

# CAS 705, Computability and Complexity

## Final Exam

Xiao Shu (0901994)  
shux@mcmaster.ca

December 11, 2009

### Question 1

As  $W = \{w_1, \dots, w_n\}$  is a base of  $\mathcal{L}$ ,  $w_1 = 1 \cdot w_1 + 0 \cdot w_2 + \dots + 0 \cdot w_n \in \mathcal{L}$ , thus there exists  $a_1, \dots, a_n \in \mathbb{Z}$  such that  $w_1 = a_1 \cdot v_1 + \dots + a_n \cdot v_n$ , since  $V = \{v_1, \dots, v_n\}$  is a base of  $\mathcal{L}$  as well. Similarly, there exists an integer linear combination of  $W$  for each  $v_i$ , i.e., there must be a matrix  $A$  over  $\mathbb{Z}$  such that  $M_W = M_W I = M_V A$ , where  $I$  is the identity matrix. For the same reason, there also exists a matrix  $B$  over  $\mathbb{Z}$  such that  $M_W B = M_V$ , hence,  $M_W = M_W B A$ , which implies  $\det(BA) = \det(B) \cdot \det(A) = 1$ . Since  $A, B \in \mathbb{Z}^{n \times n}$ ,  $\det(B)$  and  $\det(A)$  have to be integers, therefore,  $\det(A) = \pm 1$ .

### Question 2

Since  $M_V$  is constructed with independent column vectors,  $M_V$  is a nonsingular matrix and  $M_V^{-1}$  exists. Now, let  $w' = M_V^{-1}w$ ,  $v' = \lfloor w' \rfloor$  and  $t' = w' - v'$ . Obviously,  $w = t + v$ , where  $t = M_V t'$  and  $v = M_V v'$ , and by the definition of  $\mathcal{F}$  and  $\mathcal{L}$ ,  $t$  is a vector in  $\mathcal{F}$  and  $v$  is a vector in  $\mathcal{L}$ . Next, we show the uniqueness of  $t$  and  $v$  with regards to  $V$ . Suppose for another pair of vector  $s, u$ , where  $s \in \mathcal{F}$  and  $u \in \mathcal{L}$ , we have  $w = s + u$ , then  $t + v = s + u$  and  $t' - s' = u' - v'$  where  $s' = M_V^{-1}s$  and  $u' = M_V^{-1}u$ . However, since  $u' - v'$  is a vector over  $\mathbb{Z}$ ,  $t' - s'$  has to be over  $\mathbb{Z}$  as well, thus, considering that each element of  $t' - s'$  must be in  $(-1, 1)$ , the only possible value of  $t' - s'$  is a zero vector, which implies  $t = s$  and  $v = u$ .

### Question 3

By the result of Question 1, for any two bases  $W, V$  of  $\mathcal{L}$ , there exists a vector  $A$  such that  $|\det(A)| = 1$  and  $M_W = M_V A$ . Thus,  $|\det(M_W)| = |\det(M_V)| \cdot |\det(A)| = |\det(M_V)|$ , which means  $\det(L)$  is well defined, as the determinants of all the bases have the same absolute value.

### Question 4

Since  $w = M_v t$ ,  $t = M_v^{-1}w$ , thus  $v = M_v \lfloor t \rfloor = M_v \lfloor M_v^{-1}w \rfloor$ . Then, for the first case,  $v = (53159, 81818)$ , and for the second case,  $v = (56405, 82444)$ .

## Question 5

The solution with basis  $V$  finds a closer vector to  $w$  than the solution with basis  $V'$  does, and  $H(V) \approx 0.977$  is larger than  $H(V') \approx 0.077$ . It appears that, for a basis  $W$  of a given lattice  $\mathcal{L}$ , the larger  $H(W)$  is, the more likely it is to find a close vector in  $\mathcal{L}$  with  $W$ . However, we notice that this claim does not always true, and we believe that there exists a better indicator function, which estimates how close the vector that can be found with a certain basis.

Before going through the indicator we propose, we first discuss the meaning of  $H$ . By the definition of  $H$ ,

$$\begin{aligned}
 H(W) &= \left( \frac{\det(\mathcal{L})}{\|w_1\| \dots \|w_n\|} \right)^{\frac{1}{n}} \\
 &= \left( \frac{|\det(M_W)|}{\|w_1\| \dots \|w_n\|} \right)^{\frac{1}{n}} \\
 &= \left| \det(M_W) \cdot \det \left( \begin{bmatrix} \frac{1}{\|w_1\|} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\|w_n\|} \end{bmatrix} \right) \right|^{\frac{1}{n}} \\
 &= \left| \det \begin{bmatrix} w_{1,1}/\|w_1\| & w_{2,1}/\|w_2\| & \dots & w_{n,1}/\|w_n\| \\ w_{1,2}/\|w_1\| & w_{2,2}/\|w_2\| & \dots & w_{n,2}/\|w_n\| \\ \dots & \dots & \dots & \dots \\ w_{1,n}/\|w_1\| & w_{2,n}/\|w_2\| & \dots & w_{n,n}/\|w_n\| \end{bmatrix} \right|^{\frac{1}{n}}
 \end{aligned}$$

where  $W = \{w_1, \dots, w_n\}$  and  $w_{i,j}$  is the  $j$ -th element of vector  $w_i$ . Since the matrix on the right side is normalized, i.e., each column is a unit vector, and we know that absolute value of determinant is equal to the volume of the parallelepiped spanned by those vectors, it implies that the indicator  $H$  is only correlated with the angle between each pair of vectors and independent of the length of the vectors. If these unit vectors are more orthogonal to each other, the volume of the parallelepiped and  $H(W)$  will be larger.

Now, the question is, what is the relationship between  $H$  and the result of the CVP algorithm? If we look the CVP algorithm from aspect of the vector space given by  $W$ , what it does is to find the closest integer point to a given point in the vector space. However, when we evaluate the quality of the results produced by the algorithm, we measure the distance between the two points in the euclidean space. Thus, the fundamental reason that leads to the difference in the results is due to the difference between the vector space and the euclidean space. In other words, If the vector space given by  $W$  is similar to the euclidean space, i.e., vectors in  $W$  are pseudo-orthogonal, the close vector found by the CVP algorithm is indeed close enough in the euclidean space since, in this case, the CVP algorithm and the evaluation method have an agreement on the definition of distance. Therefore, large  $H(W)$  implies the orthogonality of vectors in  $W$ , which helps in searching for the closest vector in the euclidean space.

The above analysis shows that  $H$  can somehow indicates the quality of a basis in term of finding the closest vector, and we are convinced that it is good enough in the lattice based cryptography applications, however, the connection between  $H$  and the quality of a basis does not seem to be clear, and there exists counter example showing that  $H$  might not always be able to distinguish the qualities of two bases correctly. For example, in Table

$i$	$W_i$	$H(W_i)$	Probability of getting the closest vector	Expected distance to the closest vector
1	$\{(1, 0), (0, 1)\}$	1	1	0.38
2	$\{(1, 6), (2, 13)\}$	0.177	0.077	3.61
3	$\{(3, 4), (5, 7)\}$	0.146	0.143	2.37

Table 1: Compare indicator  $H$  to other indicators.

1,  $W_1, W_2, W_3$  are bases of lattice  $\mathcal{L}$ , which is the set of integer points on a euclidean plain. It is obvious that  $W_1$  is the “good” basis among the three, and it shows clearly by the indicator  $H$ . However, that is not the case for  $W_2$  and  $W_3$  — although  $H(W_2) > H(W_3)$ , the other indicators in the table do not agree that  $W_2$  is better.

Here, in Table 1, we present another two indicators which estimate the quality of a basis in solving CVP. The first one is the probability that a vector found by the CVP algorithm just happens to be the closest vector in the euclidean space. Obviously, if it is highly probable to find the closest vector with a basis, the basis is “good”. The second indicator is the expected distance from a uniformly distributed random vector to its closest vector with regard to the basis. A basis is “good”, if, with which we can find vector that is not too far from the given vector on average.

Both of these two indicators have a more straightforward and clearer connection with the quality of a basis than  $H$  does, however, we still need to know what the mathematical relationship between a indicator and a basis to make it useful. In this work, we focus our analysis on the second indicator, the expected distance, as we believe it reflects the hardness of breaking the lattice based cryptography better.

Interestingly, due to the symmetry, the expected distance from a random point to its closest vector with regard to a basis is equivalent to the expected distance from a random point in the fundamental domain to the centre of the domain. More formally, the expected distance  $G(W)$  in basis  $W = \{w_1, \dots, w_n\}$ , can be written as,

$$\begin{aligned}
G(W) &= \int_{v \in \mathcal{F}(W)} \frac{1}{\det(W)} \|v - \frac{1}{2} \sum_{i=0}^n w_i\| dv \\
&= \int_{Wu \in \mathcal{F}(W)} \frac{1}{\det(W)} \|Wu - \frac{1}{2} \sum_{i=0}^n w_i\| dWu \quad (\text{Let } v = Wu) \\
&= \int_{u \in [0,1]^n} \|W(u - \frac{1}{2})\| du
\end{aligned}$$

The analytic solution to this formula is not trivial. It is not easy even for two dimensional case. However, we can estimate the result with a simple randomized algorithm as follows,

```

1 def approximate_g(W, m):
2     n ← dimension(W)
3     s ← 0
4     for i ∈ [1...m]:
5         u ← random([0, 1]n) - 1/2
6         s ← s + ||Wu||
7     return s / m

```

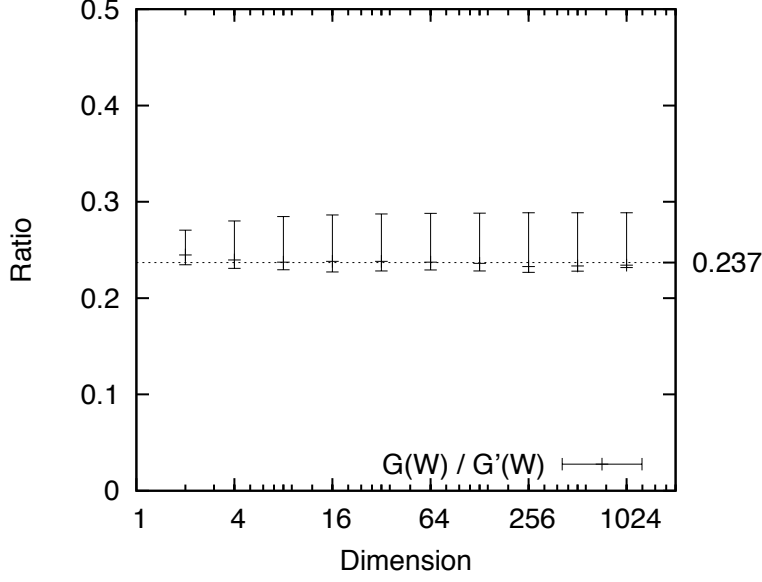


Figure 1: The ratio of  $G(W)$  and  $G'(W)$  over different dimensions.

We did a group of experiment on different bases from 2-dimensional to 1024-dimensional with this algorithm, and we notice that the value of  $G(W)$  is easy to predict with  $W$ . Suppose  $G'(W) = \sqrt{\|w_1\|^2 + \dots + \|w_n\|^2}$ , then  $G(W)/G'(W) \approx 2.37$ . Figure 1 shows the result, where the error range is small and the ratio does not affect by the dimension of  $W$  much. This gives us a strong impression that  $G(W)$  can be written as a function of  $G'(W)$ , however, with the time constrain of the project, we have not been able to find an elegant analytic solution.

In practice,  $G'(W)$  seems to be good indicator to the quality of a basis for lattice based cryptography applications. Suppose an encoded message  $e = mW + r$ , where  $m$  is the plain text,  $W$  is a “bad” basis and  $r$  is a short random vector, and suppose  $e'$  is the closest vector to  $e$  with regards to  $W$ . Since the sum of any two edges is greater than the third edge for a triangle in euclidean space, we have,

$$\begin{aligned} \|e' - mW\| + \|r\| &\geq \|e' - mW - r\| = \|e' - e\| \\ \|e' - mW\| &\geq \|e' - e\| - \|r\| \end{aligned}$$

Then, the mathematical expectation of the distance from  $e'$  to the actual closest vector in euclidean space,  $mW$ , is,

$$\begin{aligned} E[\|e' - mW\|] &\geq E[\|e' - e\|] - E[\|r\|] \\ E[\|e' - mW\|] &\geq G(W) - E[\|r\|] \quad (\text{By the definition of } G) \\ E[\|e' - mW\|] &\geq 0.237 * G(W) - E[\|r\|] \end{aligned}$$

Since  $E[\|r\|]$  is a small constant, if  $G'(W)$  is large,  $E[\|e' - mW\|]$  will be large as well. In other words, on average, the CVP algorithm can only find a lattice vector that is far from the actual  $mW$  if  $G'(W)$  is sufficiently large, thus, the algorithm does not help in breaking  $mW$ .

In conclusion, we find that  $H$  could somehow indicate the quality of a basis, however, we believe  $G'$ , which provides a bound for lattice based cryptography applications, is a better indicator.

## Question 6

If we can get back the plain text  $m$  from an encoded message  $e = mW + r$ , we actually also find the solution to the closest vector problem to vector  $e$ , i.e.,  $mW$ . Considering that the CVP is NP-hard [1], it is unlikely to compute the plain text  $m$  efficiently with a bad basis in the first place.

It is not easy to conjecture a meaningful dimension without knowing how lattice based cryptosystems are implemented and all the attacking techniques in practice. In general, since a top super-computer can do  $10^{15}$  operations per second, that is about  $10^{20}$  operations a day,  $\log(10^{20}) \approx 66$  is reasonable dimension to prevent brutal force attack by a super-computer in a day. On the other hand, encoding or decoding a message needs  $O(n^2)$  operations to do the matrix and vector multiplication, thus, it might be too slow on a personal computer or embedded device if  $n \geq 10^4$ . Another factor that limits the feasible size of the dimension is the size of public key, which is also  $O(n^2)$ . If  $n$  is larger than  $10^3$ , it would take more than one second to transfer the key over a typical broadband network.

## Question 7

Since  $e = mW + r$  and  $e' = mW + r'$ , Eve knows the arithmetical difference between the two random perturbation, i.e.,  $r - r' = e - e'$ . These information could greatly reduce the difficulty of finding  $mW$ . Suppose  $mW = \{v_1, \dots, v_n\}$ ,  $e = \{e_1, \dots, e_n\}$  and  $e'_i, r_i, r'_i$  are defined similarly. By the definition of  $e$ ,  $e_i = v_i + r_i$ , then  $v_i = e_i - r_i$ . Since  $r$  is a short vector, there exists a small positive constant  $c$ , such that,  $e_i - c \leq v_i \leq e_i + c$ . Meanwhile, as  $e'_i = v_i + r'_i$ , we also know that  $e'_i - c \leq v_i \leq e'_i + c$ . Combining these two inequations together, we have,

$$\max\{e'_i, e_i\} - c \leq v_i \leq \min\{e'_i, e_i\} + c$$

Thus, range need to search for  $mW$  becomes smaller.

We now try to answer a more general question — if the plain text message are sent repeatedly for  $k$  times, how much easier it will be to break the text? Since the width of the range for possible  $v_i$  after  $k$  same plain messages is,

$$\begin{aligned} & (\min\{e_i^{(1)}, \dots, e_i^{(k)}\} + c) - (\max\{e_i^{(1)}, \dots, e_i^{(k)}\} - c) \\ = & 2c - (\max\{e_i^{(1)}, \dots, e_i^{(k)}\} - \min\{e_i^{(1)}, \dots, e_i^{(k)}\}) \end{aligned}$$

suppose the original range is over  $[0, 1]$ , then the problem is equivalent to calculate the expected value of  $1 - (\max\{X_j\} - \min\{X_j\})$ , where  $X_j$  are independent random variable uniformly distributed over  $[0, 1]$  for each  $j \in \{1 \dots k\}$ . By the definition of mathematical

expectation,

$$\begin{aligned}
& E[\max\{X_j\} - \min\{X_j\}] \\
&= \int_0^1 \int_y^1 k(k-1)(x-y)^{k-2} \cdot (x-y) dx dy \\
&= k(k-1) \cdot \int_0^1 dy \cdot \int_y^1 (x-y)^{k-1} dx \\
&= k(k-1) \cdot \int_0^1 dy \cdot \left[ \frac{(x-y)^k}{k} \right]_y^1 \\
&= k(k-1) \cdot \int_0^1 \frac{(1-y)^k}{k} dy \\
&= k(k-1) \cdot \left[ \frac{-(1-y)^{k+1}}{k(k+1)} \right]_0^1 \\
&= \frac{k-1}{k+1}
\end{aligned}$$

Thus, the expected width of possible range after  $k$  duplicated plain messages is,

$$1 - E[\max\{X_j\} - \min\{X_j\}] = 1 - \frac{k-1}{k+1} = \frac{2}{k+1}$$

Specially, when  $k = 2$ , i.e., the case in this question, the possible range for each  $v_i$  is only two third of the original range. The message is much easier to break than before. Considering that repeated messages are fairly common in communication, it is better to add some random “salt” into the beginning or end of each message to reduce the possibility of repeat.

## Question 8

With the result from last question, the possibilities for  $r$  is,

$$\begin{aligned}
& (4 - (\max\{-9, -6\} - \min\{-9, -6\}) + 1) \\
& \times (4 - (\max\{-29, -26\} - \min\{-29, -26\}) + 1) \\
& \times (4 - (\max\{-48, -51\} - \min\{-48, -51\}) + 1) \\
& \times (4 - (\max\{18, 20\} - \min\{18, 20\}) + 1) \\
& \times (4 - (\max\{48, 47\} - \min\{48, 47\}) + 1) \\
&= 2 \times 2 \times 2 \times 3 \times 4 \\
&= 96
\end{aligned}$$

## Question 9

Since,  $e = mW + r$  and  $e' = m'W + r$ , Eve knows the arithmetical difference of the two plain messages  $m - m' = (e - e')W^{-1}$ . This is extremely dangerous in practice, because usually, some parts of a plain message are well-known public knowledge or very predictable. For example, a modern HTML file often starts with `<!DOCTYPE HTML...`. This means if  $m$  is

exposed, which is not uncommon for some type of conversation,  $m'$  is also leaked as long as we know  $m - m'$ . Even if this is not the case, i.e., we do not know one of the plain message,  $m - m'$  still gives out too much important information for guessing. For example, suppose the both  $m, m'$  are in English and encoded in ASCII, since the code of 'a'..'z' is 97..122 and white-space is 32, whenever  $m_i - m'_i \geq 97 - 32$ , it is highly probable that  $m'_i$  is a white-space and  $m_i$  is a lowercase letter due to the fact that lowercase letter and white-space are dense in English. For other types of plain text, the similar statistical results could also exist, thus, for the sake of security, the random perturbation should never be reused.

## References

- [1] Peter van Emde Boas. Another np-complete problem and the complexity of computing short vectors in a lattice iterative decoding of two-dimensional hidden markov models. Technical report, Mathematische Instituut, University of Amsterdam, 1981.