



HAL
open science

Is Quantum Tomography a difficult problem for Machine Learning?

Philippe Jacquet

► **To cite this version:**

Philippe Jacquet. Is Quantum Tomography a difficult problem for Machine Learning?. MAXENT 2022 - 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Jul 2022, Paris, France. hal-03942607

HAL Id: hal-03942607

<https://hal.science/hal-03942607>

Submitted on 17 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

Is Quantum Tomography a difficult problem for Machine Learning?

Philippe Jacquet

¹ Inria Saclay Ile-de-France, France; philippe.jacquet@inria.fr

† Presented to International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT'22), IHP, Paris, July 18-22, 2022.

Version December 8, 2022 submitted to Entropy

Abstract: One of the key issue in machine learning is the characterization of the learnability of a problem. The regret is way to quantify learnability. The quantum tomography is a special case of machine learning where the training set is a set of quantum measurements and the ground truth the results of these measurements, but nothing is known about the hidden quantum system. We will show that in some case the quantum tomography is a hard problem to learn. We consider a problem related to optical fiber communication where information are 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 descents 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 to the point that many people (among them late Stephen Hawking) start thinking that machines may supersede human intelligence thanks to the greater performance of silicon neurons over biological neurons, and maybe capable of cracking the last enigmas around the physical nature of the universe.

But we should not forget that actual Artificial Intelligence (AI) has many limitations. But due to the youth of the 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 of 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 energy consumption among information technologies after Bitcoin and data centers. Deep learning is not yet as such a good self organizing learning process as some researchers would have thought [5]. 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 the machine learning technologies are: (i) the data sparsity; (ii) the absence of 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 a 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 inherent stochastic variations (e.g., when predicting the result of

35 a quantum measurement) it may be impossible to make the loss as small as we would like. Given an
 36 automaton architecture, there exists a setting which gives the optimal average loss. But the optimal
 37 setting might be difficult to reach. However, there is still the question of the size of the training set
 38 needed to converge to the optimal settings.

39 All problems are not equal in front of learnability [9]. Some seem to be a perfect match with AI,
 40 some other are more difficult to adapt. In [2] the author shows that the random parity functions are
 41 just unlearnable. In fact in a broader perspective, the "learnability" may not be a learnable problem [1].

42 The first contribution of this paper is a new definition of learning regret with respect to a given
 43 single problem submitted to a given learning strategy. Most regret expression are infimum of regret
 44 over large class (if not universal) of problems [6] and therefore lose the specificity of individual
 45 problems.

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

51 2. Expressing the convergence regret

52 Let T be an integer and let $\mathbf{x}^T = (x_1, \dots, x_T)$ be a sequence of features which are vectors of a
 53 certain dimension which define the problem (the notation with T is not for "transpose", which should
 54 be noted ${}^T\mathbf{x}$, but for a sequence with T atoms). Each feature \mathbf{x} generates a discrete random label y . Let
 55 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
 56 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
 57 and labels define the problem for supervised learning.

58 The learning process will give as output an index $L(y^T)$ which will be taken from a set of \mathcal{L} ,
 59 such that each $L \in \mathcal{L}$ define a distribution $P_L(y^T|\mathbf{x}^T)$ (L for "learning") over the label sequence
 60 given the feature sequence. In absence of side information the learning process leads to $L(y^T) =$
 61 $\arg \max_{L \in \mathcal{L}} \{P_L(y^T|\mathbf{x}^T)\}$. 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.

62 The distance between the two distributions can be expressed by the Kullback-Liebler
 63 divergence [3]

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

64 However it should be stressed that the quantity $P_{L(y^T)}(y^T|\mathbf{x}^T)$ does not necessarily define a probability
 65 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
 66 $D(P_S \| P_L)$ is not a distance, because it can be non positive. One way to get through is to introduce
 67 $P_L^*(y^T|\mathbf{x}^T) = \frac{P_{L(y^T)}(y^T|\mathbf{x}^T)}{S(\mathbf{x}^T)}$ with $S(\mathbf{x}^T) = \sum_{y^T} P_{L(y^T)}(y^T|\mathbf{x}^T)$ which makes $P_L^*(\cdot)$ a probability distribution.
 68 Thus we will use $D(P_S \| P_L^*)$ which satisfies:

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

69 and is now a well defined semi distance which we will define as the learning regret $R(\mathbf{x}^T) =$
 70 $D(P_S \| P_L^*)$ [6].

71 3. The quantum learning on polarized photons

72 We now include pure physical measurements in the learning process. There are several application
 73 which involves physic, [8] describe a processus of deep learning over the physical layer of a wireless
 74 network. The issue with quantum physical effect is the fact that they are not reproducible and not

75 deterministic. We consider a problem related to optical fiber communication where information are
 76 encoded in photon polarizations. The photon polarization is given by a quantum wave function of
 77 dimension 2. In the binary case the bit 0 is given by polarisation angles θ_Q and the bit 1 is given by
 78 angle $\theta_Q + \pi/2$. The quantity θ_Q is supposed to be unknown by the receiver and its estimate θ_T
 79 is obtained after a training sequence via machine learning.

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

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

92 If we assume that the experiment results are delivered in batch to the training process, that is the
 93 estimate $\theta_t = \theta$ does not vary for $0 < t < T$, the learning class of probability distribution is a function
 94 of θ 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
 95 $P_S(y^T|x^T) = P_L(y^T|x^T, \theta_Q)$, thus the source distribution belongs to the class \mathcal{L} of learning distribution.
 96 For a given pair of sequence (y^T, x^T) , let θ^* be the value of θ which maximizes $P_L(y^T|x^T, \theta)$. Since we
 97 will never touch the sequence x^T which are the foundation of the experiments, we will sometimes
 98 drop the parameter x^T and denote $\ell_{y^T}(\theta) = -\log P_L(y^T|x^T, \theta)$. The quantity θ^* which maximizes
 99 $P_L(y^T|x^T, \theta)$ will satisfy $\ell'_{y^T}(\theta^*) = 0$. We have

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

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

Lemma 1. *We have the expression*

$$\ell_{y^T}(\theta^*) = \frac{1}{2\pi} \int_0^{2\pi} \ell_{y^T}(w) \ell''_{y^T}(w) dw \int_{\mathbb{R}} \exp(-i\ell'_{y^T}(w)z) dz. \quad (3)$$

103 **Proof.** Let $g_{y^T}(\theta) = \ell'_{y^T}(\theta)$ which is homomorphic and is locally invertible (since $\ell''_{y^T}(\theta)$ is never
 104 zero). Let $a \in \mathbb{R}$ we denote l_{y^T} the function $a \rightarrow \ell_{y^T}(g_{y^T}^{-1}(a))$. We have $\ell_{y^T}(\theta^*) = l_{y^T}(0)$. For $z \in \mathbb{R}$, let
 105 $\tilde{l}_{y^T}(z)$ be the Fourier transform of function $l_{y^T}(a)$. Formally we have

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

$$= \int_0^{2\pi} \ell_{y^T}(w) \ell_{y^T}''(w) e^{-i\ell'_{y^T}(w)z} dw \quad (5)$$

106 and inversely

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

107 Thus

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

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

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

108 \square

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

112 Let $p^T = (p_1, \dots, p_T)$ and $q^T = (q_1, \dots, q_T)$ be two sequence of real numbers. We denote
 113 $p(y^T) = \prod_t p_t^{1-y_t} q_t^{y_t}$.

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

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

115 For $t_1 \neq t_2$, we have

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

116 **Proof.** This is just the consequence of the finite sums via algebraic manipulations. \square

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

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

118 **Proof.** Let $C(x^T) = \sum_{y^T} P_S(y^T | x^T) \ell_{y^T}(\theta^*)$. Applying both lemma with $p_t = \cos(\theta_Q -$
 119 $x_t)^2 e^{-2i \tan(\theta - x_t)z}$ 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

$$\begin{aligned} C(x^T) &= \sum_{y^T} \frac{1}{2\pi} \int_0^{2\pi} \ell_{y^T}(\theta) \ell''_{y^T}(\theta) d\theta \int_{\mathbb{R}} \exp(-i\ell'_{y^T}(w)z) dz \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_{\mathbb{R}} (\bar{\ell}(\theta, z) \bar{\ell}''(\theta, z) + \bar{\Delta}(\theta, z)) \prod_t (p_t + q_t) dz \end{aligned}$$

120 with

$$\begin{aligned} \bar{\ell}(\theta, z) &= -2 \sum_t \frac{p_t}{p_t + q_t} \log \cos(\theta - x_t) + \frac{q_t}{p_t + q_t} \log \sin(\theta - x_t) \\ \bar{\ell}''(\theta, z) &= 2 \sum_t \frac{p_t}{p_t + q_t} \frac{1}{\cos(\theta - x_t)^2} + \frac{q_t}{p_t + q_t} \frac{1}{\sin(\theta - x_t)^2} \\ \bar{\Delta}(\theta, z) &= -2 \sum_t \frac{p_t q_t}{(p_t + q_t)^2} \left(\frac{\log \cos(\theta - x_t)}{\cos(\theta - x_t)^2} + \frac{\log \sin(\theta - x_t)}{\sin(\theta - x_t)^2} \right) \end{aligned}$$

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

$$\begin{aligned} m(\theta) &= \sum_t \tan(\theta - x_t) \cos(\theta_Q - x_t)^2 + \tan(\theta - x_t + \pi/2) \sin(\theta - x_t)^2 \\ v(\theta) &= \sum_t \tan(\theta - x_t)^2 \cos(\theta_Q - x_t)^2 + \tan(\theta - x_t + \pi/2)^2 \sin(\theta_Q - x_t)^2 \\ &\quad - \sum_t \left(\tan(\theta - x_t) \cos(\theta_Q - x_t)^2 + \tan(\theta - x_t + \pi/2) \sin(\theta_Q - x_t)^2 \right)^2 \end{aligned}$$

122 We notice that $m(\theta) \sim 2(\theta - \theta_Q)T$ and $v(\theta) = T + O(\theta - \theta_Q)$ when $\theta \rightarrow \theta_Q$. The expression is obtained
123 via saddle point method approximation, under the mild conditions being that it can be applied as in
124 the maximum likelihood problem [7] (the error term would be the smallest possible)

$$\begin{aligned} \int_{\mathbb{R}} (\bar{\ell}(\theta, z) \bar{\ell}''(\theta, z) + \bar{\Delta}(\theta, z)) \prod_t (p_t + q_t) dz &= \int_{\mathbb{R}} (\bar{\ell}(\theta, z) \bar{\ell}''(\theta, z) + \bar{\Delta}(\theta, z)) \\ &\quad \exp\left(-im(\theta)z - v(\theta)z^2/2 + O(T|z|^3)\right) dz \quad (13) \\ &= (\bar{\ell}(\theta) \bar{\ell}''(\theta) + \bar{\Delta}(\theta)) \frac{\sqrt{\pi}}{\sqrt{v(\theta)}} \exp\left(-\frac{m(\theta)^2}{v(\theta)}\right) \\ &\quad (1 + O(1/\sqrt{T})) \quad (14) \end{aligned}$$

125 with $\bar{\ell}(\theta) = \bar{\ell}(\theta, 0)$, $\bar{\ell}''(\theta) = \bar{\ell}''(\theta, 0)$ and $\bar{\Delta}(\theta) = \bar{\Delta}(\theta, 0)$ Since $\frac{m(\theta)^2}{v(\theta)} = 4(\theta - \theta_Q)^2T + O(|\theta - \theta_Q|^3T)$,
126 the factor $\prod_t(p_t + q_t)$ behaves like a gaussian function centered on θ_Q with standard deviation of order
127 $1/\sqrt{T}$. Thus via saddle point approximation again, it comes:

$$\begin{aligned} C(x^T) &= \frac{1}{2\sqrt{\pi}} \int_0^{2\pi} (\bar{\ell}(\theta) \bar{\ell}''(\theta) + \bar{\Delta}(\theta)) \frac{\sqrt{\pi}}{\sqrt{v(\theta)}} \exp\left(-\frac{m(\theta)}{v(\theta)}\right) (1 + O(1/\sqrt{T})) \\ &= \frac{1}{2\sqrt{\pi}} \int_0^{2\pi} \frac{\bar{\ell}(\theta) \bar{\ell}''(\theta) + \bar{\Delta}(\theta)}{\sqrt{v(\theta)}} \exp\left(-4(\theta - \theta_Q)^2T + O(|\theta - \theta_Q|^3T)\right) (1 + O(1/\sqrt{T})) \\ &= \frac{\bar{\ell}(\theta_Q) \bar{\ell}''(\theta_Q) + \bar{\Delta}(\theta_Q)}{2\sqrt{v(\theta_Q)}} (1 + O(1/\sqrt{T})) \\ &= h(\theta_Q) (1 + O(1/\sqrt{T})) \end{aligned}$$

128 with $h(\theta_Q) = (\bar{\ell}(\theta_Q) \bar{\ell}''(\theta_Q) - \bar{\Delta}(\theta_Q))/2T$ with $h(\theta) = \sum_t \cos(\theta - x_t)^2 \log \cos(\theta - x_t)^2 + \sin(\theta -$
129 $x_t)^2 \log \sin(\theta - x_t)^2$ 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

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

130 \square

Theorem 2. *We have*

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

131 **Remark:** this order of magnitude is much smaller than the previous order of magnitude, confirming
132 that the overall regret is indeed \sqrt{T} . The regret per measurement is $O(1/\sqrt{T})$ therefore the individual
133 regrets nevertheless tend to zero when $T \rightarrow \infty$.

134 **Proof.** It is formally a Shtarkov sum [4], [6]. Using lemma 1 and lemma 2 gives

$$S(x^T) = \sum_{y^T} P_{L(y^T)}(y^T|x^T) = \sum_{y^T} \frac{1}{2\pi} \int_0^{2\pi} P(y^T|x^T, w) \ell_{y^T}''(w) dw \int_{\mathbb{R}} \exp(-i \ell'_{y^T}(w)z) dz. \quad (16)$$

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

135 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 $p(y^T) =$
 136 $P(y^T|x^T, \theta) e^{-i \ell_{y^T}(\theta)}$; $\tilde{\ell}''(\theta, z)$ has same expression as $\bar{\ell}''(\theta, z)$ but with the new expression of p_t and q_t :

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

137 Developing further:

$$S(x^T) = \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_{\mathbb{R}} \tilde{\ell}(\theta, z) \exp(-2Tz^2 + O(T|z^3|)) dz, \quad (18)$$

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

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

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

141 4. Incremental learning and gradient descent

142 We investigate gradient descent methods to reach the value θ^* . There are many gradient
 143 strategies. The classic strategy, which we call, the slow gradient descent, where we define the loss by
 144 $\text{loss}(y_t, \theta_t|x_t) = (y_t - \sin(\theta_t - x_t))^2$, since the average value of y_t is $\sin(\theta_Q)^2$, thus the average loss is
 145 $(\sin(\theta_Q - x_t)^2 - \sin(\theta_t - x_t)^2)^2 + \frac{\sin(2\theta_Q - 2x_t)^2}{4}$ (minimized at $\theta_t = \theta_Q$) and the gradient θ_t updates is

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

146 In figure 2 we display our simulations as a sequence θ_t starting with a random initial θ_1 . We
 147 assume that for all t the transmitted bit is always 0 *i.e.* the polarization angle is always θ_Q . The learning
 148 rate is $r = 0.0002$. We simulate nine parallel gradient descents randomly initialized sharing the
 149 same random feature sequence x^T , with $T = 3,000,000$. On figure 2 we plot the parallel evolutions of
 150 quantity θ_t . The initial points are green diamond and the final points are the red diamond. Although we
 151 start with nine different positions, the trajectories converge toward $\theta_Q \pm \pi$. However the convergence
 152 is slow, confirming the $1/\sqrt{T}$ and worse rate. In fact some initial positions converge even more slowly
 153 and even after 3,000,000 trials is still very far. The reason is that the target function $\log P(y^T|x^T, \theta)$
 154 has several local maxima as it is shown in figure 1 where the x_t belongs to the set of values $2\pi k/10$
 155 for $k = 1, \dots, 10$. It is very unlikely that a communication operator would tolerate so many runs
 156 (3,000,000) in order to have a proper convergence. However it would be possible to run the gradient
 157 descents in parallel and act like with particle systems in order to select the fastest in convergence.

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

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

159 We notice that in stationary situation (where we suppose that θ_t very little varies) we have
 160 $E(\theta_{t+1}) = \theta_t + r \frac{\sin(\theta_Q - x_t)^2 - \sin(\theta_t - x_t)^2}{\frac{\partial}{\partial \theta_t} \sin(\theta_t - x_t)^2}$ which is equal to θ_t when $\theta_t = \theta_Q$. In figure 3 we display our
 161 simulations as a sequence θ_t starting with a random initial θ_1 . The learning rate is $r = 0.0002$. We
 162 simulate nine parallel fast gradient descents randomly initialized sharing the same random feature
 163 sequence x^T , with $T = 3,000,000$. The gradient descent converges fast but not converge on the good
 164 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
 165 maxima which acts like a trap for the gradient descent.

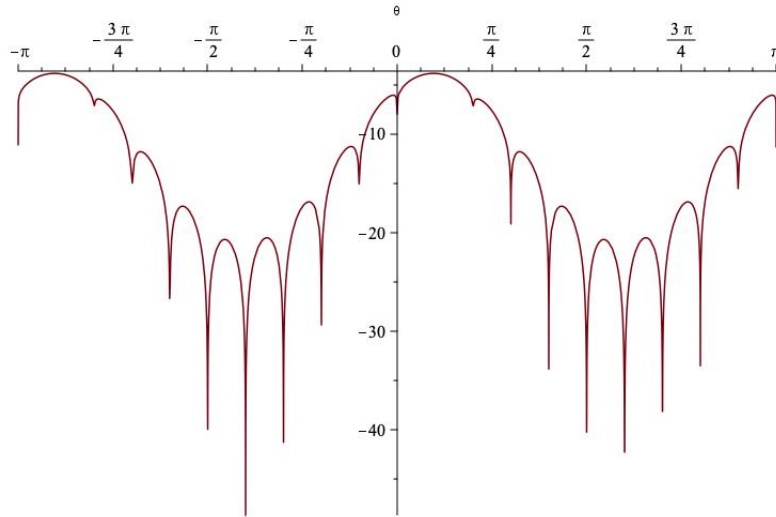


Figure 1. Target function $\sum_t \cos(\theta_Q - x_t)^2 \log \cos(\theta - x_t)^2 + \sin(\theta_Q - x_t)^2 \log \sin(\theta - x_t)^2$ as function of θ .

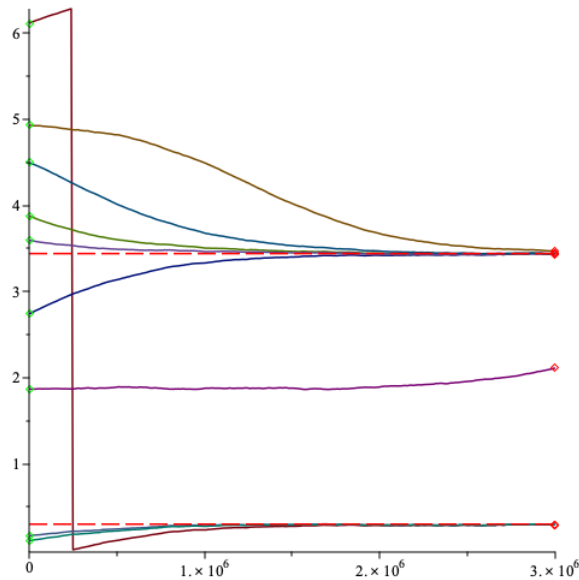


Figure 2. Angle estimate θ_t versus time of nine slow gradient descents randomly initialized. Green diamonds are starting points, red diamonds are stopping points.

166 **5. Conclusion**

167 We have presented a simple quantum tomography problem, the photon unknown polarization
 168 problem and have analyzed its learnability via AI over T runs. We have shown that the learning regret

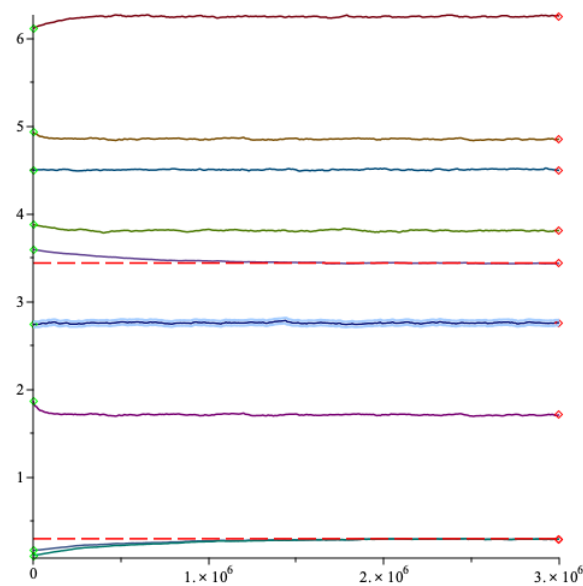


Figure 3. Angle estimate θ_t versus time of nine fast gradient descents randomly initialized. Green diamonds are starting points, red diamonds are stopping points.

169 cannot decay faster than $1/\sqrt{T}$ (i.e. a cumulative regret of \sqrt{T}). Furthermore the classic gradient
 170 descent are hampered by local extrema which may significantly impact the theoretical convergence
 171 rate.

172 References

- 173 1. Ben-David, Shai, et al. "Learnability can be undecidable." *Nature Machine Intelligence* 1.1 (2019): 44-48.
- 174 2. Abbe, Emmanuel, and Colin Sandon. "Provable limitations of deep learning." *arXiv preprint arXiv:1812.06369*
 175 (2018).
- 176 3. Van Erven, Tim, and Peter Harremo. "Rényi divergence and Kullback-Leibler divergence." *IEEE Transactions*
 177 *on Information Theory* 60.7 (2014): 3797-3820.
- 178 4. Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmis-*
 179 *sion*, 23(3):3–17, Jul.-Sep. 1987.
- 180 5. Tishby, Naftali, and Noga Zaslavsky. "Deep learning and the information bottleneck principle." *2015 IEEE*
 181 *Information Theory Workshop (ITW)*. IEEE, 2015.
- 182 6. Jacquet, Philippe, Gil Shamir, and Wojciech Szpankowski. "Precise Minimax Regret for Logistic Regression
 183 with Categorical Feature Values." *Algorithmic Learning Theory*. PMLR, 2021
- 184 7. Newey, Whitney K.; McFadden, Daniel. "Chapter 36: Large sample estimation and hypothesis testing".
 185 *Handbook of Econometrics*, 1994.
- 186 8. O'shea, Timothy, and Jakob Hoydis. "An introduction to deep learning for the physical layer." *IEEE*
 187 *Transactions on Cognitive Communications and Networking* 3.4 (2017): 563-575.
- 188 9. Bouillard, Anne, and Philippe Jacquet. "Quasi Black Hole Effect of Gradient Descent in Large Dimension:
 189 Consequence on Neural Network Learning." *ICASSP 2019-2019 IEEE International*
- 190 10. Hendrikx, Hadrien, Francis Bach, and Laurent Massoulié. "Accelerated decentralized optimization with
 191 local updates for smooth and strongly convex objectives." *The 22nd International Conference on Artificial*
 192 *Intelligence and Statistics*. PMLR, 2019.
- 193 11. N Garcia Trillos, F. Morales, J. Morales "Traditional and accelerated gradient descent for neural architecture
 194 search", *Geometric Science of Information*, 2021, Springer, page 507-514