

Practically Solving LPN in High Noise Regimes Faster Using Neural Networks

Haozhe Jiang*[†]

Kaiyue Wen*[‡]

Yilei Chen[§]

March 14, 2023

Abstract

We conduct a systematic study of solving the learning parity with noise problem (LPN) using neural networks. Our main contribution is designing families of two-layer neural networks that practically outperform classical algorithms in high-noise, low-dimension regimes. We consider three settings where the numbers of LPN samples are abundant, very limited, and in between. In each setting we provide neural network models that solve LPN as fast as possible. For some settings we are also able to provide theories that explain the rationale of the design of our models.

Comparing with the previous experiments of Esser, Kübler, and May (CRYPTO 2017), for dimension $n = 26$, noise rate $\tau = 0.498$, the “Guess-then-Gaussian-elimination” algorithm takes 3.12 days on 64 CPU cores, whereas our neural network algorithm takes 66 minutes on 8 GPUs. Our algorithm can also be plugged into the hybrid algorithms for solving middle or large dimension LPN instances.

1 Introduction

Neural networks are magical, capable of learning how to play various board games [SHM⁺16, CHhH02, Tes95] and video games [MKS⁺15, VBC⁺19, BBC⁺19], how to control fusion reactors [DFB⁺22], how to predict spatial structures of proteins [JEP⁺21], etc. In recent years, the rise of neural networks not only revolutionizes the field of artificial intelligence but also greatly impacts other fields in computer science. It is natural to ask whether neural networks can help us with cryptography.

One of the signature cryptographic hard problems is the Learning Parity with Noise problem (LPN), also known as decoding binary random linear codes, a canonical problem in coding theory. Let n be the dimension, $\tau \in (0, 0.5)$ be the error rate. Let \mathbf{s} be a secret vector in \mathbb{Z}_2^n . The LPN problem asks to find the secret vector \mathbf{s} given an oracle which, on its i^{th} query, outputs a random vector $\mathbf{a}_i \in \mathbb{Z}_2^n$, and a bit $y_i := \langle \mathbf{s}, \mathbf{a}_i \rangle + e_i \pmod{2}$, where e_i is drawn from the error distribution that outputs 1 with probability τ , 0 with probability $1 - \tau$. When τ is very small, say $\tau \in (0, 1/n)$,

*Equal Contribution.

[†]IIS, Tsinghua University. Email: jianghz20@mails.tsinghua.edu.cn.

[‡]IIS, Tsinghua University. Email: wenky20@mails.tsinghua.edu.cn.

[§]IIS, Tsinghua University, Shanghai Artificial Intelligence Laboratory, and Shanghai Qi Zhi Institute. Email: chenyilei@mail.tsinghua.edu.cn. Research supported by Tsinghua University startup funding.

then as long as we obtain $\Omega(n)$ LPN samples, we can efficiently find out which n of the samples are error-free and use Gaussian elimination to find the secret. However, for large $\tau \in (1/n^c, 0.5)$ where $0 < c < 1$, no classical or quantum algorithm is known for solving LPN in polynomial time in n . The LPN problem was proposed by machine learning experts as a conjectured hard problem for good cryptographic use [BFKL93]. Since its proposal, researchers have found numerous interesting cryptographic applications from LPN, including authentication protocols [HB01, KPC⁺11], public-key encryptions [Ale03, YZ16], identity-based encryptions [BLSV18, DGHM18], and efficient building blocks for secure multiparty computation [BCG⁺19]. The LPN problem also inspires the formulation of the learning with errors problem (LWE) [Reg09], which is more powerful in building cryptographic tools.

For LPN with constant noise rates, the asymptotically fastest algorithm, due to Blum, Kalai, and Wasserman [BKW03], takes $2^{O(\frac{n}{\log n})}$ time and requires $2^{O(\frac{n}{\log n})}$ samples. However, for cryptosystems based on LPN, the number of samples is typically a small polynomial in n . In this setting, Lyubashevsky gives a $2^{O(\frac{n}{\log \log n})}$ time algorithm [Lyu05] by first amplifying the number of samples and then running the BKW algorithm. To improve the concrete running time for solving LPN, researchers develop more sophisticated hybrid algorithms, see for example [LF06, GJL14, BTV16, ZJW16, BV16, EKM17, EHK⁺18]. The hybrid algorithms combine the BKW reduction with other tools like the “Guess-then-Gaussian-elimination” (henceforth Gauss) algorithm and decoding algorithms in coding theory [MMT11, BJMM12]. Using hybrid algorithms, Esser, Kübler, and May [EKM17] show that middle-size LPN instances are within the reach of the current computation power. For example, they show that LPN with dimension $n = 135$, noise rate $\tau = 0.25$ can be solved within 5.69 days using 64 CPU cores. The practical running time for solving a larger instance of $n = 150$, $\tau = 0.25$ is recently reported as less than 5 hours by Wiggers and Samardjiska [WS21] by using 80 CPU cores and more carefully designed reduction chains.

Machine learning and LPN. The most elementary setting of machine learning is the *supervised learning* paradigm. In this setting, we are given a set of labeled data that consists of some input data (e.g. pictures) and corresponding labels (e.g. whether the picture contains animals, not necessarily correct). The input data is presumed to be drawn from a fixed distribution and the given dataset often does not contain most of the possible inputs. The goal is to find a function that predicts the labels from inputs. The predictions are required to be accurate on both seen and unseen inputs. The size of the dataset required to find a good enough function is called the *sample complexity* and the function is often called *model*. The function is always chosen from the *function class*, a pre-determined set of functions. If the function class is simple, it would be easy to find the best function, but none of them would be sophisticated enough to perform well on complex tasks. In contrast, a complex function class holds greater expressive power but may be hard to optimize and has higher sample complexity. We often call the optimization procedure *learning*. Neural networks provide us with huge and adjustable expressiveness as well as an efficient learning algorithm. The recent success of neural networks largely hinges on their expressive power, as well as recent advances in big data and computing resources to make them useful.

The LPN problem perfectly fits into the supervised learning paradigm. Specifically, let the queries be inputs and the parity with noise be labels. If we can find a model that simulates the LPN oracle without noise, we can use this model to sample some data and use gaussian elimination to recover the secret. Thus it is tempting to try using the power of neural networks to break LPN. However, perhaps

surprisingly, there are very few instances where neural networks outperform the conventional algorithms in the cryptanalysis literature. In fact, most of the successful examples are in the attack of blockciphers, see e.g. [Ala12, Goh19, BGPT21]. The only documented attempt of using neural networks in solving problems related to LPN is actually aiming at solving LWE, conducted by Wenger et al. [WCCL22]. But their neural network algorithm is not yet competitive with the traditional algorithms. Besides, the limited theoretical understanding of neural networks we have is in stark contrast to its splendid empirical achievements. Hence this paper mainly focuses on empirically demonstrating neural networks’ usefulness in breaking the LPN problem with the most rudimentary networks. We hope this paper can serve as a starting point for breaking LPN and other applications in cryptography with neural networks. We also hope this paper can motivate further theoretical works on relevant problems.

1.1 Our contributions

Our main contribution is designing families of neural networks that practically outperform classical algorithms in high-noise, low-dimension regimes.¹ We consider three settings. In the *abundant sample setting*, the time complexity is considered the prime. In the *restricted sample setting*, efforts are made to reduce the sample complexity required by neural networks. In the *moderate sample setting*, we consider optimizing time complexity given sample complexity typically given by the reduction phase of the hybrid algorithm.

Model \mathcal{M}	Initialization	Loss	Optimizer
Base ₁₀₀₀ (Definition 2.14)	Kaiming [HZRS15]	Logistic	Adam

Table 1: Shared Features in All Three Settings

Our main experimental results suggest that two-layer neural networks with Adam optimizer and logistic loss function work the fastest in all three settings. Some shared features of our algorithms are presented in Table 1. Our neural network architectures are quite different from the ones used by Wenger et al. [WCCL22] where they use the transformer model to attack LWE. Let us also remark that previous attempts of using neural networks for solving decoding problems use more complicated network structures with more than five layers [NML⁺18, BCK18], in contrast to our design that only uses two layers.

Although the power of neural networks is usually hard to explain, for some settings we are able to explain the rationale of the design of our machine learning models in theory, in terms of their representation capability, optimization power, and the generalization effect. We present some first-step analysis in Section 5.

Comparing with the previous experiments of Esser, Kübler, and May [EKM17], for dimension $n = 26$, noise rate $\tau = 0.498$, the “Guess-then-Gaussian-elimination” algorithm takes 3.12 days on 64 CPU cores, whereas our algorithm takes 66 minutes on 8 GPUs. For large instances like $n = 125$, $\tau = 0.25$, our algorithm can also be plugged into hybrid algorithms to reduce the running time from 4.22 days²

¹We release our code in <https://github.com/WhenWen/Solving-LPN-using-Neural-Networks.git>.

²The experiment reported in [EKM17, Page 28] solved an LPN instance of $n = 135$, $\tau = 0.25$ in 5.69 days. They first spent 1.47 days to enumerate 10 bits of secrets. We didn’t repeat the enumeration step and directly start from $n = 125$, $\tau = 0.25$.

to 3.5 hours. To reach this performance, we first use BKW to reduce $1.2e10$ LPN samples of dimension 125, noise rate 0.25 to $1.1e8$ LPN samples of dimension 26, noise rate 0.498 in 40 minutes with 128 CPUs. Then we enumerate the last 6 bits and apply our neural network algorithm (Algorithm 11) with 8 GPUs on LPN with $n = 20, \tau = 0.498$ to find out 26 bits of the secret in 66 minutes. The whole process (BKW+neural network) is then repeated to solve another 26 bits in another 106 minutes. The final reduced problem has dimension 73 and noise rate 0.25 and can be solved by the MMT algorithm in 1 minute. In total, the running time is 3.55 hours.

Let us remark that we haven't compared our neural network algorithm with the more sophisticated decoding algorithms such as Walsh-Hadamard or MMT. But our experiments have already shown that neural networks can achieve competitive performance for solving LPN in the high-noise, low-dimension regime compared to the Gauss decoding, when running on hardware with similar costs, and there is a huge potential for improvement for our neural network algorithm. There is also clear room for improvement in our reduction algorithm. For example, the time complexity of the second round can be significantly reduced given that we already know 26 bits of the secret. However, our algorithm mainly focuses on accelerating the decoding phase of the hybrid algorithm. In this sense, our contributions are orthogonal to [BV16, WS21], where the improvement is made possible by constructing a better reduction chain.

More details in the three settings. As mentioned, we consider three settings, named *abundant*, *restricted*, and *moderate* sample setting. Let us now explain those three settings.

Setting 1.1 (Abundant Sample). *In the “Abundant Sample” setting, we assume an unlimited amount of fresh LPN samples are given, and we look for the algorithm that solves LPN as fast as possible. This setting has a clear definition and will serve as the starting point for our algorithms.*

Under Setting 1.1, we first present the naive algorithm that directly considers the LPN problem as a supervised learning algorithm. This algorithm ignores the special structures of the LPN problem. However, our experiments in Section 4.1 show that given enough samples, neural networks can learn the LPN encoding perfectly. One example run on an LPN problem with dimension 20 and noise rate 0.498 is shown in Figure 1. We also compare the time complexity of our algorithm with the Gauss algorithm in Table 2, showing the supremacy of our algorithm in the heavy noise regime.

Setting 1.2 (Restricted Sample). *In the “Restricted Sample” setting, sample complexity is considered the prime. For the same LPN task, an algorithm that solves the task with fewer LPN oracle queries is more desirable.*

In Setting 1.2, as the sample complexity is bounded, we would apply the *search-decision* reduction first to simplify our goal. However, even distinguishing LPN data from random data is hard in this case due to the phenomenon of *overfitting*. This means when training on a small amount of data, the dataset might render the neural network to memorize all the data, resulting in poor performance on unseen inputs. We show that the commonly used regularization method named L2 regularization, or weight decay, is significantly helpful under this setting in Section 4.2. Empirical evaluations show that our algorithms achieve comparable sample complexity with state-of-the-art algorithms, the results are shown in Table 8. Together with the positive results in Setting 1.1, we show that neural-network based algorithms have the potential to improve the breaking algorithm of cryptography primitives.

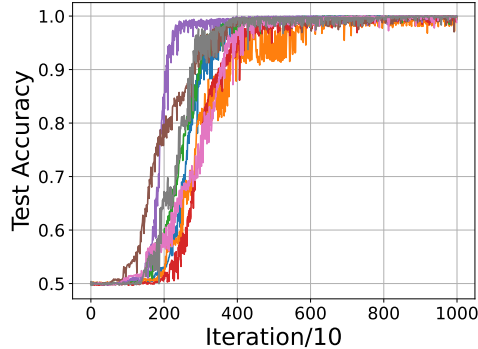


Figure 1: **Experiments on LPN_{20,∞,0.498}**. The horizontal axis represents the training iteration. One unit on this axis represents 10 iterations. The vertical axis represents the accuracy of the model on the clean test set without noise. There are 8 curves in this graph, each corresponding to a random initialization.

$n \backslash \tau$	0.4	0.45	0.49	0.495	0.498
20	39	70	197	323	730
30	139	374	1576		

(a) **Neural Network**

$n \backslash \tau$	0.4	0.45	0.49	0.495	0.498
20	0.40	5.76	22.0	312	6407
30	26.4	682			

(b) **Gaussian Elimination**

Table 2: **Time Complexity w.r.t Dimension and Noise Rate**. Each entry represents the running time (in seconds). The experiments for neural networks are performed on a single GPU. The experiments for Gaussian elimination are performed on a single 64 cores processor. For a neural network, the criterion for solving the LPN is that the accuracy of the network reaches 80% on clean data. For Gaussian Elimination, the criterion for success is to get at least 7 correct secrets out of 10 attempts. The running time here is averaged over all runs that recover the correct secret in the time limit (3 hours). For the empty cell, no runs are successful within 3 hours.

Setting 1.3 (Moderate Sample). *In the “Moderate Sample” setting, we constrain the sample complexity and seek the smallest running time. This setting is typically used as part of hybrid algorithms, where our algorithm is used to solve LPN instances reduced from BKW or other algorithms.*

To further validate our points, we consider the more refined setting 1.3, where we consider minimizing time complexity given the sample complexity. We show with experiments that with the number of samples provided by reduction algorithms like BKW, neural networks can already learn a model with moderate accuracy, even with noise as high as 0.498. This resolves the impracticability of the algorithm we design for Setting 1.1, which requires too many samples. With a combination with the traditional algorithm Gauss, we can leverage neural networks to solve low dimension, high noise LPN instances faster than previously reported [EKM17].

Concluding the three settings, we show that neural networks have huge potential in solving LPN in a practical sense, for metrics spanning from pure time complexity to pure sample complexity. We also show that we can already include neural networks as a building block of the *reduction-decoding* scheme

to accelerate the breaking of large instances of LPN problem, which to our knowledge is the first time for neural-network based algorithm to achieve comparable, or even better, performance for LPN problem.

Future directions. At the end of the introduction we would like to mention a few interesting open problems:

1. The neural network structure used by our solver is quite simple – we only use two-layer, fully-connected networks. Are there any neural networks with more dedicated structures that help to solve LPN?
2. Is it possible to use our technique to solve the LWE problem? Compared to LPN, LWE uses a large modulus and uses ℓ_2 norm to measure the length of the noise. Are those differences crucial for the competitiveness of neural networks?
3. In addition to the previous works that use neural networks in designing decoding algorithms in coding theory [NML⁺18, BCK18], our work shows the neural network has a plausible advantage in decoding binary random linear codes in the high noise regimes. Can we develop practically fast neural network decoding algorithms for other codes?

Organization. The paper is organized as follows. In Section 2 we fix some notations throughout the paper. This section also contains a brief tutorial on machine learning and neural networks. In Section 3 we explain our techniques and algorithms used for breaking LPN. In Section 4 we conduct experiments to demonstrate the usefulness of techniques used and compare the performance of our algorithms to SOTA prior algorithms. We also provide a guideline for tuning hyperparameters for our algorithm. In Section 5 we provide theories that explain the rationales of our network architectures.

2 Preliminary

Notations and terminology. In cryptography, the security parameter is a variable that is used to parameterize the computational complexity of the cryptographic algorithm or protocol, and the adversary’s probability of breaking security.

Let $\mathbb{R}, \mathbb{Z}, \mathbb{N}$ be the set of real numbers, integers, and positive integers. For $q \in \mathbb{N}_{\geq 2}$, denote $\mathbb{Z}/q\mathbb{Z}$ by \mathbb{Z}_q . For $n \in \mathbb{N}$, $[n] := \{1, \dots, n\}$. A vector in \mathbb{R}^n (represented in column form by default) is written as a bold lower-case letter, e.g. \mathbf{v} . For a vector \mathbf{v} , the i^{th} component of \mathbf{v} will be denoted by v_i . A matrix is written as a bold capital letter, e.g. \mathbf{A} . The i^{th} column vector of \mathbf{A} is denoted \mathbf{a}_i . The length of a vector is the ℓ_p -norm $\|\mathbf{v}\|_p := (\sum v_i^p)^{1/p}$, or the infinity norm given by its largest entry $\|\mathbf{v}\|_\infty := \max_i \{|v_i|\}$. The length of a matrix is the norm of its longest column: $\|\mathbf{A}\|_p := \max_i \|\mathbf{a}_i\|_p$. By default, we use ℓ_2 -norm unless explicitly mentioned. For a binary vector \mathbf{v} , let $\text{HW}(\mathbf{v})$ denote the Hamming weight of \mathbf{v} . Let B_p^n denote the open unit ball in \mathbb{R}^n in the ℓ_p norm. We will write $x \text{ e } y$ as short hands for $x \times 10^y$.

When a variable v is drawn uniformly random from the set S we denote it as $v \leftarrow U(S)$. When a function f is applied on a set S , it means $f(S) := \sum_{x \in S} f(x)$.

2.1 Learning Parity with Noise

The learning parity with noise problem (LPN) is defined as follows

Definition 2.1 (LPN [BFKL93, BKW03]). *Let $n \in \mathbb{N}$ be the dimension, $m \in \mathbb{N}$ be the number of samples, $\tau \in (0, 1/2)$ be the error rate. Let η_τ be the error distribution that output 1 with probability τ , 0 with probability $1 - \tau$. A set of m LPN samples is obtained from sampling $\mathbf{s} \leftarrow U(\mathbb{Z}_2^n)$, $\mathbf{A} \leftarrow U(\mathbb{Z}_2^{n \times m})$, $\mathbf{e} \leftarrow \eta_\tau^m$, and outputting $(\mathbf{A}, \mathbf{y}^t := \mathbf{s}^t \mathbf{A} + \mathbf{e}^t \pmod{2})$.*

We say that an algorithm solves $\text{LPN}_{n,m,\tau}$ if it outputs \mathbf{s} given \mathbf{A} and \mathbf{y} with non-negligible probability.

An algorithm solves the decisional version of LPN if it distinguishes the LPN sample $\text{LPN}_{n,m,\tau}$ from random samples over $\mathbb{Z}_2^{n \times m} \times \mathbb{Z}_2^m$ with probability greater than $1/2 + 1/\text{poly}(n)$. The decisional LPN problem is as hard as the search version of LPN [BFKL93].

The LPN problem reduces to a variant of LPN where the secret is sampled from the error distribution [ACPS09]. The reduction is simple and important for our application so we sketch the theorem statement and the proof here.

Lemma 2.2. *If $\text{LPN}_{n,m,\tau}$ is hard, then so is the following variant of LPN: we sample each coordinate of the secret $\mathbf{s} \in \mathbb{Z}_q^n$ from the same distribution as the error distribution, i.e., η_τ , and then output $m - n$ LPN samples.*

Proof. Given m standard LPN samples, denoted as $(\mathbf{A}, \mathbf{y}^t := \mathbf{s}^t \mathbf{A} + \mathbf{e}^t \pmod{2})$. Write $\mathbf{A} = [\mathbf{A}_1 \mid \mathbf{A}_2]$ where $\mathbf{A}_1 \in \mathbb{Z}_2^{n \times n}$. Without a loss of generality, assume \mathbf{A}_1 is invertible (if not, pick another block of n full-rank columns from \mathbf{A} as \mathbf{A}_1). Write $\mathbf{y}^t = [\mathbf{y}_1^t \mid \mathbf{y}_2^t]$ where $\mathbf{y}_1 \in \mathbb{Z}_2^n$. Let $\bar{\mathbf{A}} := -\mathbf{A}_1^{-1} \cdot \mathbf{A}_2$. Let $\bar{\mathbf{y}}^t := \mathbf{y}_1^t \cdot \bar{\mathbf{A}} + \mathbf{y}_2^t$. Then $\bar{\mathbf{y}}^t = (\mathbf{s}^t \mathbf{A}_1 + \mathbf{e}_1^t) \cdot (-\mathbf{A}_1^{-1} \cdot \mathbf{A}_2) + (\mathbf{s}^t \mathbf{A}_2 + \mathbf{e}_2^t) = \mathbf{e}_1^t \cdot \bar{\mathbf{A}} + \mathbf{e}_2^t$, meaning that $\bar{\mathbf{A}}, \bar{\mathbf{y}}^t$ is composed of $m - n$ LPN samples where the secret is sampled from the error distribution. \square

2.2 Machine Learning

Supervised Learning The goal of supervised learning is to learn a function that maps inputs to labels. The input $x \in \mathcal{X}$ and the label $y \in \mathcal{Y}$ are usually assumed to obey a fixed distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$. Usually, \mathcal{D} is not directly accessible to the learner, instead, another distribution $\mathcal{D}_{\text{data}}$, known as empirical distribution, is provided to the learner. This distribution is usually a uniform distribution over a finite set of inputs and labels $\mathcal{S}_{\text{Train}} \triangleq \{(x_i, y_i)\}_{i \in [1:N]}$. This set $\mathcal{S}_{\text{Train}}$ is usually named *training set* and (x_i, y_i) is assumed to obey \mathcal{D} independently.

The goal of *learning* is to choose from a function class \mathcal{H} a function $f : X \rightarrow Y$ given $\mathcal{D}_{\text{data}}$. To measure the quality of f , *loss function* $\ell : Y \times Y \rightarrow \mathbb{R}^{\geq 0}$ is often considered. We now provide some examples of loss functions that will be used in our paper.

Definition 2.3 (Zero-one Loss). $\ell_{0-1}(y_1, y_2) = 1[y_1 \neq y_2]$.

Definition 2.4 (Logistic Loss). $\ell_{\log}(y_1, y_2) = -y_2 \log(1 - y_1) - (1 - y_2) \log y_1, y_2 \in \{0, 1\}, y_1 \in [0, 1]$.

Definition 2.5 (Mean Absolute Error Loss). $\ell_{\text{mae}}(y_1, y_2) = |y_1 - y_2|$.

Definition 2.6 (Mean Square Error Loss). $\ell_{\text{mse}}(y_1, y_2) = |y_1 - y_2|^2$.

Definition 2.7. Given a loss function $\ell : Y \times Y \rightarrow \mathbb{R}^{\geq 0}$, the population loss $\mathcal{L}_{\mathcal{D}} : \mathcal{H} \rightarrow \mathbb{R}^{\geq 0}$ is defined as

$$\mathcal{L}_{\mathcal{D}}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(f(x), y)].$$

For zero-one loss, 1 minus the expected loss is also called *accuracy*. We usually abuse this notation when f 's co-domain is $[0, 1]$ by calling the accuracy of the rounding of f as the accuracy of f . The *training accuracy* is defined as accuracy with the underlying population as the uniform distribution over the training set. The goal of learning can then be rephrased to find $f \in \mathcal{H}$ with low population loss. Two questions then naturally arise, (1) How to evaluate population loss? and (2) How to effectively minimize population loss?

To evaluate the loss, the *test set* $\mathcal{S}_{\text{Test}} \triangleq \{(x_i, y_i)\}_{i \in [N+1, N+M]}$ is usually considered. The element in $\mathcal{S}_{\text{Test}}$ is also assumed to obey \mathcal{D} independently and is also independent of the elements in $\mathcal{S}_{\text{Train}}$.

Definition 2.8. Given a loss function $\ell : Y \times Y \rightarrow \mathbb{R}^{\geq 0}$, the test loss $\mathcal{L}_{\text{Test}} : \mathcal{H} \rightarrow \mathbb{R}^{\geq 0}$ is defined as

$$\mathcal{L}_{\text{Test}}(f) = \frac{1}{M} \sum_{i=N+1}^{N+M} \ell(f(x_i), y_i).$$

The test accuracy is defined as accuracy with the underlying population as the uniform distribution over the test set.

When f is chosen by the learning algorithm given the training data, $\mathcal{L}_{\text{Test}}(f)$ can then serve as an unbiased estimator of $\mathcal{L}_{\mathcal{D}}(f)$. In the traditional machine learning community, $\mathcal{L}_{\text{Test}}$ is usually only measured once after training, and another set called *validation set* is used to track the performance of the algorithm through the course of the training. However, this boundary is blurred in modern literature and we ignore this subtlety here because our final objective is to utilize machine learning to solve LPN secrets instead of fitting the data.

To effectively minimize the loss, the learner would use a learning algorithm \mathcal{A} that maps training distribution to a function $f \in \mathcal{H}$ (usually with randomness). As the learner only has access to the data distribution, \mathcal{A} is usually designed to minimize *training loss*.

Definition 2.9. Given a loss function $\ell : Y \times Y \rightarrow \mathbb{R}^{\geq 0}$, the training loss $\mathcal{L}_{\text{Train}} : \mathcal{H} \rightarrow \mathbb{R}^{\geq 0}$ is defined as

$$\mathcal{L}_{\text{Train}}(f) = \frac{1}{M} \sum_{i=1}^M \ell(f(x_i), y_i).$$

When trying to characterize the gap between the learned function and the best available function in the function class, the following decomposition is common in machine learning literature.

$$\begin{aligned} & \mathcal{L}_{\mathcal{D}}(\mathcal{A}(\mathcal{D}_{\text{data}})) - \underbrace{\min_{f \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(f)}_{\text{Representation Gap}} \\ &= \underbrace{\mathcal{L}_{\mathcal{D}}(\mathcal{A}(\mathcal{D}_{\text{data}})) - \mathcal{L}_{\text{Train}}(\mathcal{A}(\mathcal{D}_{\text{data}}))}_{\text{Generalization Gap}} + \underbrace{\mathcal{L}_{\text{Train}}(\mathcal{A}(\mathcal{D}_{\text{data}})) - \mathcal{L}_{\text{Train}}(f^*)}_{\text{Optimization Gap}} \\ &+ \underbrace{\mathcal{L}_{\text{Train}}(f^*) - \mathcal{L}_{\mathcal{D}}(f^*)}_{\text{Stochastic Error}}, \quad f^* = \arg \min_{f \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(f). \end{aligned} \tag{1}$$

The three gaps in the above equation characterize different aspects of machine learning.

1. The function class needs to be chosen to be general enough to minimize the representation gap.
2. The learning algorithm needs to be chosen to find the best trade-off between the generalization gap and the optimization gap given the function class.

In the recent revolution brought by neural networks, it is shown that choosing the function class as $\{f \mid f \text{ can be represented by a fixed neural architecture}\}$ and learning algorithm as the gradient-based optimization method can have surprising effects over various domains. We will now briefly introduce neural networks and gradient-based optimization algorithms.

Neural Networks Neural networks are defined by *architecture*, which maps differentiable weights to a function from \mathcal{X} to \mathcal{Y} . This function is called the *neural network* and the weights are called the *parameterization* of the network. The most simple architecture is *Multi-Layer Perceptron (MLP)*.

Definition 2.10 (MLP). *Multi-Layer Perceptron is defined as a mapping \mathcal{M} from \mathbb{R}^{d_1} to $\mathbb{R}^{d_{L+1}}$, with*

$$\mathcal{M}[\theta_1, \dots, \theta_L](x) = (\sigma_L \circ T[\theta_L] \circ \dots \circ \sigma_1 \circ T[\theta_1])(x),$$

where $\theta_i = (W_i, b_i)$, $W_i \in \mathbb{R}^{d_{i+1} \times d_i}$, $b_i \in \mathbb{R}^{d_{i+1}}$ and $T[\theta_i]$ is an affine function with $T[\theta_i](x) = W_i x + b_i$. $\sigma_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a function that is applied coordinate-wise and is called activation function. L and $L-1$ are called the number of layers and depth of the MLP, and $\{d_i\}_{i=2, \dots, L}$ is called the widths of the MLP.

We now provide some examples of activation functions that will be used in our work.

Definition 2.11 (ReLU). $\text{ReLU} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as $(\text{ReLU}(x))_i = x_i 1[x_i \geq 0]$.

Definition 2.12 (Sigmoid). $\text{Sigmoid} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as $(\text{Sigmoid}(x))_i = \frac{1}{1+e^{-x_i}}$.

Definition 2.13 (Cosine). $\text{Cos} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as $(\text{Cos}(x))_i = \cos(x_i)$.

The base model we used in this work is defined as followed.

Definition 2.14 (Base Model). *Our base model is defined as MLP with depth 1 with activation $\sigma_1 = \text{ReLU}$ and $\sigma_2 = \text{Sigmoid}$. We will denote this model as Base_d with d specifying d_2 . We will write Base with activation σ to indicate replacing σ_1 by σ .*

Gradient-based Optimization It is common to use gradient-based optimization methods to optimize the neural network. A standard template is shown in Algorithm 1. The unspecified parameters such as \mathcal{M} , W , and ℓ in the algorithm are often called *hyperparameters*.

Definition 2.15 (Sampler). *A sampler is a finite-state machine, on each call of method GetData , it will return a set of B samples $\{(x_i, y_i)\}$ satisfying $x_i, y_i \sim \mathcal{D}_{\text{data}}$. The number B is called batch size and the B samples are called a batch.*

We hereby provide two examples of samplers that will be used in our papers. Algorithm 2 is the sampler used in Settings 1.2 and 1.3 and Algorithm 3 is the sampler used in Setting 1.1.

Algorithm 1 Gradient-based Optimization

Require: A neural network architecture \mathcal{M}

Require: An initialization parameter for the model W_0

Require: A differentiable loss function ℓ

Require: A Stop Criterion \mathcal{SC}

Require: A Data Sampler \mathcal{DS}

▷ See Definition 2.15

Require: An Optimizer \mathcal{O}

▷ See Definition 2.16

Require: A Regularization Function \mathcal{R}

▷ See Definition 2.17

$step \leftarrow 0$

while \mathcal{SC} is not reached **do**

▷ See Definition 2.20

$f \leftarrow \mathcal{M}[W_{step}]$

$\mathcal{S}_{Batch} = \{(x_i, y_i)\}_{i=1, \dots, B} \leftarrow \mathcal{DS}.GetData()$.

$\mathcal{L}_{Batch} \leftarrow \frac{1}{B} \sum_{i=1}^B \ell(f(x_i), y_i) + \mathcal{R}(W_{step})$.

▷ Calculate regularized loss

$g_W \leftarrow -\frac{\partial \mathcal{L}_{Batch}}{\partial W} |_{W=W_{step}}$.

▷ Calculate gradient w.r.t the model parameter

$W_{step+1} \leftarrow \mathcal{O}.GetUpdate(W_{step}, g_W)$.

$step \leftarrow step + 1$.

end while

return $\mathcal{M}[W_{step}]$

Algorithm 2 Batch Sampler

parameters

$\mathcal{S}_{Train} = \{(x_i, y_i)\}_{i=1, \dots, N}$, training set

$B \leq N$, batch size

end parameters

procedure GETDATA()

Sample i.i.d from $\{1, 2, \dots, N\}$ B index to form index set I_{Batch}

$\mathcal{S}_{Batch} = \{(x_i, y_i) \mid i \in I_{Batch}\}$.

return \mathcal{S}_{Batch} .

end procedure

Algorithm 3 Oracle Sampler

parameters

\mathcal{D} , the underlying distribution

$B \leq N$, batch size

end parameters

procedure GETDATA()

$\mathcal{S}_{Batch} = \{(x_k, y_k)\}_{k=1 \dots B}$ with (x_k, y_k) i.i.d sampled from \mathcal{D} .

return \mathcal{S}_{Batch} .

end procedure

Definition 2.16 (Optimizer). *An optimizer is an automaton, on each call of method `GetUpdate`, it will update states given the current parameter and gradient, and return an updated parameter.*

We now provide some examples of optimizers. The *stochastic gradient descent (SGD) Optimizer* is shown in Algorithm 4. When the batch size equals the size of the training set, this algorithm is often called *gradient descent* directly. In our paper, we use a more complicated optimizer named *Adam*, as shown in Algorithm 5. This optimizer, although poorly understood theoretically, has been widely applied across domains by the current machine-learning community.

Algorithm 4 SGD Optimizer

parameters

η , learning rate

λ , weight decay

end parameters

procedure GETUPDATE(W, g_W)

return $W - \eta(\lambda W + g_W)$.

end procedure

Definition 2.17 (Regularization). *A regularization function is defined as a mapping from the parameter space to \mathbb{R} .*

Theoretically and empirically, a proper choice of a regularization function can improve the generalization of the learned model in previous literature. We hereby provide two examples of regularization functions that will be used in our paper. One can easily notice that L2 regularization (Definition 2.18) applied in the gradient-based optimization method is simply another form of weight decay. L1 regularization (Definition 2.19) applied with the linear model is known as *LASSO* and can induce sparsity in the model parameters (meaning the model parameters contain more zeroes).

Definition 2.18 (L2 Regularization). *L2 Regularization $R_2 : (\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_k}) \rightarrow \mathbb{R}$ with penalty factor λ is defined as $R_2(w_1, \dots, w_d) = \frac{\lambda}{2} \sum_i \|w_i\|_2^2$.*

Definition 2.19 (L1 Regularization). *L1 Regularization $R_1 : (\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_k}) \rightarrow \mathbb{R}$ with penalty factor λ is defined as $R_1(w_1, \dots, w_d) = \lambda \sum_i \|w_i\|_1$.*

Definition 2.20 (Stop Criterion). *A stop criterion is defined as a function that returns True or False determining whether the procedure should terminate.*

There are typically three kinds of stop criteria, which we list as below.

Definition 2.21 (Stop-by-time). *A stop-by-time criterion $\mathcal{SC}_{\text{time}}(t_{\text{time}})$ returns true if physical running time exceeds threshold t .*

Definition 2.22 (Stop-by-step). *A stop-by-step criterion $\mathcal{SC}_{\text{step}}(t_{\text{step}})$ returns true if the weight update step exceeds threshold T .*

Definition 2.23 (Stop-by-accuracy). *A stop-by-accuracy criterion $\mathcal{SC}_{\text{acc}}(\mathcal{S}, \gamma)$ returns true if the accuracy of the learned function on \mathcal{S} exceeds threshold γ .*

Algorithm 5 Adam Optimizer

state variables

m , first moment, initialized to be 0.
 v , second moment, initialized to be 0.
 ϵ , a small positive constant, by default $1e - 8$

end state variables**parameters**

η , learning rate
 λ , weight decay
 β_1, β_2 , moving average factor for moments, by default 0.9, 0.999.

end parameters**procedure** GETUPDATE(W, g_W)

$dW \leftarrow \lambda W + g_W$.
 $m \leftarrow \beta_1 m + (1 - \beta_1) dW$.
 $v \leftarrow \beta_2 m + (1 - \beta_2) dW^2$.
 $\hat{m} \leftarrow m / (1 - \beta_1)$.
 $\hat{v} \leftarrow v / (1 - \beta_2)$.
return $W - \eta \hat{m} / (\sqrt{\hat{v}} + \epsilon)$

end procedure

3 Methods

In this section, we will introduce our methods under the three different settings introduced in Section 1. This section is organized as follows. Section 3.1 introduces our methods when the data distribution coincides with the population distribution and the time complexity is a major concern. In other words, we assume an unrestricted number of fresh LPN samples are given and our goal is to learn the secret as fast as possible. We will show that under this setting, direct application of the gradient-based optimization (cf. Algorithm 1) works well both empirically and theoretically. Section 3.2 introduces our methods when the sample complexity is greatly limited. Under this setting, reduction to decision version LPN and proper regularization is essential. Section 3.3 introduces our methods when the sample complexity and the time complexity need a subtle balance such that our method can be utilized as a building block in the classical *reduction-decoding* scheme of the LPN problem. Under this setting, we will use the samples to approximate the population distribution with a bootstrapping technique.

Due to Lemma 2.2, we will always assume secret \mathbf{s} has Hamming weight $\lfloor n\tau \rfloor$.³ We will first highlight the common hyperparameters for all three settings in Table 3. The first row is applied in all our implemented algorithms and the second row is used in our theoretical analysis in Section 5.

We further outline other hyperparameters specified for each algorithm in Table 4. Readers should note that the hyperparameters are tuned specifically towards the typical problem specification we list out in Table 5 on which our hyperparameters selection is conducted on. The power of our method is not limited to the typical specification listed and we provide generic meta algorithms 8 and 10 for hyperparameter selection in other cases.

³This is the expected hamming weight of a secret. We fix the sparsity here mainly for the convenience of comparing and adjusting other parameters.

	Model \mathcal{M}	Initialization	Regularization	Loss	Optimizer
Practical	Base ₁₀₀₀ (Definition 2.14)	Kaiming	None	Logistic	Adam
Theory	Base _{d} with smooth activation σ	Any	None	MAE	SGD

Table 3: Shared Hyperparameters

	Sampler	Learning Rate	Weight Decay	Batch Size	Stop Criterion
Algorithm 6	Oracle	$2e - 4$ to $6e - 3$	0	131072 to 1048576	$\mathcal{SC}_{\text{time}}(t_{\text{time}})$
Algorithm 7	Oracle	Any	Any	Any	$\mathcal{SC}_{\text{step}}(t_{\text{step}})$
Algorithm 9	Fix Batch	$2e - 5$ to $1e - 4$	$2e - 3$	Training Set Size	$\mathcal{SC}_{\text{step}}(t_{\text{step}})$
Algorithm 11	Fix Batch	$2e - 3$	0	1048576	$\mathcal{SC}_{\text{time}}(t_{\text{time}})$

Table 4: Hyperparameters for Different Algorithms

3.1 Abundant Sample

In Setting 1.1, we assume the learner has access to the Oracle Sampler (Algorithm 3) with underlying distribution following LPN instances with a fixed secret, a fixed noise rate τ , and any batch size B . Under this setting, the memory of the devices is the key constraint and the goal of this algorithm is to reduce the time complexity. We propose to use a direct variant of the gradient-based optimization algorithm Algorithm 1, which is shown in Algorithm 6.

One may notice that there are some unspecified hyperparameters, including the time threshold t_{time} in Algorithm 6. Also for settings outside the scope of Table 5, some currently fixed parameters such as the learning rate η and the batch size B may also require tuning based on the problem setting, i.e, n and τ . Under such case, we propose to use Algorithm 8, a meta algorithm for hyperparameter selection. In Section 4.1.2, we show that Algorithm 8 effectively finds hyperparameters that leads to low time complexity of Algorithm 6, for example, Algorithm 6 with $\eta = 6e - 3$ and $B = 1048576$, can solve $n = 20, \tau = 0.495$ in 6 minutes with a single GPU. An example running results of Algorithm 8 is shown in Table 7.

We also perform some theoretical analysis on the time complexity of our method, showing that the dependency on τ is merely $(\frac{1}{2} - \tau)^{-2}$. However, due to the complexity of our choice of the optimizer and the loss function in Algorithm 6, our analysis is conducted on a simpler version (Algorithm 7) with more clear theoretical structure. The analysis is deferred to Section 5.2.

	Dimension n	Noise Rate τ	Sample Size
Algorithm 6	20 to 40	0.4 to 0.498	Abundant
Algorithm 9	20 to 50	0.2 to 0.3	Less than 1e6
Algorithm 11	20	0.498	8e7

Table 5: Typical Problem Specification for Different Algorithms

Algorithm 6 Abundant Sample Algorithm: Practical Version

Require: n , the dimension, τ the error rate.

Run Algorithm 1 with hyperparameters specified in Tables 3 and 4 to get a learned model $\mathcal{M}[W_T]$.

Set \hat{s} as all zero vector in \mathbb{R}^n .

for $i \in \{1, 2, \dots, n\}$ **do**

Set $\hat{s}[i] = 1[\mathcal{M}[W_T][e_i] > 0.5]$ with $e_i \in \mathbb{R}^n$ as the unit vector with the i -th coordinate being 1.

end for

return \hat{s} .

Algorithm 7 Abundant Sample Algorithm: Theoretical Version

Require: n , the dimension, τ the error rate.

Run Algorithm 1 with hyperparameters specified in Tables 3 and 4 to get a learned model $\mathcal{M}[W_T]$ and return the model.

Algorithm 8 Abundant Sample Meta Algorithm For Hyperparameters Selection

Require: n , the dimension, τ the error rate.

Require: γ , the accuracy threshold, fixed as 80% in our experiments.

Require: *Repeat*, the number of different datasets to estimate t_{time} , fixed as 3 in our experiments

Require: A set of hyperparameters profiles \mathcal{P} .

Initialize an empty hashmap Map_t to record running time.

for Profile $P \in \mathcal{P}$ **do**

Set $\text{Map}_{t_{\text{time}}}[P] = \infty$.

for $r \leq \text{Repeat}$ **do**

Randomly generate secret s with Hamming weight $\lfloor n\tau \rfloor$.

Create an oracle sampler \mathcal{DS} with batch size B and error rate τ corresponding to s .

Sample a test dataset $\mathcal{S}_{\text{Test}}$ using \mathcal{DS} , with size $O(n)$.

▷ Fixed as 131072 in our experiments.

Set stop criterion \mathcal{SC}' as $\mathcal{SC}_{\text{acc}}(\mathcal{S}_{\text{Test}}, \gamma)$.

Run a variant of Algorithm 6 with stop criterion \mathcal{SC}' and profile P and record running time

t_{time} .

Set $\text{Map}_{t_{\text{time}}}[(B, \eta)] = \min\{t_{\text{time}}, \text{Map}_{t_{\text{time}}}[(B, \eta)]\}$.

▷ Only require solving by constant probability.

end for

end for

return $(P, \text{Map}_{t_{\text{time}}}[P])$ with the smallest $\text{Map}_{t_{\text{time}}}[P]$ for $P \in \mathcal{P}$.

3.2 Restricted Sample

In Setting 1.2, the sample complexity is assumed to be highly limited. Due to the restriction on sample complexity, we could no longer expect the learned model to fully recover the secret directly as in Section 3.1. Hence we design our algorithm to solve the decisional version of LPN and use neural network as a reduction method instead. We present our algorithm in Algorithm 9. The returning value of Algorithm 9 is with high probability the value of secret s at index $n - 1$. In practice, the inner for loop of Algorithm 9 is parallelized across multiple GPUs.

Algorithm 9 Restricted Sample Algorithm

Require: \mathcal{S} with m samples drawn from LPN with dimension n and error rate τ .

Require: *repeat* the number of random initializations to try, typically set as 8.

Require: γ , the accuracy threshold.

Split \mathcal{S} into $\mathcal{S}_{\text{Train}}$ and $\mathcal{S}_{\text{Test}}$, each with $\lfloor m/2 \rfloor$ sample.

for Guess $g \in \{0, 1\}$ **do**

 Generate $\mathcal{S}_{\text{Train}}^g = \{(x[1 : d - 1], y + x[d] \times g \bmod 2) \mid (x, y) \in \mathcal{S}_{\text{Train}}\}$.

 Generate $\mathcal{S}_{\text{Test}}^g = \{(x[1 : d - 1], y + x[d] \times g \bmod 2) \mid (x, y) \in \mathcal{S}_{\text{Test}}\}$.

 Set sampler \mathcal{DS} as batch sampler with batch size $\lfloor m/2 \rfloor$ on $\mathcal{S}_{\text{Train}}^g$.

for $r \leq \textit{repeat}$ **do**

 Run Algorithm 1 with the hyperparameters specified in Tables 3 and 4 to get a learned model $\mathcal{M}[W_T]$.

if the accuracy of $\mathcal{M}[W_T]$ on $\mathcal{S}_{\text{Test}}^g$ exceeds γ **then return** g .

end if

end for

end for

Again, there are some unspecified hyperparameters in Algorithm 9 that need to be pivoted down by experiments. In our experiments, we mainly consider $n \leq 50$, $\tau \leq 0.3$ and $m \in [2^{10}, 2^{20}]$. In this regime, we propose to set $t_{\text{step}} = 300k$ and γ as $1/2 + \sqrt{\log(20)/m}$. The dependency on τ is dropped here because the noise we considered here is relatively small. We observe in practice that the accuracy of test set on successful runs almost always exceeds this threshold by a large margin. For other hyperparameters including the exact value of sample size m , we apply meta algorithm Algorithm 10.

We would like to stress a few major differences in hyperparameters selection between Setting 1.1 and Setting 1.2.

1. L2 regularization, or weight decay, is not helpful in Setting 1.1, but can reduce the sample complexity significantly in Setting 1.2.
2. Under Setting 1.2, it is generally better to use the whole dataset as a batch instead of using a smaller batch size.
3. The learning rate required by Algorithm 9 is typically smaller than the learning rate required by Algorithm 6 by a factor of 10 to 100.

Algorithm 10 Restricted Sample Meta Algorithm For Hyperparameters Selection

Require: n , the dimension, τ the error rate.

Require: A set of sample numbers $m_0 < m_1 < m_2 < \dots < m_L$.

Require: $Repeat$, the number of different datasets we tested on, fixed as 3 in our experiments

Require: A set of hyperparameters profiles \mathcal{P} .

for Profile $P \in \mathcal{P}$ **do**

Set $\text{Map}_m[P] = 0$ to record the sample needed to perform reduction.

Set low and high index to perform a binary search, $Left = 0$ and $Right = L$.

while $Left \neq Right$ **do**

 Middle index $Mid = \lfloor (Left + Right)/2 \rfloor$.

 Set sample number $m = m_{Mid}$.

 Set success Count $c = 0$.

for $r \leq Repeat$ **do**

 Randomly generate secret s with Hamming weight $\lfloor n\tau \rfloor$

 Sample a dataset \mathcal{S} generated with secret s and error rate τ , with size m .

 Run a variant of Algorithm 9 with hyperparameter profile P

 Increment success count c by 1 if the return value equals $s[d - 1]$.

end for

if $c \geq \lfloor 2Repeat/3 \rfloor$ **then** $Right = Mid$

end if

if $c < \lfloor 2Repeat/3 \rfloor$ **then** $Left = Mid + 1$

end if

end while

 Set $\text{Map}_m[P] = m_{Left}$.

end for

return $(P, \text{Map}_m[P])$ with the smallest $\text{Map}_m[P]$.

3.3 Moderate Sample

Setting 1.2 and Setting 1.1 can be viewed as two extreme settings where in the first one only time complexity is considered and in the second one only sample complexity is considered. Although they both show interesting properties and strong performances, difficulties are faced when trying to fit them in the classical *reduction-decoding* paradigm of LPN algorithms, in which the sample complexity after reduction is mediocre. This calls for investigating whether and how we can apply machine learning methods as a part of the *reduction-decoding* paradigm, especially as the decoding algorithm.

Our work gives an affirmative answer to the first question and proposes the following Algorithm 11 as a candidate. We would like to point out that the algorithm design process of Algorithm 11 centered around LPN problems with medium dimension $n \approx 20$ and heavy error rate $\tau > 0.495$, as this is the setting after the reduction phase where classical algorithms like Pooled Gaussian and MMT require huge time complexity. Similar to the previous section, a meta algorithm is required to determine the running time t_{time} and predicted accuracy τ' . The algorithm is analogous to Algorithm 8 and is omitted here.

One can see the major differences between Algorithm 6 and Algorithm 11 is that

Algorithm 11 Moderate Sample Algorithm

Require: \mathcal{S} with m samples drawn from LPN with dimension n and error rate τ .

Require: Repeat number $repeat$ for starting with different initializations, set to 1 in our experiments.

Require: Repeat number $repeat_{post}$ for post processing run, set to 20 in our experiments.

Require: The size of boosting set m' for post processing run.

Require: The hypothesis test error rate threshold τ' .

$$\text{Set } m_1 = \frac{2n}{(1/2-\tau)^2}.$$

▷ See Lemma 3 in [EKM17]

Split the dataset with a training set $\mathcal{S}_{\text{Train}}$ of size $m - m_1$ and a test set $\mathcal{S}_{\text{Test}}$ of size m_1 .

for $r \leq repeat$ **do**

 Run Algorithm 1 with the parameters specified in Tables 3 and 4 to get a learned model $\mathcal{M}[W_T]$.

 Randomly generate m' boolean vector $x_i \in \{0, 1\}^{n+1}$ and use $\mathcal{M}[W_T]$ to predict the pseudo label $1[\mathcal{M}[W_T](x_i[1 : n]) > 0.5] + x_i[n + 1] \pmod 2$ to form $\mathcal{S}_{\text{Boost}}$.

▷ Rebalance Step

 Run Pooled Gaussian algorithm [EKM17] on $\mathcal{S}_{\text{Boost}}$ for $repeat_{post}$ number of times to get a set of possible secret (discard the last bit) using hypothesis test error rate threshold τ' .

▷ Post Processing Step

 Return the secret if one of them reaches accuracy $1 - \tau - \sqrt{\frac{3(\frac{1}{2}-\tau)n}{m_1}}$ on $\mathcal{S}_{\text{Test}}$.

end for

1. The sampler is replaced by the fixed batch sampler. We use this bootstrapping technique to simulate sampling from \mathcal{D} with the sample we possess.
2. The direct inference is replaced by a pooled Gaussian step. This is because the training part of Algorithm 11 typically returns a model with low accuracy, for example around 52% on clean data when $n = 20, \tau = 0.498$ and $m - m_1 = 1e8$. We empirically observe that the rebalance and post-processing step usually takes less than 2 minutes to recover the correct secret under the typical setting. The experiment details are deferred to Section 4.

To simulate the training dynamics of Algorithm 11 with Algorithm 6, we propose to use the same hyperparameter setting except for the sampler. The initial sample m under this setting is mostly determined by the reduction phase. For example, in our case study in Section 4, it is set to $1e8$ as the BKW algorithm reduces LPN problem with $n = 125, \tau = 0.2, m = 1.2e10$ to LPN problem with $n = 26, p = 0.498, m = 1.1e8$ in about 0.5 hours with 128 cores server.

4 Experiment

In this section, we perform experiments on our methods and classical methods like BKW and Gauss as well. Without otherwise mention, the experiment results we report are conducted on 8 NVIDIA 3090 GPUs, which cost approximately 10000 dollars. For other experiments using CPUs, we use a 128 CPU cores server with 496GB memory. Specifically, the server contains two AMD EPYC 7742 64-Core Processors, which cost approximately 10000 dollars as well.

This section is organized as follows. The results for three settings are shown respectively in Sections 4.1 to 4.3. In each subsection, we will first show our method’s overall performance with a comparison

$n \backslash \tau$	0.4	0.45	0.49	0.495	0.498
20	39	70	197	323	730
30	139	374	1576		

(a) **Neural Network**

$n \backslash \tau$	0.4	0.45	0.49	0.495	0.498
20	0.40	5.76	22.0	312	6407
30	26.4	682			

(b) **Gaussian Elimination**

Table 6: **Time Complexity w.r.t. Dimension and Noise Rate.** Each entry represents the running time (in seconds) for the corresponding algorithm to solve the corresponding LPN instance with constant probability. For a neural network, the criterion for solving the LPN is that the accuracy of the network reaches 80% on clean data. For Gaussian Elimination, the criterion for success is to get at least 7 correct secrets out of 10 attempts. The experiments presented in Table 6a are performed on a single GPU and in Table 6b on a single 64 cores Processor. The running time here is averaged over all runs that recover the correct secret in the time limit (3 hours). For the empty cell, no runs are successful within 3 hours.

to classical algorithms. Then we will present a case study section, in which pilot experiments are demonstrated to show the experiment observation we have under the setting. It will be followed by a hyperparameter selection section, in which we showcase how our meta hyperparameter selection algorithm is performed and also our conclusion on the importance of different hyperparameters.

4.1 Abundant Sample Setting

In this section, we study the empirical performance of Algorithm 6. The performance of Algorithm 6 is summarized in Table 6a. As a comparison, we list the performance of Gaussian Elimination in Table 6b. The hyperparameters may not be optimal for all n and τ . Nevertheless, it loosely reflects how the performance of Algorithm 6 varies with n and τ . A key observation is that the performance does not deteriorate much when we raise the noise. For instance, the runtime only triples when raising τ from 0.45 to 0.49. However, the running time rises quickly when we raise the dimension. In contrast, the running time of Gaussian Elimination rockets with both dimension and noise rate. These features makes Algorithm 6 competitive in medium dimension with super high noise cases (note that LPN instances with medium dimension and super high noise often appear at the final stage of the BKW reduction or other dimension-reduction algorithms). In Section 4.1.1, we will first illustrate typical phenomena that arise when running the algorithm. In Section 4.1.2, we show how to tune the algorithm to its best performance.

4.1.1 Case Study

We run Algorithm 6 on $\text{LPN}_{44,\infty,0.2}$. The training accuracy and the test accuracy are shown in Figure 2.

Empirical Conclusion 1. *Under Setting 1.1, we find that (i) The randomness in initialization has little impact on the running time or the final converged accuracy of the model. (ii) The training and test accuracy tie closely and test accuracy usually reaches 100% eventually on clean data after it departs from 50%.*

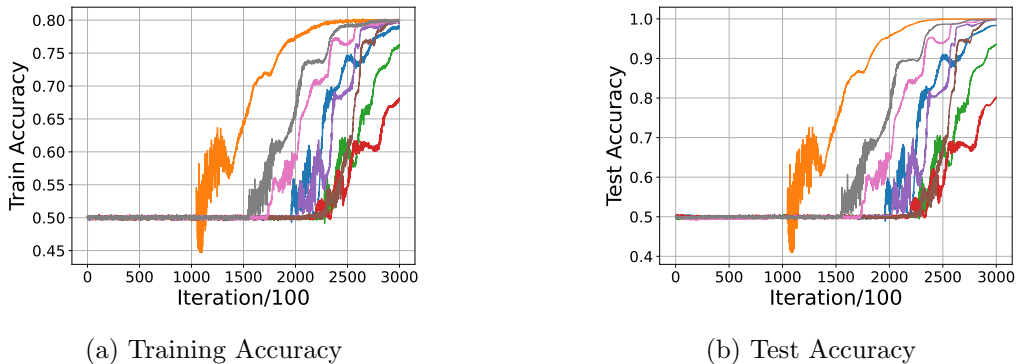


Figure 2: **Experiments on $\text{LPN}_{44, \infty, 0.2}$.** Figures (a) and (b) show the training accuracy and the evaluation accuracy respectively. The horizontal axis represents the training iteration. One unit on this axis represents 100 iterations. The vertical axis represents the accuracy of the model on the corresponding dataset. The training accuracy here is calculated on the batch used for training the network before the gradient is used to update the weight. There are 8 curves in every graph, each corresponding to a random initialization. The same type of graphs will appear frequently in this paper.

Initialization We use Kaiming initializations. As depicted in Figure 2, it is observed that whether a model can learn the task is independent of the randomness in the Kaiming initialization. We also note that the running time needed to learn the LPN instance does not differ much among different initializations. Considering this phenomenon, we do not try multiple initializations in designing Algorithm 6.

Matching Training and Test Accuracy. One can observe an almost identical value between the training and test accuracy in Figure 2. This is expected because, on each train iteration, fresh data is sampled so the training accuracy, similar to the test accuracy, reflects the population accuracy. It is further observed that after the test accuracy starts to depart from 50%, it always reaches 100% accuracy eventually (here 100% refers to the test accuracy on noiseless data). This phenomenon suggests that most local minima of $\mathcal{L}_{\mathcal{D}}(f)$ found by algorithm 1 are global minima of $\mathcal{L}_{\mathcal{D}}(f)$.

Direct Inference of the Secret We experiment with models with accuracy 80% on the post processing step of Algorithm 6 and find out that the post processing step returns secret with high probability with typical time complexity less than 1s.

4.1.2 Hyperparameter Selection

In this section, we fix an LPN problem setup with $n = 20$ and $\tau = 0.498$ and showcase the application of our Algorithm 8. Recalling our goal is to find hyperparameters that minimize the time complexity, we report the minimal running time in seconds for the model to reach accuracy 80% on noiseless data for one out of the four datasets we experimented on. For each experiment, we fix all except the investigated hyperparameters the same as in the corresponding column of Algorithm 6 in Tables 3 and 4.

Empirical Conclusion 2. *Under Setting 1.1, we find that (i) Larger batch size has tolerance for larger learning rate, however, all sufficiently large batch size yields similar running time. (ii) The depth of the*

network should be fixed to 1 and the width should be carefully tuned.

Learning Rate and Batch Size The performance of algorithm Algorithm 6 under different learning rates and batch sizes are listed in Table 7. For a fixed batch size, if the learning rate is too small, the model will fit slowly to prolong the learning time. If on contrary, the learning rate is too large, the gradient steps would be so large that the learner cannot locate the local minima of the loss. The upper threshold of the learning rate increases with respect to the batch size because, for a fixed learning rate, a small batch size would result in an inaccurate gradient estimate and render the learning to fail. On the contrary, a large batch size can make the running time unnecessarily long. The best running times for all batch sizes are similar and for all learning rates are similar as well. As a result, we would recommend one first finds a batch size and learning rate combination that can fit the data and then tune one of the factors while fixing the other for higher performance.

Learning Rate \ \log_2 Batch Size	17	18	19	20
6×10^{-5}	1807	2405	3569	4960
2×10^{-4}	806	1140	1389	1650
6×10^{-4}	> 4000	> 4000	692	970
2×10^{-3}	> 4000	> 4000	> 4000	750
6×10^{-3}	> 4000	> 4000	> 4000	723

Table 7: **Time Complexity w.r.t. Learning Rate and Batch Size.** Each entry represents the running time (in seconds) for Algorithm 6 with corresponding hyperparameters to solve LPN_{20,∞,0.498} with probability approximately 1/4.

L2 Regularization L2 regularization does not help to learn at all and increases running time over all settings we experimented on. As mentioned in Section 4.1.1, generalization is not an issue under this setting hence L2 regularization is unnecessary.

Width and Depth of the Model We experiment with different architectures including MLP models with different widths in {500, 1000, 2000} and depths in {1, 2, 3}. The base model we apply outperforms other architectures significantly and returns the correct secret at least 5 times faster. This experiment, alongside the following architecture experiment in Section 4.2.2, shows that one should tend to use a shallow neural network with depth 1 and carefully tune the width of the model.

4.2 Restricted Sample Setting

In this section, we study empirically the minimal sample complexity required by neural networks to learn LPN problems. The investigation here is mostly limited to the case where the noise is low, as a complement to the cases in Settings 1.1 and 1.3. As mentioned in Section 3.2, we utilize Algorithm 9, which aims to solve the decision version of the problem. The performance of Algorithm 9 is shown in Table 8. These sample complexities are typically comparable with the classical algorithm, such as

BKW. In Section 4.2.1, we will show some important observations that guide us in designing Algorithm 9. In Section 4.2.2, we will show how to use Algorithm 10 to find hyperparameters of the algorithm 9.

Dimension	Error Rate		
	0.1	0.2	0.3
25	7	10.5	13
30	9.5	12.5	16
35	9.5	15	17.5
40	10.5	16	20.5

Table 8: **Sample Complexity w.r.t. Dimension and Noise Rate.** Each entry represents the logarithm of the minimal number of samples with base 2 for Algorithm 9 to return the correct guess of the last bit of the secret with a probability of approximately $2/3$.

4.2.1 Case Study

Setup	n	τ	m	Sparsity
1	44	0.2	2^{17}	0.2
2	29	0.2	2^{17}	0.2
3	30	0.2	2^{17}	0.5

Table 9: **Three Different Setups Shown in Figure 3.** The sparsity column means the hamming weight of the secret over n .

We first illustrate the common phenomena observed under Setting 1.2 that guide us in designing Algorithm 9.

We conducted experiments with neural networks using Algorithm 1 under three setups specified in Table 9. The hyperparameters follows Tables 3 and 4, and the step threshold t_{step} is set to $300k$. To test the effectiveness of using sparse secrets (cf. Lemma 2.2), we vary the sparsity of the secret (Hamming weight / dimension) in the third setup. We plot the test accuracy of the model with respect to training iterations in Algorithm 6. Notice here we test on LPN data directly hence the highest possible population accuracy is 80%.

Empirical Conclusion 3. *Under Setting 1.2, we find that (i) The sparse secret makes training easier. (ii) Initialization significantly affects the performance of the converged trained model. (iii) The reduction to the decision problem enables us to learn LPN instances with larger n . (iv) Typically there is a significant gap between training and test accuracy under this setting, which calls for methods to improve generalization.*

Sparsity of Secret Comparing the first and the last row of Figure 3, Algorithm 9 succeeded in solving $\text{LPN}_{44,2^{17},0.2}$ with secrets of sparsity 0.2 for two out of three datasets, while completely fail on solving $\text{LPN}_{30,2^{17},0.2}$ with secrets of sparsity 0.5. Hence, using sparse secret greatly reduce the hardness

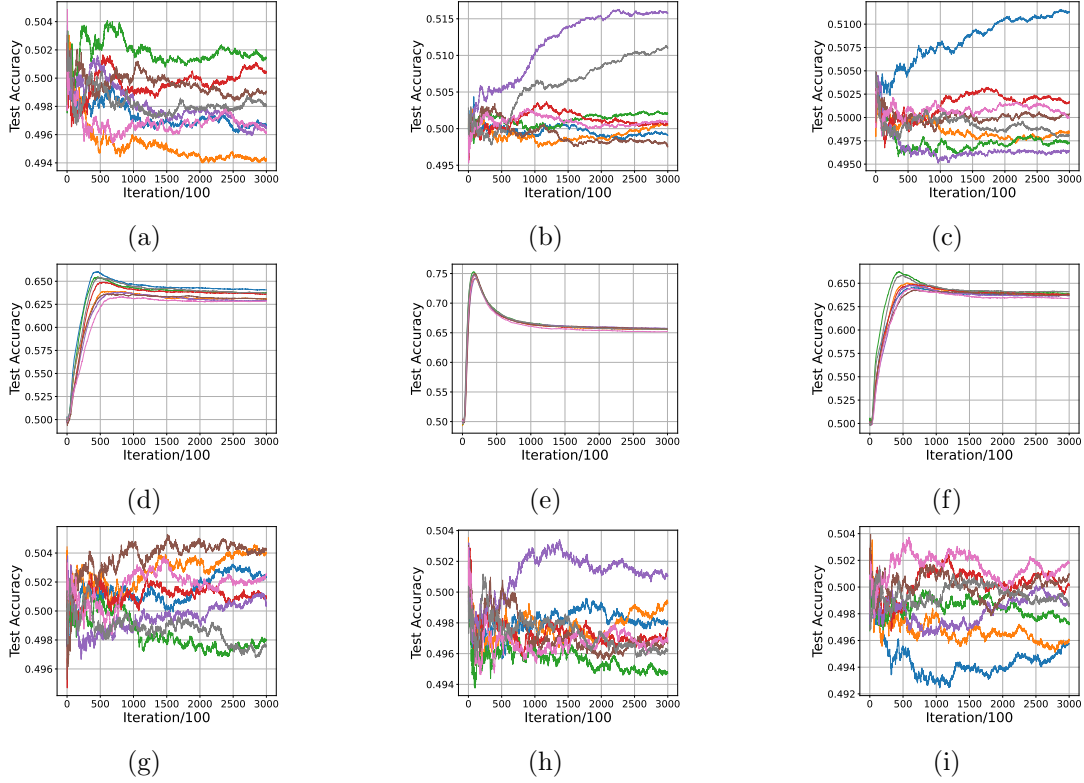


Figure 3: **Performance of the Models** These nine graphs show the performance of models during training under three different setups specified in Table 9. The first, second, and third row corresponds to Setup 1, 2, and 3 respectively. Different pictures in the same row correspond to different datasets. Different colors in the same picture correspond to different initializations of the networks.

of learning. This implies we should always apply Lemma 2.2 to reduce the secret to a sparse secret when the noise level is low.

Initialization Matters Graph (a),(b), and (c) in Figure 3 show that initialization can affect whether a neural network can distinguish the correct and the wrong guesses. This shows the necessity to try different initialization on the same dataset (possibly in parallel).

Necessity to Transfer to Decision Problem Previously we have established that Algorithm 1 can solve the decision version of $\text{LPN}_{44,m,0.2}$. Graphs (d), (e), (f) show that even if we reduce n to 29, direct learning cannot yield a model that perfectly simulates the LPN oracle without noise, in which case the test performance should get close to 0.8. However, all the trained models only yield an accuracy lower than 0.7. Hence with the reduction, we can solve an LPN instance with larger n with the same m and τ .

Divergence of Training and Test Accuracy In Figure 4, we plot a run of Algorithm 1 on LPN data when the sample complexity is greatly limited. Because the size of the training set is very limited,

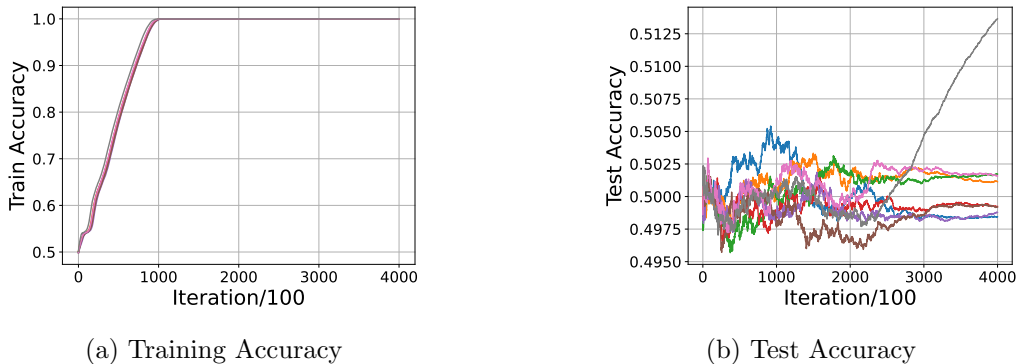


Figure 4: **Experiments on $\text{LPN}_{44,185362,0.2}$** The legend in this figure is the same as in Figure 2. We can observe a large deviation between the training and test accuracy curve. In fact, for the initialization that the accuracy boosts the highest, the accuracy increases after all the training data are correctly predicted, or in other words, memorized.

the model easily reaches 100% accuracy on the training set. This phenomenon is known as *overfitting* in the machine learning community. However, one can observe that the test accuracy of one of the model in (b) boost after the model overfits the training set. Though strange at first sight, this phenomenon, known as *grokking* where validation accuracy suddenly increases after the training accuracy plateau is common in deep learning experiments and repetitively appears in our experiments on LPN data. The exact implication of this phenomenon is still opaque. At the very least, we can infer from the figures that *generalization*, instead of optimization is the key obstacle in Setting 1.2.

4.2.2 Hyperparameter Selection

In this subsection, we fix an LPN problem setup with $n = 44$ and $\tau = 0.2$ and illustrate the conclusion of our hyperparameter selection process based on Algorithm 10. Recalling our goal is to find hyperparameters that minimize the sample complexity, we report $\log_2 \# \text{ Sample}$ for our reduction algorithm to successfully return the secret with probability approximately equal to $2/3$ following the convention of [BTV16]. The hyperparameter profiles we considered in our experiments are identical except for the investigated hyperparameter as Tables 3 and 11. The aggregated results are shown in Table 10.

Empirical Conclusion 4. *Under Setting 1.2, the architecture of the model, learning rate, and L2 regularization are the most important components in determining the sample complexity. In the meantime, tuning batch size or applying other regularization methods are in general not helpful.*

Depth of Network In contrast to the trend of using *deep* neural network, the depth of the network on LPN problem should not be large. Based on Table 10a, we find depth 1 neural network performs significantly better than any larger depth network.

Width of Network Similar to the case in Section 4.1, width of the network is critical, as shown in Table 10b. However, the blessing here is that the width is not very versatile in this setting. The

1	2	3	500	800	1000	1200	1500	2000	ReLU	Sigmoid	COS
17	19	>19	>19	17.5	17	17.5	18.5	17.5	17	18.5	18
(a) Depth			(b) Width					(c) Activation			
MSE	Logistic	0	$5e-4$	$1e-3$	$2e-3$	$3e-3$	$6e-3$	Full	Full/2	Full/4	
18	17	17	17	16.5	16	16.5	> 19	17	17	17	
(d) Loss			(e) L2 regularization					(f) Batch Size			
$1e-5$	$2e-5$	$1e-4$	$2e-4$	$1e-3$	$2e-3$	$1e-2$	$2e-2$	$1e-1$			
18.5	17	17	18	18	18.5	18	18.5	19			
(g) Learning Rate											

Table 10: **Time Complexity w.r.t. Multiple Hyperparameters.** Each entry represents the logarithm of sample complexity with base 2 for Algorithm 9 with corresponding hyperparameters to solve $\text{LPN}_{44,m,0.2}$ with probability approximately $2/3$. The hyperparameter profile we considered in our experiments always shares all except the investigated hyperparameter as Tables 3 and 11.

Sampler	Learning Rate	Weight Decay	Batch Size	Step Threshold t_{step}
Batch	$2e-5$	0	Size of Train Set	$300k$

Table 11: Base Hyperparameters

width must exceed a lower bound for better performance while all the widths larger than that show similar performance. However, the optimal width is still a finite number, in this case, near 1000.

Activation Function We experiment with replacing the activation function of the first layer and find that while all the activation function can make Algorithm 9 work given enough sample, the ReLU activation performs better than other candidates. The results are shown in Table 10c.

Loss We experiment with two types of loss, the MSE loss and the logistic loss in Table 10d. Training using these two loss functions are usually referred to as *regression* and *classification* in machine learning literature. We find out that logistic loss performs slightly better than MSE loss.

Batch Size One may naturally wonders why in Setting 1.3 and Setting 1.1, we both use fix batch size, while in the Setting 1.2, we recommend using full batch training. One important issue to notice is that the batch size in the two other settings is typically larger than the total dataset itself in the current setting. In our experiments with smaller batch sizes, the results are almost identical to those using the full training set size.

Learning Rate We observe a similar phenomenon as in Section 4.1.2 for the learning rate, i.e., it should be neither too large nor too small to achieve the best performance. However, it is worth noticing

that the best learning rate in this regime is typically a magnitude smaller than the best learning rate for Algorithm 6. We have two hypotheses for the reason behind this. First, loosely based on our theory prediction in Theorem 5.3, a higher noise rate calls for a higher learning rate. Second, the dependency between the model weight and the sample we used for each step calls for a smaller learning rate to reduce the caveat of overfitting.

Weight Decay As mentioned in Section 4.2.1, generalization is the key obstacle for neural networks to learn to distinguish LPN data from random data in Setting 1.2. We experiment with several different regularization functions and find out that weight decay, or L2 regularization, is of significance in reducing the sample size. This coincides with previous theoretical prediction Theorem 5.5 in a simpler setting. The best weight decay parameter we find for our settings is $2e - 3$. We observe empirically that when increasing this parameter to $6e - 3$, the weight of the model will quickly collapse to all zeroes. In this sense, the weight decay parameter needs to be large, but not to the extent that interferes with optimization.

Sparsity Regularization We also try regularization methods like L1 regularization and architectures with hard-coded sparsity (by fixing some of the parameters to 0 or using a convolutional neural network), the results are not shown in Table 10 because all these methods increase the required sample complexity by a large amount. Although in our Theorem 5.1, the ground truth model we constructed is sparse when the secret is sparse, the ways we utilize sparsity hurt the optimization performance of Algorithm 1. It remains open whether one can design better architectures or find other regularization functions that better suit the LPN problems.

4.3 Moderate Sample Setting

In Setting 1.3, we aim at finding ways to use neural networks as a part of the classical *reduction-decoding* scheme of LPN solving. We have observed in Section 4.1 that neural networks show resilience to noise when the dimension is small, which is further validated by Theorem 5.3. It would then be tempting to use neural networks in the decoding phase where typically the dimension is low and the noise rate is high. However, to make neural networks feasible for solving the problem after reduction, we can no longer assume the number of training samples is infinite. Instead, we should utilize all the training samples generated after the reduction phases and try reducing the time complexity based on the samples we are given.

In this section, we use Algorithm 11 to solve for the last 26 bits of the secret for LPN instance $\text{LPN}_{125,1.2e10,0.25}$ in **106 minutes**. As a comparison, in [EKM17], more than 3 days are taken to recover the 26 bits of the secret. We believe this result shed light on the potential of utilizing neural networks and GPUs as components of faster solving algorithms of LPN in practice.

In Section 4.3.1, we show the exact time component for solving the last 26 bits as well as important phenomena we observe under Setting 1.3. In Section 4.3.2, we mainly discuss the hyperparameter tuning required for the post-processing part of Algorithm 11. We will show that the hypothesis testing threshold τ' is not very versatile for the algorithm performance and hence is easy to tune.

4.3.1 Case Study

We will first introduce how we solve for the last 26 bits of the LPN instance $\text{LPN}_{125,1.2e10,0.25}$. On our server with 128 cores, we can reduce this problem to $\text{LPN}_{26,1.1e8,0.498}$ in 40 minutes. We will then enumerate the last 6 bits of the secret to further reduce the problem to $\text{LPN}_{20,1.1e8,0.498}$ and apply our Algorithm 11 on it. We specified the time constraint t_{time} in Table 4 for Algorithm 11 as 20 minutes and hypothesis testing threshold τ' as 0.483 under this setting. We further restricted the running time of the rebalance and post-processing step by 2 minutes. By using 8 3090 GPUs to enumerate secret in parallel, we try the correct postfix in the third round of enumeration and the post-processing step return the right secret in the pool of secrets. The applications of enumeration and Algorithm 11 take less than 66 minutes. In total, we solve for the last 26 bits in less than 106 minutes.

In designing Algorithm 11, two primary factors affect the performance. The first factor is the probability of Algorithm 1 procedure returns model with high accuracy on test data, which we dub as *success rate*. The second factor is the accuracy the model can reach given the time limit under such a scenario, which we dub as *mean accuracy*. The repeat number for starting with different initializations *repeat* mainly depends on the success rate and the mean accuracy affects the time complexity of our post-processing step. We will now introduce some important phenomena we observe in applying Algorithm 1 on LPN instance with dimension 20 and error rate 0.498 that guides us in selecting the hyperparameters of Algorithm 11.

Empirical Conclusion 5. *Under Setting 1.3, we find that (i) Both the success rate and the mean accuracy increases with the size of the training set. (ii) An imbalance in the output distribution exists in all models we trained in this setting (predicting one outcome with a significantly higher probability than the other over random inputs) (iii) Different than the case in Setting 1.2, the training accuracy typically plateau at a low level.*

The Size of Training Set We first show sample complexity affects these two factors mutually. As most of the samples consumed are used to perform Algorithm 1, we experiment with varying the sample in Table 12. It is observed that both the success rate and the mean accuracy increase with the sample complexity. Under the scenario where $1e8$ samples are provided, the success rate reaches 87.5%, allowing us to set *repeat* as 1 to get a final algorithm that succeeds with high probability. Given the randomness here is over the initialization of the model instead of the dataset itself, our results also show that we can compensate for the lack of samples by trying more initializations.

Discussion on Output Distribution Reader may find the rebalance step in Algorithm 11 seemingly unnecessary. However, we observe in experiments that the trained neural network typically has a bias towards 1 or 0. The proportion of two different results given random inputs can be as large as 1.5. This fact makes the testing of Gaussian Elimination steps hard and leans toward providing a wrong secret. We mitigate this effect by performing the rebalance step and believe that if this phenomenon can be addressed through better architecture designs or training methods, the performance of neural network based decoding method can be further boosted.

Optimization Obstacle We plot the training and test accuracy curve for applying Algorithm 1 over $\text{LPN}_{20,1e8,0.498}$ in Figure 5. Here the test is performed on clean data to provide better visualization.

m	Success Rate	Mean Accuracy
$4e7$	25.0%	51.1%
$6e7$	62.5%	51.7%
$8e7$	62.5%	51.8%
$1e8$	87.5%	52.2%

Table 12: **Success Rate and Mean Accuracy w.r.t. Training Set Size m on $\text{LPN}_{20,m,0.498}$.** We define a successful run as a model with random initialization that converges to a model with an accuracy greater than 51.5% on clean data in the time limit. The success rate means the probability of a run being successful over the randomness of initialization. The mean accuracy is only calculated on successful runs. We clearly observe that with an increasing number of samples, both the success rate and the mean accuracy increase significantly.

Different from Setting 1.2, where the training set is greatly limited and the network can overfit the training set, here we observe that the network can hardly fit the training data. This fact has two sides. Firstly, it shows that overfitting ceases given enough samples and generalization won't be a severe issue. However, as we observe in Figure 4, the generalization performance may boost when overfitting happens, hence the low training accuracy may also be a reason behind the low final converged accuracy.

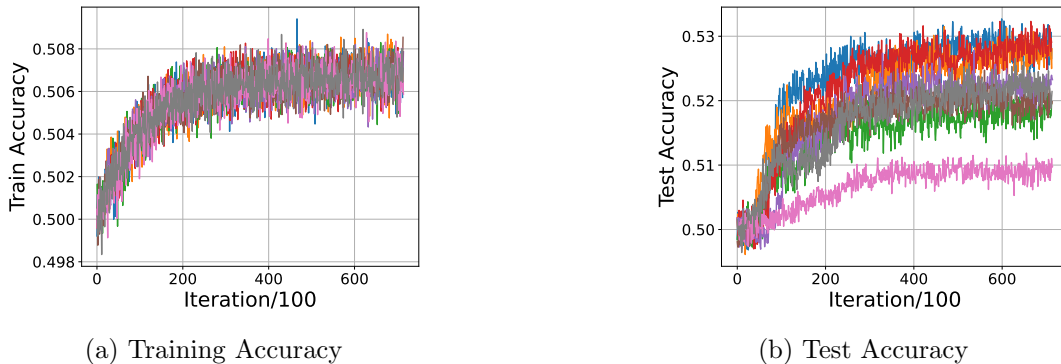


Figure 5: **Experiments on $\text{LPN}_{20,1e8,0.498}$** We observe that the training accuracy is low, showing a vast contrast with Setting 1.2. One should notice here the test accuracy is over clean data while the training accuracy is over the noisy training set.

4.3.2 Hyperparameter Selection

For the ease of hyperparameter tuning, we use hyperparameters for Algorithm 1 we found in Section 4.1.2 and our experiments show that this choice can already let us get models with enough accuracy to perform boosting. Hence here we will focus on discussing the hyperparameters for the post processing the Gaussian Elimination step.

Empirical Conclusion 6. *Under Setting 1.3, we find that (i) Hyperparameters for the network training can follow the hyperparameters under Setting 1.1. (ii) The hypothesis testing threshold τ in Algorithm 11 has large tolerance and has little impact on time complexity even when it's 1% larger than the ground-truth error rate in the test set for the post processing step.*

Hypothesis Testing Threshold In the post processing step, the samples are labeled by the neural networks and are in this sense *free*. Under our experiments, we choose the size of boosting set $m' = 231072$, with 131072 samples as the sample pool where Gaussian elimination input are sampled and 100000 samples for hypothesis testing, which are more than sufficient for the estimated error rate ≈ 0.48 .

The remaining question for hyperparameter tuning may be that how should we set the hypothesis testing threshold given that it may be hard to estimate on the fly. We propose to use a meta run to estimate the converged accuracy. As one can observe that from Figure 5, the final converged accuracy of successful run (with final accuracy greater than 51.5%) center closely.

We further show through experiments that the hypothesis testing threshold τ' in the Gaussian step are not critical for the performance of Algorithm 11 in Table 13. The estimation of hypothesis threshold can be off by almost 1% while still returning the correct secret in 20 repetitive pooled Gaussian run. However, the results in this table also shows that the final testing on noisy real data is necessary as all the pools of secrets contain wrong secrets. The reader should notice that the final testing step takes almost negligible time, as for each model, at most 20 secrets need to be tested.

0.480	0.483	0.486	0.489
6	6	1	2

Table 13: **Number of Occurrences of the Correct Secret in 20 runs w.r.t. Estimate Error Rate τ'** A subtlety in hyperparameter tuning for Algorithm 11 is that in general τ' is hard to know precisely. We propose to use a meta run to first estimate τ' of the converged model. As one can infer from Figure 5, the test accuracy of the converged model in successful runs tends to be stable. With the ground truth error rate in the test set being about 0.479, our experiments shows that in fact the post-processing step has high tolerance for the hypothesis testing threshold, ranging from 0.480 to 0.489.

5 Theoretical Understanding

In this section, we will show some primary efforts on understanding the effect of neural networks on LPN problems. Because the general understanding of the optimization power and generalization effect of the neural network is very limited, the results in this section can't fully explain all the features of our algorithms and our hyperparameter choices. However, this theoretical analysis improves our understanding of our empirical findings and hence is recorded here.

This section is organized according to the decomposition in Equation (1). Section 5.1 shows that the representation gap of the model architecture MLP over all loss is optimal over the whole function class $f : \mathbb{R}^d \rightarrow [0, 1]$. Section 5.2 shows that despite the inevitable exponential dependency on the dimension, the time complexity of Algorithm 7 scaled optimally with respect to noise. Section 5.3 introduces some prior results that partially explain why weight decay is a powerful regularization as discovered by our ablation study in Section 4.2. Section 5.4 provides a detailed discussion on the hardness of LPN problem using gradient-based methods.

5.1 Representation Power

The main result of this section is the following theorem, which shows that an MLP with a width equivalent to the input dimension is sufficient to represent the best prediction of the LPN inputs.

Theorem 5.1. *For any continuous loss function $l : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}^+$, dimension n and error rate τ , there exists a weight W for a depth 1 MLP \mathcal{M} with width n , $\sigma_1 = \text{ReLU}$, $\sigma_2 = \text{Sigmoid}$, such that the representation gap $\mathcal{L}_{\mathcal{D}}(f)$ of the specified function $f = \mathcal{M}[W]$ is approximately minimized over the function class $\mathcal{H} = \{f' : \mathbb{R}^d \rightarrow [0, 1]\}$. Quantitatively, for any $\epsilon > 0$, there exists W_ϵ , such that*

$$\mathcal{L}_{\mathcal{D}}(\mathcal{M}(W_\epsilon)) - \min_{f \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(f) \leq \epsilon.$$

Proof. For LPN problems with dimension n , secret s and error rate τ , it holds that

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(f) &= \mathbb{E}_{x \sim U(\mathbb{Z}_2^n)} [\tau \ell(f(x), s^t x \bmod 2) + (1 - \tau) \ell(f(x), (s^t x + 1) \bmod 2)] \\ &\geq \mathbb{E}_{x \sim U(\mathbb{Z}_2^n)} \left[\min_{p \in [0, 1]} \tau \ell(p, s^t x \bmod 2) + (1 - \tau) \ell(p, (s^t x + 1) \bmod 2) \right]. \end{aligned}$$

We will now show that in fact this lower bound can be reached approximately when f is specified by a weight W for a depth 1 MLP \mathcal{M} with width n , $\sigma_1 = \text{ReLU}$, $\sigma_2 = \text{Sigmoid}$.

By Lemma 5.2, there exists a weight W' for a depth 1 MLP \mathcal{M}' with width n , $\sigma_1 = \text{ReLU}$, $\sigma_2 = \mathcal{I}$ such that

$$\forall x \in \{0, 1\}^n, \mathcal{M}'[W'](x) = s^t x \bmod 2.$$

Given l is continuous, for $b \in \{0, 1\}$, we can find $a_{b, \epsilon} \in (0, 1)$, such that

$$\tau \ell(a_{b, \epsilon}, b) + (1 - \tau) \ell(a_{b, \epsilon}, 1 - b) \leq \min_{p \in [0, 1]} \ell(p, b) + (1 - \tau) \ell(p, 1 - b) + \epsilon.$$

Further define $\gamma_{b, \epsilon}$ satisfies $\text{Sigmoid}(\gamma_{b, \epsilon}) = a_{b, \epsilon}$. For W satisfying

$$\begin{aligned} W_1 &= W'_1, \\ b_1 &= b'_1, \\ W_2 &= (\gamma_{1, \epsilon} - \gamma_{0, \epsilon}) W'_2, \\ b_2 &= \gamma_{0, \epsilon}, \end{aligned}$$

it holds that $\mathcal{M}[W](x) = a_{s^t x \bmod 2, \epsilon}$. This implies $\mathcal{L}_{\mathcal{D}}(\mathcal{M}[W]) \leq \epsilon + \min_{f \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(f)$. The proof is complete. \square

The above proof relies on Lemma 5.2, whose proof is as follows.

Lemma 5.2. *For any secret $s \in \{0, 1\}^n$, there exists a weight W for a depth 1 MLP \mathcal{M} with width n , $\sigma_1 = \text{ReLU}$, $\sigma_2 = \mathcal{I}$, such that*

$$\forall x \in \{0, 1\}^n, \mathcal{M}[W](x) = s^t x \bmod 2$$

Proof. Recall in Definition 2.10, the weight W is a list of two matrices tuples specifying the affine transformation on each layer.

We will choose $W_1 = es^t$ with e as the all 1 n -dimensional vector, $b_1 = [0, -1, \dots, -n + 1]^t$. Then we would have

$$\sigma_1(T[W_1, b_1](x)) = \sigma_1((s^t x)e - b_1) = [\max\{(s^t x) - i + 1, 0\}]_{i \in [1:n]}^t.$$

We will further define $b_2 = 0$ and recursively $W_2 \in \mathbb{R}^n$ as

$$\begin{aligned} W_{2,1} &= 1. \\ W_{2,i} &= i \pmod 2 - \sum_{j=1}^{i-1} W_{2,j}(i - j + 1). \quad d - 1 \geq i \geq 2 \end{aligned}$$

It is then easy to check

$$T[W_2, b_2](\sigma_1(T[W_1, b_1](x))) = \sum_{i=1}^n W_{2,i} \max\{(s^t x) - i + 1, 0\} = s^t x \pmod 2.$$

The proof is then complete. □

5.2 Optimization Power

While Section 5.1 shows that an MLP with width n is sufficient to minimize the representation gap, it does not show how may we find such representation. As mentioned in Section 2, it is common practice to use the gradient method to optimize neural networks. Pitifully, [AS20] has shown that it requires at least $n^{\Omega(\text{Hamming weight of secret})}$ time complexity to use any gradient method to solve LPN instance. [BEG⁺22] shows through experimental and theoretical analysis that the time complexity to solve LPN instance with noise rate 0 (known as the *parity problem*) with neural networks seems to match this lower bound closely. In this section, we show that despite this exponential reliance on the problem dimension, the training of neural networks is robust to the noise rate τ in Setting 1.1 by the following theorem.

Theorem 5.3. *If there exists weight initialization W_0 , learning rate η , weight decay parameter $\lambda < 1/\eta$, and step threshold T such that Algorithm 7 can return a model with accuracy at least $\gamma > \frac{1}{2}$ on LPN problem with dimension n , secret s and noise rate 0, for any batch size B greater than threshold B_{th} with constant probability $p > 0$, then for any $\gamma' < \gamma, p' < p$, there exists another threshold $B_{th,\tau} = \max\left(O\left(\frac{1}{(1-2\tau)^2}\right), B_{th}\right)$, such that Algorithm 7 can return a model with accuracy at least $\gamma'\tau + (1 - \gamma')(1 - \tau)$ on LPN problem with dimension n , secret s and noise rate τ when batch size $B \geq B_{th,\tau}$ and learning rate $\eta' = \frac{\eta}{1-2\tau}$ with probability p' with all other hyperparameters fixed.*

As the time complexity of Algorithm 7 is $O(BT)$, Theorem 5.3 shows that with neural network can solve the LPN problem with noise rate τ in time complexity $O\left(\frac{1}{(1-2\tau)^2}\right)$ under Setting 1.1, given the underlying parity problem can be solved by the same neural network with constant probability. This rate coincides with the sample complexity of hypothesis testing on whether a boolean vector is s and is in this sense *optimal*.

One would naturally ask whether corresponding results exist under the two other settings. We observe empirically that as apposed to the high converging accuracy in Setting 1.1, under Setting 1.3, the final converging model would have low train accuracy (though greater than 50%) despite the training time. In Setting 1.1, however, the training accuracy can tend to 1 in a short period of time while the testing accuracy increases slowly, in many cases after the convergence of training accuracy. These interesting phenomenons show that the optimization dynamics may be highly different under the three settings and it remains an open question to fully characterize how different sample complexity shapes the optimization landscape of neural networks in the LPN problem.

Proof of Theorem 5.3. We will denote the weight sequence generating by applying Algorithm 7 with batch size B and learning rate η on LPN instance with dimension n and noise rate τ as W_t . Assuming the corresponding batch is $\mathcal{S}_t = \{(x_{t,i}, y_{t,i})\}$ for $t \leq T$. We will further define $\mathcal{S}_{t,\tau} = \{(x_{t,i}, y_{t,i} + f_{t,i} \bmod 2)\}$ where $f_{t,i}$ are independent boolean variables which equals to 1 with probability τ . Finally, we will define the coupled weight sequence $W_{t,\tau}$ as weight sequence generating by applying Algorithm 7 with learning rate $\eta/(1 - 2\tau)$ and batches $\mathcal{S}_{t,\tau}$.

Consider the following sequence of weights,

$$\begin{aligned}\tilde{W}_0 &= W_0 \\ \tilde{W}_t &= \tilde{W}_{t-1} - \eta\lambda\tilde{W}_{t-1} - \eta\nabla_W \mathbb{E}_{x \sim U(\mathbb{Z}_2^n)} l(\mathcal{M}[\tilde{W}_{t-1}](x), s^t x \bmod 2), 1 \leq t \leq T, t \in \mathbb{Z}.\end{aligned}$$

By standard approximation theorem, when the batch size B tends to infinity, we would have $W_t \rightarrow \tilde{W}_t$ in probability. This implies \tilde{W}_t corresponds to a function with accuracy at least γ on LPN data without noise. We can choose r small enough such that for any weight W in r neighbors of \tilde{W}_T all correspond to a function with accuracy at least γ' on LPN data without noise. Suppose now when $B \geq B_r > B_{th}$, we would have with probability $1 - \frac{p-p'}{2}$, $\|W_t - \tilde{W}_t\| \leq \frac{r}{2}$. Here \tilde{W}_l and \tilde{b}_l are the corresponding weight and bias of \tilde{W} . We would now choose a compact convex set \mathcal{C} large enough such that it contains $\{\tilde{W}_t\}_{t \in [0:T]}$ and their r -neighbor. By our assumption, we would have for any fixed $x \in \{0, 1\}^n, y \in \{0, 1\}^n$, and $w \in \{W_1, W_2, b_1, b_2\}$, $\frac{\partial l(\mathcal{M}[W](x), y)}{\partial w}$ is first-order differentiable functions, hence we can assume that there exists constant C_1 and C_2 , such that,

$$\begin{aligned}C_1 &\geq \max_{W=(W_1, b_1), (W_2, b_2)} \in \mathcal{C} \max_{w \in \{W_1, W_2, b_1, b_2\}} \max_{x \in \{0, 1\}^n, y \in \{0, 1\}} \left\| \frac{\partial \mathcal{M}[W](x_i)}{\partial w} \right\|_{\infty} \\ C_2 \|W_a - W_b\|_2 &\geq \max_{x \in \{0, 1\}^n, y \in \{0, 1\}} \left\| \frac{\partial l(\mathcal{M}[W](x), y)}{\partial W} \Big|_{W=W_a} - \frac{\partial l(\mathcal{M}[W](x), y)}{\partial W} \Big|_{W=W_b} \right\|\end{aligned}$$

We will first fix ϵ be a small constant such that $e^{C_2 T} \epsilon \leq \frac{r}{2}$ and then choose $B_{th,\tau} = \max\left\{\frac{2 \ln T + \ln nd - \ln(p-p')}{(1-2\tau)^2} \frac{32ndC_1^2}{\epsilon^2}, B_r\right\}$.

When $B \geq B_{th,\tau}$, we will inductively prove that for step $t \leq T$, event $E_t : \|W_{t,\tau} - W_t\|_2 \leq e^{\eta C_2 t} \epsilon$ happens with probability $1 - \frac{(p-p')t}{2T} - \frac{(p-p')}{2}$.

We would first suppose $\|W_t - \tilde{W}_t\| \leq \frac{r}{2}$ by the definition of B_r . Suppose E_t happens, as $e^{\eta C_2 t} \epsilon \leq \frac{r}{2}$, we would have the $W_{t,\tau} \in \mathcal{C}$. By Lemma 5.4, it holds that with probability $1 - (p - p')/2T$, for any parameter w in $W_{t,\tau}$, it holds that

$$\left\| \frac{1}{1-2\tau} \frac{\partial \frac{1}{B} \sum_i l(\mathcal{M}[W](x_{t,i}), f_{t,i} + y_{t,i} \bmod 2)}{\partial w} - \frac{\partial \frac{1}{B} \sum_i l(\mathcal{M}[W](x_{t,i}), y_{t,i})}{\partial w} \right\|_2 \leq \frac{\epsilon}{4}.$$

Considering the update rule of the SGD optimizer,

$$\left\| W_{t+1,\tau} - W_{t,\tau} - \eta \lambda W_{t,\tau} - \eta \frac{\partial \frac{1}{B} \sum_i l(\mathcal{M}[W](x_{t,i}), y_{t,i})}{\partial W} \Big|_{W=W_{t,\tau}} \right\|_2 \leq \epsilon.$$

This then implies

$$\begin{aligned} \|W_{t+1,\tau} - W_{t+1}\|_2 &\leq \epsilon + \|W_{t,\tau} - W_t\|_2 + \eta C_2 \|W_{t,\tau} - W_t\|_2 \\ &\leq \epsilon + \exp(\eta C_2) \exp(\eta C_2 t) \epsilon \\ &\leq e^{\eta C_2(t+1)} (t+1) \epsilon. \end{aligned}$$

The induction is then complete.

Considering the induction conclusion when $t = T$ and combining the definition of r , the proof is complete. \square

Lemma 5.4. *With loss function l being the MAE loss and \mathcal{M} being a one-layer MLP with weight d and smooth activation function, suppose $(x_i, y_i)_{i \in [1:B]}$ are i.i.d sample from LPN problem with dimension n and noise rate 0. Suppose further f_i are i.i.d random variables following distribution $p(f_1 = 0) = 1 - p(f_1 = 1) = \tau$, for any weight $W = ((W_1, b_1), (W_2, b_2))$ for \mathcal{M} , it holds that for $w \in \{W_1, W_2, b_1, b_2\}$, with probability $1 - 8nd \exp\left(\frac{-\epsilon^2(1-2\tau)^2 B}{2ndC^2}\right)$,*

$$\left\| \frac{1}{1-2\tau} \frac{\partial \frac{1}{B} \sum_{i=1}^B l(\mathcal{M}[W](x_i), f_i + y_i \bmod 2)}{\partial w} - \frac{\partial \frac{1}{B} \sum_{i=1}^B l(\mathcal{M}[W](x_i), y_i)}{\partial w} \right\|_2 \leq \epsilon.$$

Here $C = \max_{w \in \{W_1, W_2, b_1, b_2\}} \max_{x \in \{0,1\}^n, y \in \{0,1\}} \left\| \frac{\partial \mathcal{M}[W](x_i)}{\partial w} \right\|_\infty$.

Proof. Denote $n_i = 2f_i - 1 \in \{-1, 1\}$. The key observation of this proof is that for the MAE loss, it holds that

$$\frac{\partial l(\mathcal{M}[W](x_i), f_i + y_i \bmod 2)}{\partial w} = n_i \frac{\partial l(\mathcal{M}[W](x_i), f_i)}{\partial w}.$$

Now consider any index k of w (k can be two dimensional for a matrix), $n_i \frac{\partial l(\mathcal{M}[W](x_i), f_i)}{\partial w_k}$ are bounded random variable in $[-C, C]$, then by Hoeffding's bound, it holds for any $t > 0$

$$\begin{aligned} &\text{Prob} \left(\left| \sum_{i=1}^B n_i \frac{\partial l(\mathcal{M}[W](x_i), f_i)}{\partial w_k} - \mathbb{E} \left(\sum_{i=1}^B n_i \frac{\partial l(\mathcal{M}[W](x_i), f_i)}{\partial w_k} \right) \right| \geq t \right) \\ &\leq 2 \exp \left(-\frac{t^2}{2BC^2} \right). \end{aligned} \tag{2}$$

Now we have

$$\begin{aligned} \sum_{i=1}^B n_i \frac{\partial l(\mathcal{M}[W](x_i), f_i)}{\partial w_k} &= \frac{\partial \frac{1}{B} \sum_{i=1}^B l(\mathcal{M}[W](x_i), f_i + y_i \bmod 2)}{\partial w_k}. \\ \mathbb{E} \left[\sum_{i=1}^B n_i \frac{\partial l(\mathcal{M}[W](x_i), f_i)}{\partial w_k} \right] &= (1-2\tau) \frac{\partial \frac{1}{B} \sum_{i=1}^B l(\mathcal{M}[W](x_i), y_i)}{\partial w_k}. \end{aligned}$$

By Equation (2), by choosing $t = (1-2\tau)\frac{B\epsilon}{\sqrt{nd}}$, we would have with probability $1 - 8nd \exp\left(\frac{-\epsilon^2(1-2\tau)^2B}{2ndC^2}\right)$, it holds that

$$\left\| \frac{1}{1-2\tau} \frac{\partial \frac{1}{B} \sum_{i=1}^B l(\mathcal{M}[W](x_i), f_i + y_i \bmod 2)}{\partial w} - \frac{\partial \frac{1}{B} \sum_{i=1}^B l(\mathcal{M}[W](x_i), y_i)}{\partial w} \right\|_2 \leq \epsilon.$$

The proof is then complete. \square

5.3 Generalization Effect

In modern deep learning theory, explaining the generalization effect of neural network is a long standing open problem. Under Setting 1.1, the generalization gap is naturally zero as the data distribution coincides with the population distribution. However, under Setting 1.3 and Setting 1.2, it would be necessary for us to consider generalization effect. As we observed in Section 4, under Setting 1.3, the generalization gap is still small and the key difficulty lies in the optimization of the neural network. Under Setting 1.2, however, the gap between training accuracy and test accuracy is typically large. This result is expected because as the sample complexity decreases, the optimization landscape over population distribution and data distribution tends to differ. Our experiment shows that applying L2 regularization, or equivalently, weight decay, helps to mitigate this problem and reduce the learning rate.

Although the optimization dynamics of neural networks over the general LPN problem is hard to track, existing literature contains results showing the provable benefit of weight decay on sample complexity under a special setting. We include this result in this subsection for completeness. It remains an open problem to extend this result to secret s with hamming weight proportional to the problem dimension.

Theorem 5.5 (Informal Version of Theorem 1.1 in [WLLM19]). *If a gradient-based optimization algorithm uses SGD optimizer, logistic loss, MLP with depth 1 initialized by Kaiming initialization, weight decay 0, and a small enough learning rate, the sample complexity required to learn the secret of a parity problem with dimension n and Hamming weight 2 secret is $\Omega(n^2)$. However, with proper weight decay constant $\lambda > 0$, the sample complexity of the same problem could be reduced to $\tilde{O}(n)$.*

5.4 Discussion on the Hardness of Using Neural Networks to Solve LPN

As mentioned in Section 5.2, it is proved in [AS20, SSSS17] that it requires $n^{\Omega(k)}$ sample complexity to solve the parity problem without noise using the full batch gradient descent method, which is also the regime our Theorem 5.3 falls into.

However, this hardness constraint fails to hold for the stochastic gradient descent method. In fact [AS20] show that if a family of distribution is polynomial-time learnable, then there exists a polynomial-size neural network architecture, with a polynomial-time computable initialization that only depends on the family of the distribution, such that when stochastic gradient descent with batch size 1 is performed on the network, it can learn the distribution in polynomial time complexity. We would like to remark some implications of these results.

- These results suggest that whether there exists an architecture and initialization scheme such

that Algorithm 7 can solve the LPN problem in polynomial time remains open and is inherently equivalent to whether LPN is in P.

- The construction of the neural network that simulates the polynomial time learning algorithm given by [AS20] relies heavily on deterministic initialization and it conjectures that SGD on random initialization will still require super polynomial time or sample complexity.
- Although the full-batch gradient method will require $n^{\Omega(k)}$ time and sample complexity to solve the parity problem, this does not exclude the gradient method from being used to solve the LPN problem. First, it is still possible to use the gradient method as a building block of a larger algorithm that can effectively solve LPN. Second, given the best exponent component is not known, it is still possible that neural networks will show supreme performance over other classical algorithms, especially in the medium dimension with high noise regime, which our Algorithm 11 tries to address.

References

- [ACPS09] Benny Applebaum, David Cash, Chris Peikert, and Amit Sahai. Fast cryptographic primitives and circular-secure encryption based on hard learning problems. In *CRYPTO*, volume 5677 of *Lecture Notes in Computer Science*, pages 595–618. Springer, 2009.
- [Ala12] Mohammed M Alani. Neuro-cryptanalysis of des and triple-des. In *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part V 19*, pages 637–646. Springer, 2012.
- [Ale03] Michael Alekhnovich. More on average case vs approximation complexity. In *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings*, pages 298–307. IEEE Computer Society, 2003.
- [AS20] Emmanuel Abbe and Colin Sandon. Poly-time universality and limitations of deep learning. *arXiv preprint arXiv:2001.02992*, 2020.
- [BBC⁺19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *CoRR*, abs/1912.06680, 2019.
- [BCG⁺19] Elette Boyle, Geoffroy Couteau, Niv Gilboa, Yuval Ishai, Lisa Kohl, Peter Rindal, and Peter Scholl. Efficient two-round ot extension and silent non-interactive secure computation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 291–308, 2019.
- [BCK18] Amir Bennatan, Yoni Choukroun, and Pavel Kisilev. Deep learning for decoding of linear codes—a syndrome-based approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1595–1599. IEEE, 2018.

- [BEG⁺22] Boaz Barak, Benjamin L Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *arXiv preprint arXiv:2207.08799*, 2022.
- [BFKL93] Avrim Blum, Merrick L. Furst, Michael J. Kearns, and Richard J. Lipton. Cryptographic primitives based on hard learning problems. In *CRYPTO*, volume 773 of *Lecture Notes in Computer Science*, pages 278–291. Springer, 1993.
- [BGPT21] Adrien Benamira, David Gerault, Thomas Peyrin, and Quan Quan Tan. A deeper look at machine learning-based cryptanalysis. In *Advances in Cryptology–EUROCRYPT 2021: 40th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, October 17–21, 2021, Proceedings, Part I 40*, pages 805–835. Springer, 2021.
- [BJMM12] Anja Becker, Antoine Joux, Alexander May, and Alexander Meurer. Decoding random binary linear codes in $2^{n/20}$: How $1 + 1 = 0$ improves information set decoding. In *EUROCRYPT*, volume 7237 of *Lecture Notes in Computer Science*, pages 520–536. Springer, 2012.
- [BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.
- [BLSV18] Zvika Brakerski, Alex Lombardi, Gil Segev, and Vinod Vaikuntanathan. Anonymous ibe, leakage resilience and circular security from new assumptions. In *EUROCRYPT (1)*, volume 10820 of *Lecture Notes in Computer Science*, pages 535–564. Springer, 2018.
- [BTV16] Sonia Bogos, Florian Tramer, and Serge Vaudenay. On solving lpn using bkw and variants. *Cryptography and Communications*, 8(3):331–369, 2016.
- [BV16] Sonia Bogos and Serge Vaudenay. Optimization of \mathsf{LPN} solving algorithms. In *ASIACRYPT (1)*, volume 10031 of *Lecture Notes in Computer Science*, pages 703–728, 2016.
- [CHhH02] Murray Campbell, A. Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1):57–83, 2002.
- [DFB⁺22] Jonas Degraeve, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, Feb 2022.
- [DGHM18] Nico Döttling, Sanjam Garg, Mohammad Hajiabadi, and Daniel Masny. New constructions of identity-based and key-dependent message secure encryption schemes. In *Public Key Cryptography (1)*, volume 10769 of *Lecture Notes in Computer Science*, pages 3–31. Springer, 2018.

- [EHK⁺18] Andre Esser, Felix Heuer, Robert Kübler, Alexander May, and Christian Sohler. Dissection-bkw. In *CRYPTO (2)*, volume 10992 of *Lecture Notes in Computer Science*, pages 638–666. Springer, 2018.
- [EKM17] Andre Esser, Robert Kübler, and Alexander May. LPN decoded. In *CRYPTO (2)*, volume 10402 of *Lecture Notes in Computer Science*, pages 486–514. Springer, 2017.
- [GJL14] Qian Guo, Thomas Johansson, and Carl Löndahl. Solving LPN using covering codes. In *ASIACRYPT (1)*, volume 8873 of *Lecture Notes in Computer Science*, pages 1–20. Springer, 2014.
- [Goh19] Aron Gohr. Improving attacks on round-reduced speck32/64 using deep learning. In *CRYPTO (2)*, volume 11693 of *Lecture Notes in Computer Science*, pages 150–179. Springer, 2019.
- [HB01] Nicholas J. Hopper and Manuel Blum. Secure human identification protocols. In *ASIACRYPT*, volume 2248 of *Lecture Notes in Computer Science*, pages 52–66. Springer, 2001.
- [HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [JEP⁺21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstern, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.
- [KPC⁺11] Eike Kiltz, Krzysztof Pietrzak, David Cash, Abhishek Jain, and Daniele Venturi. Efficient authentication from hard learning problems. In *EUROCRYPT*, volume 6632 of *Lecture Notes in Computer Science*, pages 7–26. Springer, 2011.
- [LF06] Éric Levieil and Pierre-Alain Fouque. An improved LPN algorithm. In *SCN*, volume 4116 of *Lecture Notes in Computer Science*, pages 348–359. Springer, 2006.
- [Lyu05] Vadim Lyubashevsky. The parity problem in the presence of noise, decoding random linear codes, and the subset sum problem. In *APPROX-RANDOM*, volume 3624 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2005.
- [MKS⁺15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmarajan Subramanian, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015.

- [MMT11] Alexander May, Alexander Meurer, and Enrico Thomae. Decoding random linear codes in $\tilde{O}(2^{0.054n})$. In *ASIACRYPT*, volume 7073 of *Lecture Notes in Computer Science*, pages 107–124. Springer, 2011.
- [NML⁺18] Eliya Nachmani, Elad Marciano, Loren Lugosch, Warren J Gross, David Burshtein, and Yair Be’ery. Deep learning methods for improved decoding of linear codes. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):119–131, 2018.
- [Reg09] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*, 56(6):34:1–34:40, 2009.
- [SHM⁺16] David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016.
- [SSSS17] Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, pages 3067–3075. PMLR, 2017.
- [Tes95] Gerald Tesauro. Td-gammon: A self-teaching backgammon program. 1995.
- [VBC⁺19] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, Nov 2019.
- [WCCL22] Emily Wenger, Mingjie Chen, François Charton, and Kristin E. Lauter. SALSA: attacking lattice cryptography with transformers. *IACR Cryptol. ePrint Arch.*, page 935, 2022.
- [WLLM19] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.
- [WS21] Thom Wiggers and Simona Samardjiska. Practically solving lpn. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2399–2404. IEEE, 2021.
- [YZ16] Yu Yu and Jiang Zhang. Cryptography with auxiliary input and trapdoor from constant-noise LPN. In *CRYPTO (1)*, volume 9814 of *Lecture Notes in Computer Science*, pages 214–243. Springer, 2016.
- [ZJW16] Bin Zhang, Lin Jiao, and Mingsheng Wang. Faster algorithms for solving lpn. In *Advances in Cryptology—EUROCRYPT 2016: 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8–12, 2016, Proceedings, Part I 35*, pages 168–195. Springer, 2016.