } \sum_{i=1}^n w_i l_i \leq W + \sum w_i,$$

which is equivalent to

$$\max \sum_{i=1}^n v_i l_i, \text{ s.t. } l_i \in \{1, 2\} \text{ and } \sum_{i=1}^n w_i l_i \leq W + \sum w_i.$$ 

This proves part (3) of Theorem 1. Part (1) is clear since the number of vertices in $G$ is bounded by $2K(\log(T) + \log(W')) + \sum_{i=1}^n \log(w_i) + 2\sum_{i=1}^n \log(\lambda_i)$. Hence it is logarithm in the parameters $v_i, w_i, W$ and linear in the number of variables $n$. For (2), note that we set $k_m = 1$, so it suffices to prove it for $k_r$. By construction, $k_r$ is also bounded by a polynomial in $v_i, w_i, W$. This completes the proof. \qed

Corollary 1. The relinearize problem is NP-hard.
4 An Simple Case

Assume we are in the situation where each non-input vertex in the circuit has two inputs and at most one output. In this case, we have a polynomial time algorithm for the relinearize problem. For a vertex $i$, define $M(i, \ell)$ to be the minimal cost to compute the circuit up to vertex $i$, so that the new length of $i$ is $\ell$.

Recall that $p_1(i)$ and $p_2(i)$ denote the parents of $i$. If $i$ is a multiplicative vertex, we have

$$M(i, \ell) = \min_{\ell_1, \ell_2} \{ M(p_1(i), \ell_1) + M(p_2(i), \ell_2) + k_r(\ell_1 + \ell_2 - \ell) + k_m(\ell_1 + \ell_2) \}.$$  

If $i$ is an addition vertex, we have

$$M(i, \ell) = \min_{\ell_1, \ell_2} \{ M(p_1(i), \ell_1) + M(p_2(i), \ell_2) + k_r(\max\{\ell_1, \ell_2\} - \ell) \}.$$  

Here it is important that the vertices all only have a single output, since otherwise $p_1(i)$ and $p_2(i)$ might have a common ancestor, in which case relinearizing this ancestor might benefit both of them.

Claim 1. Suppose $N = |V| \geq 2$. Then in the above formulae, it suffices to take the minimum over range $2 \leq \ell_1, \ell_2 \leq N$.

Proof. For the input vertices, the lengths is at most 2. For any non-input vertex $v$, we prove inductively that its length cannot exceed its number of ancestors. The length is at most $l(p_1(v)) + l(p_2(v)) - 1$, and by inductive hypothesis, both $l(p_1(v))$ and $l(p_2(v))$ are at most their number of ancestors (or plus one if it happens to be an input vertex). That is, $l(p_1(v)) + l(p_2(v)) - 1 \leq n_1 + n_2 + 1 = n$. Here $n_1, n_2, n$ denote the number of ancestors for $p_1(v), p_2(v), v$, respectively.

Now our algorithm proceeds as follows. We traverse the $N$ vertices. At each vertex, we compute $M(i, \ell)$ for $O(N)$ values of $\ell$, and each computation requires $O(N^2)$ operations. Thus the total running time is $O(N^4)$. Finally, the optimal cost is given by $\min_{2 \leq \ell \leq N} M(v, \ell)$, where $v$ is the output node of the graph $G$.

5 Conclusion and Future Work

Fully homomorphic encryption evaluates boolean circuits, and relinearization is a standard technique to reduce the ciphertext sizes after evaluation. In this paper, we consider the goal of optimizing where and how much to perform the relinearization operation in any given circuit, in order to minimize the total computational cost. We formalized it as a discrete optimization problem, and proved that the problem is NP-hard. In the special case where every node has at most one output node, we give a polynomial time algorithm.

For future directions, it is of interest to design fast approximate algorithms for the relinearization problem. Also, one can aim at optimizing specific circuits that appear in the literature for applications of FHE. Examples include components of the AES encryption/decryption circuit and machine learning models such as logistic regression or neural network.

References

[AMBG+16] Carlos Aguilar-Melchor, Joris Barrier, Serge Guelton, Adrien Guinet, Marc-Olivier Killijian, and Tancrede Lepoint. Nfllib: Ntt-based fast lattice library. In Cryptographers’ Track at the RSA Conference, pages 341–356. Springer, 2016.
[BCIV17] Joppe W Bos, Wouter Castryck, Ilia Iliashenko, and Frederik Vercauteren. Privacy-friendly forecasting for the smart grid using homomorphic encryption and the group method of data handling. In *International Conference on Cryptology in Africa*, pages 184–201. Springer, 2017.

[BGV14] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (Leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT)*, 6(3):13, 2014.

[BLMZ17] Fabrice Benhamouda, Tancrède Lepoint, Claire Mathieu, and Hang Zhou. Optimization of bootstrapping in circuits. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2423–2433. SIAM, 2017.

[CDS15] Sergiu Carpov, Paul Dubrulle, and Renaud Sirdey. Armadillo: a compilation chain for privacy preserving applications. In *Proceedings of the 3rd International Workshop on Security in Cloud Computing*, pages 13–19. ACM, 2015.

[CLP] Hao Chen, Kim Laine, and Rachel Player. Simple encrypted arithmetic library-SEAL v2.

[CLR17] Hao Chen, Kim Laine, and Peter Rindal. Fast private set intersection from homomorphic encryption. *IACR Cryptology ePrint Archive*, 2017:299, 2017.

[FV12] Junfeng Fan and Frederik Vercauteren. Somewhat practical fully homomorphic encryption. *IACR Cryptology ePrint Archive*, 2012:144, 2012.

[GBDL+16] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pages 201–210, 2016.

[HS14] Shai Halevi and Victor Shoup. Algorithms in helib. In *International Cryptology Conference*, pages 554–571. Springer, 2014.

[LP13] Tancrède Lepoint and Pascal Paillier. On the minimal number of bootstrappings in homomorphic circuits. In *International Conference on Financial Cryptography and Data Security*, pages 189–200. Springer, 2013.

[MBFK16] Carlos Aguilar Melchor, Joris Barrier, Laurent Fousse, and Marc-Olivier Killijian. Xpir: Private information retrieval for everyone. *Proceedings on Privacy Enhancing Technologies*, 2016:155–174, 2016.