SECURITY OF NEURAL CRYPTOGRAPHY

R. Mislovaty [1]*, E. Klein*[1]*, I. Kanter*[1]*, W. Kinzel*[2]

[1]Bar-Ilan University, Minerva Center and Department of Physics
Ramat-Gan 52900, Israel
[2]Institut für Theoretische Physik, Universität Würzbur
Am Hubland 97074, Würzbur, Germany

ABSTRACT

In this paper we analyze the security of Neural Cryptography, a novel key-exchange protocol based on synchronization of Neural Networks[1]. Various attacks on this protocol were suggested by Shamir et al., and the protocol was shown to be secure against them[2]. A new attack strategy involving a large number of cooperating attackers, that succeeds to reveal the encryption key was recently found[3].

1. INTRODUCTION

The learning abilities of artificial neural networks with complicated input-output relations have attracted a lot of attention in the recent years [4, 5, 6]. The ability of such networks to synchronize by mutual learning was extensively studied as well [7, 8]. Naturally, this raised the question of comparison between synchronization and learning abilities of networks with identical architectures. Special attention was given to Tree Parity Machines (TPM) with K hidden units. This interest was driven by fast synchronization ability of these networks, versus their much longer learning time. It was suggested to use this property in order to construct a novel cryptographic scheme, which is radically different from standard ones. The idea is to exploit the fast synchronization ability of TPM, and use their synchronized weight vectors as an encryption key. The input/output relations would be transmitted over a public channel. Since learning times are much longer, a third eavesdropping network trying to reveal the key by learning from the transmissions would be left behind [1]. Such protocol based on neural networks would use only simple arithmetics, and no injective trapdoor functions, whereas traditional cryptography is based on number theoretic functions, and their irreversibility when using large prime integers. The theoretical possibility of constructing a secure cryptographic scheme not using such trapdoors is an open question in cryptography. Therefore, a discussion of Neural Cryptography's security is presented in this paper, along with some new results that were recently discovered.

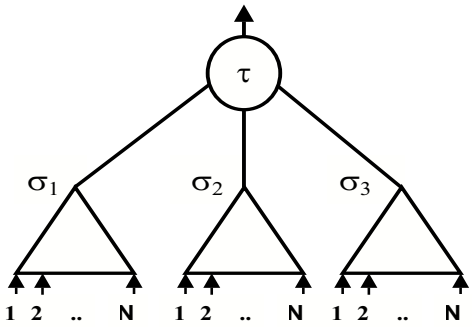

Figure 1: A tree parity machine with K=3

2. THE MODEL

The networks are Tree Parity Machines (TPM) with K hidden units $\sigma_i = \pm 1, \quad i = 1, ..., K$ feeding a binary output, $\tau = \prod_{i=1}^{K} \sigma_i$, as shown in Figure 1. The networks consist of a discrete coupling vector $\mathbf{w}_i = w_{i1}, ..., w_{iN}$ and disjointed sets of inputs $\mathbf{x}_i = x_{i1}, ..., x_{iN}$ containing N elements each. The input elements are random variables $x_{ij} = \pm 1$. Each component of the weight vector can take certain discrete values $w_{ij} = \{\pm L, \pm(L-1), ..., \pm 1, 0\}$ and is initiated randomly from a flat distribution.

The local field in the ith hidden unit is defined as

$$h_i = \mathbf{w}_i \cdot \mathbf{x}_{i,} \tag{1}$$

and the output in the ith hidden unit is the sign of the local field. The output of the tree parity machine is therefore given by

$$\tau = \prod_{i=1}^{K} sign(h_i) = \prod_{i=1}^{K} \sigma_i \tag{2}$$

During the mutual learning process, the two machines A and B, exchange their output values $\tau^{A/B}$. They update their weights using the hebbian learning rule, only in case

0-7803-8715-5/04/$20.00 ©2004 IEEE.

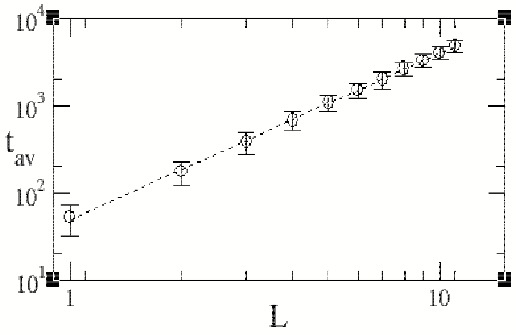

Figure 2: The average synchronization time t_{av} as function of L, for TPM with $K = 3$. Simulation results obtained using $N = 10^3$, averaged over 10^3 samples. The equation of the fitted line is $t_{av} = 50L^{1.91}$.

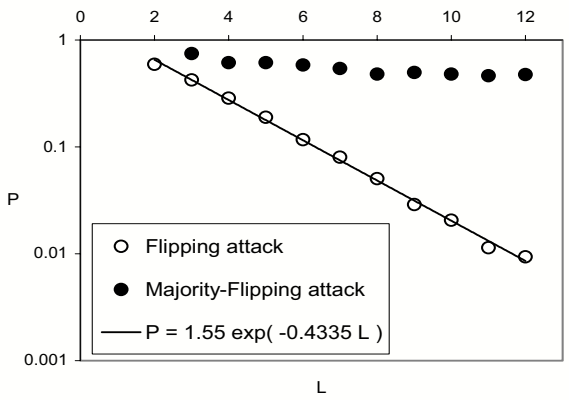

Figure 3: The attacker's success probability P as a function of L, for the flipping attack and the majority-flipping attack. Simulations were carried out using $N = 10^3$, M=100, and averaged over 1000 samples. Waiting time was equal to about 1/3 of total synchronization time.

their outputs agree, and only in hidden units which agree with the output

$$\mathbf{w}_i^A(t+1) = \mathbf{w}_i^A(t) + \mathbf{x}_i\tau^A\theta(\tau^A\sigma_i^A)\theta(\tau^A\tau^B) \qquad (3)$$
$$\mathbf{w}_i^B(t+1) = \mathbf{w}_i^B(t) + \mathbf{x}_i\tau^B\theta(\tau^B\sigma_i^B)\theta(\tau^A\tau^B).$$

This leads them to a parallel state in which $W^A = W^B$.

The synchronization is due to existence of the absorbing boundaries -L and L, as can be explained by the theory of bounded random walks. When updated, the wight element w_{ij} is moved in the direction determined by the product $\tau^{A/B}x_{ij}$, the same for A and B by its definition. If a weight tries to step beyond the boundary, it remains stuck there, while the other party's weight moves towards it. Synchronization time, therefore, should scale with the weights depth L like L^2. This observation is supported by simulation results shown in Figure 2.

Obviously parties weights approach only if the update step happens simultaneously in the same hidden unit - an attractive step. Since there is 2^{K-1} degeneracy of $\{\sigma_i\}$ for each τ value, repulsive steps (update of different hidden units) are present as well. However, it was shown that attractive forces overcome repulsive for TPM with $K = 3$ hidden units, and synchronization of A and B is achieved [8]. In the rest of the paper we will consider a TPM with K=3 hidden units.

3. SECURITY ANALYSIS

The attacker, C, tries to learn the weight vector of one of the two machines, say A. The values of N, and L are public, as well as all the transmissions through the cannel: inputs x_{ij} and outputs $\tau^{A/B}$. The information C lacks in each learning step are the values $\{\sigma_i\}$ of A's hidden units, i.e which of the 2^{K-1} possible updating scenarios A performs. Unlike the simple teacher-student scenario [9, 10], the teacher's weights in this case are time-dependant, therefore the attacker must use some attack strategy in order to follow the teacher's steps. The most trivial is the *The Naive Attack* in which the attacker imitates one of the parties: when $\tau^A = \tau^B$ and learning step takes place, the naive attacker performs learning as well, using his own internal representations $\{\sigma_i^C\}$. Simulations and analytic results show that this attacker's success rate drops very quickly when increasing L. Clearly, the naive attacker fails to reveal the correct internal representations of the imitated party, and does too many repulsive steps. The most successful attack suggested so far [11] is the *Flipping Attack* (Geometric attack), in which the attacker imitates one of the parties, but in steps in which his output disagrees with the imitated party's output ,$\tau^C \neq \tau^A$, attacker surely knows that either one or all three of his hidden units are wrong (the first option is more probable). In order to achieve $\tau^C = \tau^A$ he negates ("flips") the sign of one of his hidden units. Since $\sigma = sign(h)$, the unit most likely to be wrong is the one with the minimal $|h|$, therefore that is the unit which is flipped. This strategy results a great improvement in the attacker's success. The Flipping Attacker's success rate, measured in simulations, is shown in Figure 3. It can be seen that the success rate is quite high for all L values presented, but it drops exponentially as L increases. On the other hand parties synchronization time increases like L^2, and therefore we conclude that in the limit of large L values the suggested neural cryptography scheme is secure against the Flipping Attack [2, 12]. These results were supported analytically as well [8].

The attackers mentioned above try to imitate the parties, each using different heuristics. They use an ensem-

ble of independent attackers. It has been shown that among a group of Ising vector students which perform learning, and have an overlap R with the teacher, the best student is the center of mass vector (which was shown to be an Ising vector as well) which has an overlap $R_{cm} \propto \sqrt{R}$, for $R \in [0:1]$[13]. Therefore letting the attackers cooperate throughout the process may be to their advantage. A new attack strategy was presented recently, suggesting to use a large group of M attackers that cooperate with each other. We can understand why such a strategy might be more successful from the following argument: The flipping attacker's probability to guess correctly A's internal representation is some function $P_{correct}(\rho)$ of its overlap ρ with A, starting from $P_{correct}(\rho = 0) = 0.25$. Assume we have a group of M *independent* flipping attackers, each having overlap ρ with A. They will split into 4 groups, one for each possible internal representation. Since $P_{correct} > 0.25$ for all $\rho > 0$, the number of attackers having the correct internal representation, $M \cdot P_{correct}$, will be bigger than the number of attackers in the other 3 groups, for all $\rho > 0$. Therefore, the internal representation resulted from the majority discission of M *independent* flipping attackers would *always* be the correct one! From this argument we conclude that the attack should use $M >> 2^{K-1}$ attackers, which would simultaneously develop an overlap with the parties, trying to remain as independent as possible. The Majority Flipping Attack procedure is to start from independent flipping attackers and let them act separately for some initial number of time steps. Then, the majority procedure is applied: we count how many attackers have each of the 4 possible internal representation, and assign the majority's internal representation to all the M attackers. To prevent the similarity between the attackers from developing too quickly, this majority procedure is applied only on even time steps. However, the attackers make many coherent moves, and unavoidable overlap is developed between them as well. Therefore we do not have a group of *independent* attackers, but of attackers with an overlap between them. This overlap diminishes the efficiency of the attack, and it is not *always* successful as a majority attack of M independent attackers would be.

The simulation results for the Majority Flipping Attack's success rate are shown in Figure 3, next to the results of the Flipping Attack. Unlike the Flipping Attack, the Majority Flipping Attack's success rate remains constant around 0.5, even when L is increased. Therefore, the Majority Flipping Attack manages to break the key exchange protocol, leaving the presented cryptographic scheme insecure. These results were supported analytically as well [3].

4. CONCLUSIONS

We have presented a bridge between the theory of neural networks and cryptography. An encryption scheme based on neural networks was presented, and its security was discussed in details. Particularly, we presented a new attack strategy which acts through many cooperating attackers, and leaves the presented scheme insecure. There is still an open question wether a different model can be developed, which would remain secure against the majority attack. Our understanding for the reason of the majority attack's success implies that we should be looking for algorithms where the overlap between the attackers would develop much faster than their overlap with the parties. This might be achieved by using larger K values, or some other modifications. These models are still under consideration.

5. REFERENCES

[1] I. Kanter, W. Kinzel and E. Kanter, Europhys. Lett., **57**, 141 (2002).

[2] R. Mislovaty, Y. Perchenok, I. Kanter and W. Kinzel, Phys. Rev. E **66**, 066102 (2002).

[3] L. Shacham, E. Klein, R. Mislovaty, I. Kanter and W. Kinzel, cond-mat/0312068 (2003).

[4] D. Saad and S. Solla, Phys. Rev. E. **52**, 4225 (1995).

[5] J. A. Hertz, A. Krogh and R. G. Palmer, *Introduction to the Theory of Neural Computation*, (Addison-Wesley, Redwood City, CA, 1991).

[6] M. Biehl, P. Reigler and C. Wöhler, J. Phys. A **29**, 4769 (1996).

[7] R. Metzler, W. Kinzel and I. Kanter, Phys. Rev. E **62**, 2555 (2000) and W. Kinzel, R. Metzler and I. Kanter, J. Phys. A. **33** L141 (2000).

[8] M. Rosen-Zvi, E. Klein, I. Kanter and W. Kinzel, Phys.Rev.E **66** 066135 (2002).

[9] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[10] M. Biehl and P. Reigler, Europhys. Lett. **28**, 525 (1994).

[11] A.Klimov, A. Mityagin, A. Shamir, ASIACRYPT 2002. Several attacks were suggested, and the Flipping Attack is the most successful of them.

[12] R. Mislovaty, E. Klein, I. Kanter and W. Kinzel, Phys. Rev. Lett. **91**, 118701(2003).

[13] M. Copelli, M. Boutin, C. Van Der Broeck and B. Van Rompaey, Europhys. Lett., **46**, 139 (1999).