



# Proceeding Paper Is Quantum Tomography a Difficult Problem for Machine Learning?<sup>†</sup>

**Philippe Jacquet** 

Inria Saclay Ile-de-France, 91120 Palaiseau, France; philippe.jacquet@inria.fr

+ Presented at the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris, France, 18–22 July 2022.

**Abstract:** One of the key issues in machine learning is the characterization of the learnability of a problem. Regret is a way to quantify learnability. Quantum tomography is a special case of machine learning where the training set is a set of quantum measurements and the ground truth is the result of these measurements, but nothing is known about the hidden quantum system. We will show that in some case quantum tomography is a hard problem to learn. We consider a problem related to optical fiber communication where information is encoded in photon polarizations. We will show that the learning regret cannot decay faster than  $1/\sqrt{T}$  where *T* is the size of the training dataset and that incremental gradient descent may converge worse.

Keywords: machine learning; artificial intelligence; photon polarization; quantum tomography

## 1. Introduction: Supervised Learning in General

With the invention of deep neural learning, the general public thinks there is a glimpse of a universal machine learning technology capable of solving arbitrary problems without any specific preparation on training data and learning strategy. Everything "is" be solvable as long as there are enough layers, enough processing power and enough training data. We arrived at the point that many people (among them the late Stephen Hawking) start thinking that machines may supersede human intelligence thanks to the greater performance of silicon neurons over biological neurons, and may be capable of cracking the last enigmas around the physical nature of the universe.

However, we should not forget that actual Artificial Intelligence (AI) has many limitations. However, due to the youth of technology, many of the present limits might be of teething nature. To learn a language the present algorithms need to be trained over millions of texts which is equivalent to a training period of 80 years if it were done at the learning pace of a child! Presently, deep neural training is very demanding in processing and it is the third major source of energy consumption among information technologies after Bitcoin and data centers. Deep learning is not yet such a good self-organizing learning process as some researchers would have thought [1]. There is also the obstacle of data sparsity to learning (the machine only recognizes the data on which it has been trained over and over as if a reader could only understand the texts on which (s)he has been trained).

To make it short the main limitations of machine learning technologies are: (i) the data sparsity; (ii) the absence of a computable solution to learn (e.g., the program halting problem); (iii) the presence of hard-to-learn algorithms in the solution. My present paper will address the third limitation.

A supervised learning problem can be viewed as a set of training data and ground truths. The machine acts as an automaton whose aim is to predict the ground truth from data. The *loss* measures the difference between the prediction and the ground truth and can be established under an arbitrary metric. The general objective of supervised machine learning is to minimize the average loss, but since the ground truth might contain some



Citation: Jacquet, P. Is Quantum Tomography a Difficult Problem for Machine Learning? *Phys. Sci. Forum* 2022, *5*, 47. https://doi.org/10.3390/ psf2022005047

Academic Editors: Frédéric Barbaresco, Ali Mohammad-Djafari, Frank Nielsen and Martino Trassinelli

Published: 7 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). inherent stochastic variations (e.g., when predicting the result of a quantum measurement) it may be impossible to make the loss as small as we would like. Given an automaton architecture, there exists a setting that gives the optimal average loss. However, the optimal setting might be difficult to reach. However, there is still the question of the size of the training set needed to converge to the optimal settings.

All problems are not equal in front of learnability [2]. Some seem to be a perfect match with AI, some others are more difficult to adapt. In [3], the author shows that the random parity functions are just unlearnable. In fact, in a broader perspective, the "learnability" may not be a learnable problem [4].

The first contribution of this paper is a new definition of learning regret with respect to a given single problem submitted to a given learning strategy. Most regret expressions are infimum of regret over a large class (if not universal) of problems [5] and therefore lose the specificity of individual problems.

The second contribution is the application of this new regret definition to a quantum tomography problem. The specificity of the problem is that the hidden source probability distribution is indeed contained in the learning distribution class. The surprising result is that the regret is at least in the square root of the number of runs, hinting at a poor convergence rate of the learned distribution toward the hidden distribution. We conclude with numerical experiments with gradient descents.

### 2. Expressing the Convergence Regret

Let *T* be an integer and let  $\mathbf{x}^T = (\mathbf{x}_1, \dots, \mathbf{x}_T)$  be a sequence of features which are vectors of a certain dimension which define the problem (the notation with *T* is not for "transpose", which should be noted <sup>*T*</sup> $\mathbf{x}$ , but for a sequence with *T* atoms). Each feature  $\mathbf{x}$  generates a discrete random label *y*. Let denote  $P_S(y|\mathbf{x})$  (*S* for "source") the probability to have label *y* given the feature  $\mathbf{x}$ . If  $y^T$  is the sequence of random labels given the sequence of feature  $\mathbf{x}^T$ :  $P_S(y^T|\mathbf{x}^T) = \prod_t P_S(y_t|\mathbf{x}_t)$ . The sequence of features and labels defines the problem for supervised learning.

The learning process will give as output an index  $L(y^T)$  which will be taken from a set of  $\mathcal{L}$ , such that each  $L \in \mathcal{L}$  define a distribution  $P_L(y^T | \mathbf{x}^T)$  (L for "learning") over the label sequence given the feature sequence. In absence of side information the learning process leads to  $L(y^T) = \arg \max_{L \in \mathcal{L}} \{P_L(y^T | \mathbf{x}^T)\}$ . Our aim is find how close  $P_{L(y^T)}(y^T | \mathbf{x}^T)$  is to  $P_S(y^T | \mathbf{x}^T)$  when  $y^T$  varies.

The distance between the two distributions can be expressed by the Kullback–Leibler divergence [6]

$$D(P_S || P_L) = \sum_{y^T} P_S(y^T | \mathbf{x}^T) \log \frac{P_S(y^T | \mathbf{x}^T)}{P_{L(y^T)}(y^T | \mathbf{x}^T)}$$
(1)

However, it should be stressed that the quantity  $P_{L(y^T)}(y^T | \mathbf{x}^T)$  does not necessarily define a probability distribution since  $L(y^T)$  may vary when  $y^T$  varies, making  $\sum_{y^T} P_{L(y^T)}(y^T | \mathbf{x}^T)$  equal to 1 unlikely. Thus,  $D(P_S || P_L)$  is not a distance, because it can be non-positive. One way to get through is to introduce  $P_L^*(y^T | \mathbf{x}^T) = \frac{P_{L(y^T)}(y^T | \mathbf{x}^T)}{S_L(\mathbf{x}^T)}$  with  $S(\mathbf{x}^T) = \sum_{y^T} P_{L(y^T)}(y^T | \mathbf{x}^T)$  which makes  $P_L^*()$  a probability distribution. Thus, we will use  $D(P_S || P_L^*)$  which satisfies:

$$D(P_S || P_L^*) = \sum_{y^T} P_S(y^T | \mathbf{x}^T) \log \frac{P_S(y^T | \mathbf{x}^T)}{P_L^*(y^T | \mathbf{x}^T)} = D(P_S || P_L) + \log S(\mathbf{x}^T),$$
(2)

and is now a well-defined semi distance which we will define as the learning regret  $R(x^T) = D(P_S || P_L^*)$  [5].

### 3. The Quantum Learning on Polarized Photons

We now include pure physical measurements in the learning process. There are several applications that involve physic, ref. [7] describes a process of deep learning over

the physical layer of a wireless network. The issue with quantum physical effects is the fact that they are not reproducible and not deterministic. We consider a problem related to optical fiber communication where information is encoded in photon polarizations. The photon polarization is given by a quantum wave function of dimension 2. In the binary case, the bit 0 is given by polarisation angles  $\theta_Q$  and the bit 1 is given by angle  $\theta_Q + \pi/2$ . The quantity  $\theta_Q$  is supposed to be unknown by the receiver and its estimate  $\theta_T$  is obtained after a training sequence via machine learning.

For this purpose, the sender sends a sequence of *T* equally polarized photons, along angle  $\theta_Q$ , the receiver measures these photons over a collection of *T* measurement angles  $x_1, x_2, \ldots, x_T$ , called the featured angles. They are pure scalar and are not vector (d = 1), therefore we will not depict them in bold font as in the previous section which is therefore of dimension 1. The labels, or ground truths,  $y_1, \ldots, y_T$  are the sequence of binary measurement obtained,  $y_t \in \{0, 1\}$ , there are  $2^T$  possible label sequences.

This problem is the most simplified version of tomography on quantum telecommunication since it relies on a single parameter. More realistic and more complicated situations will occur when noisy circular polarization is introduced within a more complex combination of polarizations within groups of photons. This will considerably increase the dimension of the feature vectors and certainly will make our results on the training process more critical. However, in the situation analyzed in our paper, we show that this simple system is difficult to learn.

If we assume that the experiment results are delivered in batches to the training process, that is the estimate  $\theta_t = \theta$  does not vary for 0 < t < T, the learning class of probability distribution is a function of  $\theta$  with  $P_L(y^T | x^T, \theta) = \prod_{y_t=0} \cos(\theta - x_t)^2 \prod_{y_t=1} \sin(\theta - x_t)^2$ . The source distribution is indeed  $P_S(y^T | x^T) = P_L(y^T | x^T, \theta_Q)$ , thus the source distribution belongs to the class  $\mathcal{L}$  of learning distribution. For a given pair of sequence  $(y^T, x^T)$ , let  $\theta^*$  be the value of  $\theta$  which maximizes  $P_L(y^T | x^T, \theta)$ . Since we will never touch the sequence  $x^T$  which are the foundation of the experiments, we will sometimes drop the parameter  $x^T$  and denote  $\ell_{y^T}(\theta) = -\log P_L(y^T | x^T, \theta)$ . The quantity  $\theta^*$  which maximizes  $P_L(y^T | x^T, \theta)$  will satisfy  $\ell'_{u^T}(\theta^*) = 0$ . We have

$$\begin{cases} \ell_{y^T}(\theta) &= -2\sum_t \log|\cos(\theta - x_t + y_t \pi/2)| \\ \ell'_{y^T}(\theta) &= 2\sum_t \tan(\theta - x_t + y_t \pi/2) \\ \ell''_{y^T}(\theta) &= 2\sum_t \frac{1}{\cos(\theta - x_t + y_t \pi/2)^2} \end{cases}$$

we notice that for all  $\theta \ell''_{y^T}$  is always strictly positive (but  $\ell''$  and  $\ell'$  are not continuous so  $\ell$  is not convex). We now turn to displaying and proving our main results (two theorems), whose proof would need the following two next lemmas.

Lemma 1. We have the expression

$$\ell_{y^{T}}(\theta^{*}) = \frac{1}{2\pi} \int_{0}^{2\pi} \ell_{y^{T}}(w) \ell''_{y^{T}}(w) dw \int_{\mathbb{R}} \exp(-i\ell'_{y^{T}}(w)z) dz.$$
(3)

**Proof.** Let  $g_{y^T}(\theta) = \ell'_{y^T}(\theta)$  which is homomorphic and is locally invertible (since  $\ell''_{y^T}(\theta)$  is never zero). Let  $a \in \mathbb{R}$  we denote  $l_{y^T}$  the function  $a \to \ell_{y^T}(g_{y^T}^{-1}(a))$ . We have  $\ell_{y^T}(\theta^*) = l_{y^T}(0)$ . For  $z \in \mathbb{R}$ , let  $\tilde{l}_{y^T}(z)$  be the Fourier transform of function  $l_{y^T}(a)$ . Formally we have

$$\widetilde{l}_{y^T}(z) = \int_{\mathbb{R}} l_{y^T}(a) e^{-iaz} da$$
(4)

$$= \int_{0}^{2\pi} \ell_{y^{T}}(w) \ell_{y^{T}}''(w) e^{-i\ell'_{y^{T}}(w)z} dw$$
(5)

and inversely

$$l_{y^{T}}(a) = \frac{1}{2\pi} \int_{\mathbb{R}} \tilde{l}_{y^{T}}(\mathbf{z}) e^{iaz} dz$$
(6)

Thus

$$\ell_{y^{T}}(\theta^{*}) = \frac{1}{2\pi} \int_{\mathbb{R}} \widetilde{l}_{y^{T}}(z) dz$$
(7)

$$= \frac{1}{2\pi} \int_0^{2\pi} \ell_{y^T}(w) \ell_{y^T}''(w) dw$$
 (8)

$$\times \int_{\mathbb{R}} e^{-i\ell'_{y^T}(w)z} dz.$$
<sup>(9)</sup>

In fact, the function  $\ell_{y^T}(\theta)$  may have several extrema as we will see in the next section, thus  $\ell'_{y^T}(\theta)$  may have several roots, thus  $g_{y^T}^{-1}(a)$  is polymorphic. In order to avoid the secondary roots which contribute to the non-optimal extrema, we will concentrate on the main root in the vicinity of  $\theta_Q$ .

Let  $p^T = (p_1, \dots, p_T)$  and  $q^T = (q_1, \dots, q_T)$  be two sequence of real numbers. We denote  $p(y^T) = \prod_t p_t^{1-y_t} q_t^{y_t}$ .

**Lemma 2.** For any  $1 \le t_0 \le T$  we have the identity

$$\sum_{y^T} y_{t_0} p(y^T) = q_{t_0} \prod_{t \neq t_0} (p_t + q_t).$$
(10)

For  $t_1 \neq t_2$ , we have

$$\sum_{y^T} y_{t_1} y_{t_2} p(y^T) = q_{t_1} q_{t_2} \prod_{t \neq t_1, t_2} (p_t + q_t).$$
(11)

**Proof.** This is just the consequence of the finite sums via algebraic manipulations.  $\Box$ 

**Theorem 1.** Under mild conditions, we have the estimate

$$\sum_{y^T} P(y^T | x^T) \log \frac{P_S(y^T | x^T)}{P_{L(y^T)}(y^T | x^T)} = O(\sqrt{T})$$
(12)

**Proof.** Let  $C(x^T) = \sum_{y^T} P_S(y^T | x^T) \ell_{y^T}(\theta^*)$ . Applying both lemma with  $p_t = \cos(\theta_Q - x_t)^2 e^{-2i \tan(\theta - x_t) z}$  and  $q_t = \sin(\theta_Q - x_t)^2 e^{-2i \tan(\theta - x_t + \pi/2) z}$ , thus  $p(y^T) = P_S(y^T | x^T) e^{-i\ell_y T(\theta)}$  we get

$$C(x^{T}) = \sum_{y^{T}} \frac{1}{2\pi} \int_{0}^{2\pi} \ell_{y^{T}}(\theta) \ell''_{y^{T}}(\theta) d\theta \int_{\mathbb{R}} \exp(-i\ell'_{y^{T}}(w)z) dz$$
$$= \frac{1}{2\pi} \int_{0}^{2\pi} d\theta \int_{\mathbb{R}} (\overline{\ell}(\theta, z)\overline{\ell}''(\theta, z) + \overline{\Delta}(\theta, z)) \prod_{t} (p_{t} + q_{t}) dz$$

with

$$\overline{\ell}(\theta, z) = -2\sum_{t} \frac{p_t}{p_t + q_t} \log \cos(\theta - x_t) + \frac{q_t}{p_t + q_t} \log \sin(\theta - x_t)$$

$$\overline{\ell}''(\theta, z) = 2\sum_{t} \frac{p_t}{p_t + q_t} \frac{1}{\cos(\theta - x_t)^2} + \frac{q_t}{p_t + q_t} \frac{1}{\sin(\theta - x_t)^2}$$

$$\overline{\Delta}(\theta, z) = -2\sum_{t} \frac{p_t q_t}{(p_t + q_t)^2} \left( \frac{\log \cos(\theta - x_t)}{\cos(\theta - x_t)^2} + \frac{\log \sin(\theta - x_t)}{\sin(\theta - x_t)^2} \right)$$

We notice that  $\prod_t (p_t + q_t) = \exp(2im(\theta)z + v(\theta)z^2 + O(z^3T))$  with

$$m(\theta) = \sum_{t} \tan(\theta - x_{t}) \cos(\theta_{Q} - x_{t})^{2} + \tan(\theta - x_{t} + \pi/2) \sin(\theta - x_{t})^{2}$$
  

$$v(\theta) = \sum_{t} \tan(\theta - x_{t})^{2} \cos(\theta_{Q} - x_{t})^{2} + \tan(\theta - x_{t} + \pi/2)^{2} \sin(\theta_{Q} - x_{t})^{2}$$
  

$$-\sum_{t} \left( \tan(\theta - x_{t}) \cos(\theta_{Q} - x_{t})^{2} + \tan(\theta - x_{t} + \pi/2) \sin(\theta_{Q} - x_{t})^{2} \right)^{2}$$

We notice that  $m(\theta) \sim 2(\theta - \theta_Q)T$  and  $v(\theta) = T + O(\theta - \theta_Q)$  when  $\theta \to \theta_Q$ . The expression is obtained via saddle point method approximation, under the mild conditions being that it can be applied as in the maximum likelihood problem [8] (the error term would be the smallest possible)

$$\int_{\mathbb{R}} (\bar{\ell}(\theta, z)\bar{\ell}''(\theta, z) + \bar{\Delta}(\theta, z)) \prod_{t} (p_{t} + q_{t}) dz = \int_{\mathbb{R}} (\bar{\ell}(\theta, z)\bar{\ell}''(\theta, z) + \bar{\Delta}(\theta, z)) \\ \exp\left(-im(\theta)z - v(\theta)z^{2}/2 + O(T|z|^{3})\right) dz$$
(13)  
$$= (\bar{\ell}(\theta)\bar{\ell}''(\theta) + \bar{\Delta}(\theta)) \frac{\sqrt{\pi}}{\sqrt{v(\theta)}} \exp\left(-\frac{m(\theta)^{2}}{v(\theta)}\right) \\ (1 + O(1/\sqrt{T}))$$
(14)

with  $\overline{\ell}(\theta) = \overline{\ell}(\theta, 0)$ ,  $\overline{\ell}''(\theta) = \overline{\ell}''(\theta, 0)$  and  $\overline{\Delta}(\theta) = \overline{\Delta}(\theta, 0)$  Since  $\frac{m(\theta)^2}{v(\theta)} = 4(\theta - \theta_Q)^2 T + O(|\theta - \theta_Q|^3 T)$ , the factor  $\prod_t (p_t + q_t)$  behaves like a gaussian function centered on  $\theta_Q$  with standard deviation of order  $1/\sqrt{T}$ . Thus, via saddle point approximation again, it comes:

$$\begin{split} C(x^{T}) &= \frac{1}{2\sqrt{\pi}} \int_{0}^{2\pi} (\bar{\ell}(\theta)\bar{\ell}''(\theta) + \bar{\Delta}(\theta)) \frac{\sqrt{\pi}}{\sqrt{v(\theta)}} \exp(-\frac{m(\theta)}{v(\theta)}) (1 + O(1/\sqrt{T})) \\ &= \frac{1}{2\sqrt{\pi}} \int_{0}^{2\pi} \frac{\bar{\ell}(\theta)\bar{\ell}''(\theta) + \bar{\Delta}(\theta)}{\sqrt{v(\theta)}} \exp\left(-4(\theta - \theta_{Q})^{2}T + O(|\theta - \theta_{Q}|^{3}T)\right) (1 + O(1/\sqrt{T})) \\ &= \frac{\bar{\ell}(\theta_{Q})\bar{\ell}''(\theta_{Q}) + \bar{\Delta}(\theta_{Q})}{2\sqrt{v(0)}} (1 + O(1/\sqrt{T})) \\ &= h(\theta_{Q})(1 + O(1/\sqrt{T})) \\ &\text{ with } h(\theta_{Q}) = (\bar{\ell}(\theta_{Q})\bar{\ell}''(\theta_{Q}) - \bar{\Delta}(\theta_{Q}))/2T \text{ with } h(\theta) - \sum_{t} \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t}) + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x_{t})^{2} + \frac{1}{2} \log \cos(\theta - x_{t})^{2} \log \cos(\theta - x$$

with  $h(\theta_Q) = (\ell(\theta_Q)\ell^{-1}(\theta_Q) - \Delta(\theta_Q))/21$  with  $h(\theta) - \sum_t \cos(\theta - x_t)^2 \log \cos(\theta - x_t)^2 + \sin(\theta - x_t)^2 \log \sin(\theta - x_t)^2$  is clearly O(T). Furthermore,  $h(\theta_Q) = -\sum_{y^T} P_S(y^T | x^T) \log P_S(y^T | x^T)$ , thus we have

$$\sum_{y^T} P(y^T | x^T) \log \frac{P_S(y^T | x^T)}{P_{L(y^T)}(y^T | x^T)} = O\left(\frac{h(\theta_Q)}{\sqrt{T}}\right) = O(\sqrt{T}).$$

5 of 9

**Theorem 2.** We have

$$\log S(x^{T}) = \log \left( \sum_{y^{T}} P_{L(y^{T})}(y^{T} | x^{T}) \right) = \frac{1}{2} \log T + O(1).$$
(15)

**Remark 1.** This order of magnitude is much smaller than the main order of magnitude provided in Theorem 1, confirming that the overall regret is indeed  $\sqrt{T}$ . The regret per measurement is  $O(1/\sqrt{T})$  therefore the individual regrets nevertheless tend to zero when  $T \to \infty$ .

**Proof.** It is formally a Shtarkov sum [5,9]. Using Lemmas 1 and 2 gives

$$S(x^{T}) = \sum_{y^{T}} P_{L(y^{T})}(y^{T}|x^{T}) = \sum_{y^{T}} \frac{1}{2\pi} \int_{0}^{2\pi} P(y^{T}|x^{T}, w) \ell_{y^{T}}''(w) dw \int_{\mathbb{R}} \exp(-i\ell'_{y^{T}}(w)z) dz.$$
(16)  
$$= \frac{1}{2\pi} \int_{0}^{2\pi} d\theta \int_{0}^{2\pi} \tilde{\ell}''(\theta, z) \Pi(w, z, z) dz.$$
(17)

$$= \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_{\mathbb{R}} \tilde{\ell}''(\theta, z) \prod_t (p_t + q_t) dz$$
(17)

with  $p_t = \cos(\theta - x_t)^2 e^{-2i\tan(\theta - x_t)z}$  and  $q_t = \sin(\theta - x_t)^2 e^{-2i\tan(\theta - x_t + \pi/2)z}$ , thus  $p(y^T) = P(y^T | x^T, \theta) e^{-i\ell_y T(\theta)}$ ;  $\tilde{\ell}''(\theta, z)$  has same expression as  $\bar{\ell}''(\theta, z)$  but with the new expression of  $p_t$  and  $q_t$ :

$$\widetilde{\ell}''(\theta, z) = 2\sum_t \frac{p_t}{p_t + q_t} \frac{1}{\cos(\theta - x_t)^2} + \frac{q_t}{p_t + q_t} \frac{1}{\sin(\theta - x_t)^2}$$

Developing further:

$$S(x^{T}) = \frac{1}{2\pi} \int_{0}^{2\pi} d\theta \int_{\mathbb{R}} \widetilde{\ell}(\theta, z) \exp\left(-2Tz^{2} + O(T|z^{3}|)\right), \tag{18}$$

via the saddle point estimate (which consists to do a change of variable  $z \rightarrow \frac{1}{\sqrt{T}}z'$  under the same conditions of Theorem 1 we get

$$S(x^T) = \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_{\mathbb{R}} \widetilde{\ell}(\theta, 0) \frac{\sqrt{\pi}}{\sqrt{2T}} (1 + O(1/\sqrt{T})).$$
(19)

We terminate with the evaluation  $\tilde{\ell}''(\theta, 0) = 4T$ , thus  $S(x^T) = \frac{\sqrt{T}}{\sqrt{\pi/2}}(1 + O(1/\sqrt{T}))$ .

### 4. Incremental Learning and Gradient Descent

We investigate gradient descent methods to reach the value  $\theta^*$ . There are many gradient strategies. The classic strategy, which we call, the slow gradient descent, where we define the loss by  $\log(y_t, \theta_t | x_t) = (y_t - \sin(\theta_t - x_t)^2)^2$ , since the average value of  $y_t$  is  $\sin(\theta_Q)^2$ , thus the average loss is  $(\sin(\theta_Q - x_t)^2 - \sin(\theta_t - x_t)^2)^2 + \frac{\sin(2\theta_Q - 2x_t)^2}{4}$  (minimized at  $\theta_t = \theta_Q$ ) and the gradient  $\theta_t$  updates is

$$\theta_{t+1} = \theta_t - r \frac{\partial}{\theta_t} loss(y_t, \theta_t | x_t).$$
(20)

In Figure 1 we display our simulations as a sequence  $\theta_t$  starting with a random initial  $\theta_1$ . We assume that for all *t* the transmitted bit is always 0 i.e., the polarization angle is always  $\theta_Q$ . The learning rate is r = 0.0002. We simulate nine parallel gradient descents randomly initialized sharing the same random feature sequence  $x^T$ , with T = 3,000,000. On Figure 1 we plot the parallel evolutions of quantity  $\theta_t$ . The initial points are green diamonds and the final points are red diamonds. Although we start with nine different positions, the trajectories converge toward  $\theta_Q \pm \pi$ . However, the convergence is slow, confirming the  $1/\sqrt{T}$  and worse rate. In fact, some initial positions converge even more slowly, and even after 3,000,000 trials, are still very far. The reason is that the target function log  $P(y^T | x^T, \theta)$ 

has several local maxima as it is shown in Figure 2 where the  $x_t$  belongs to the set of values  $2\pi k/10$  for k = 1, ..., 10. It is very unlikely that a communication operator would tolerate so many runs (3,000,000) in order to have a proper convergence. However, it would be possible to run the gradient descents in parallel and act like with particle systems in order to select the fastest in convergence.







**Figure 2.** Target function  $\sum_t \cos(\theta_Q - x_t)^2 \log \cos(\theta - x_t)^2 + \sin(\theta_Q - x_t)^2 \log \sin(\theta - x_t)^2$  as function of  $\theta$ .

A supposedly faster gradient descent would be defined by the inverse derivative

$$\theta_{t+1} = \theta_t + r \frac{y_t - \sin(\theta_t - x_t)^2}{\frac{\partial}{\partial \theta_t} \sin(\theta_t - x_t)^2}$$
(21)

We notice that in stationary situation (where we suppose that  $\theta_t$  very little varies) we have  $E(\theta_{t+1}) = \theta_t + r \frac{\sin(\theta_Q - x_t)^2 - \sin(\theta_t - x_t)^2}{\frac{\partial}{\partial \theta_t} \sin(\theta_t - x_t)^2}$  which is equal to  $\theta_t$  when  $\theta_t = \theta_Q$ . In Figure 3,

we display our simulations as a sequence  $\theta_t$  starting with a random initial  $\theta_1$ . The learning rate is r = 0.0002. We simulate nine parallel fast gradient descents randomly initialized sharing the same random feature sequence  $x^T$ , with T = 3,000,000. The gradient descent converges fast but does not converge on the good value  $\theta_Q \pm \pi$ . Again it is due to the fact that the target function log  $P(y^T | x^T, \theta)$  has several local maxima which act like a trap for the gradient descent.



**Figure 3.** Angle estimate  $\theta_t$  versus time of nine fast gradient descents randomly initialized. Green diamonds are starting points, red diamonds are stopping points.

#### 5. Conclusions

We have presented a simple quantum tomography problem, the photon unknown polarization problem and have analyzed its learnability via AI over *T* runs. We have shown that the learning regret cannot decay faster than  $1/\sqrt{T}$  (i.e., a cumulative regret of  $\sqrt{T}$ ). Furthermore, the classic gradient descent is hampered by local extrema which may significantly impact the theoretical convergence rate.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

### References

- Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015.
- Bouillard, A.; Jacquet, P. Quasi Black Hole Effect of Gradient Descent in Large Dimension: Consequence on Neural Network Learning. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
- 3. Abbe, E.; Sandon, C. Provable limitations of deep learning. *arXiv* 2018, arXiv:1812.06369.
- Ben-David, S.; Hrubeš, P.; Moran, S.; Shpilka, A.; Yehudayoff, A. Learnability can be undecidable. *Nat. Mach. Intell.* 2019, 1, 44–48. [CrossRef]
- Jacquet, Philippe, Gil Shamir, and Wojciech Szpankowski. Precise Minimax Regret for Logistic Regression with Categorical Feature Values. In Algorithmic Learning Theory; PMLR: New York City, NY, USA, 2021.

- Van Erven, T.; Harremos, P. Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* 2014, 60, 3797–3820. [CrossRef]
- O'shea, T.; Hoydis, J. An introduction to deep learning for the physical layer. *IEEE Trans. Cogn. Commun. Netw.* 2017, 3, 563–575. [CrossRef]
- 8. Newey, W.K.; McFadden, D. Chapter 36: Large sample estimation and hypothesis testing. In *Handbook of Econometrics*; Elsevier: Amsterdam, The Netherlands, 1994.
- 9. Shtarkov, Y.M. Universal sequential coding of single messages. Probl. Inf. Transm. 1987, 23, 3–17.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.