Introduction to Quantum Learning Theory

Last compiled on May 24, 2022

These notes are based on the lectures given by Prof. Marco Tomamichel as part of the course *Advanced Topics in Quantum Information Theory*. The notes were typewritten by Jan Seyfried and revised by Prof. Marco Tomamichel.[1]

Contents

[1] Recordings of the lecture are available at: https://video.ethz.ch/lectures/d-phys/2022/spring/402-0462-00L.html

1 Introduction

1.1 Overview

Learning theory can be seen as the theoretical underpinning of machine learning. We analyze simplified models which hopefully allow for a rigorous analysis. This creates some tension with practice, as there models are more complex and rigorous guarantees are superfluous as long as "things work". Still, analytical results can give us understanding of why more complex models with practical relevance work.

Quantum learning theory is at the intersection of machine learning, statistics and quantum information and is a nice application area of (quantum) information theory tools. There are different aspects of quantum learning theory, depending on whether quantum information is applied in the learning algorithm, or if the data under consideration is a quantum system. The table below gives an overview over keywords and links to the relevant sections.

algorithm \ data	classical	quantum
c	classical learning theory	state tomography multiarmed quantum bandits
q	quantum PAC learning	?

Learning something from a quantum system using classical computation is relevant for experimental sciences and the first topic we will consider. Chapter 2 will treat *quantum state tomography*, the problem of learning the density matrix of a state. Closer to machine learning is the case where quantum algorithms are used to learn classical data - with the hope of obtaining an advantage with respect to classical algorithms. Chapter 3 provides an introduction to *quantum PAC learning*. The case where we apply quantum algorithms to quantum data is not treated here.

1.2 Notation

Familiarity with the following objects is assumed:

- **Schatten p-norm**: $||T||_p = \left(\operatorname{tr}[|T|^p]\right)^{1/p}$, where $|T| = \sqrt{(T^*T)}$. We mostly use
  - $||X||_1 = \sum_{i=1}^{d} |\lambda_i(X)|$, given that X is a normal operator, for $\lambda_i(X)$ the ith eigenvalue of X,
  - $||X||_2 = \sqrt{\operatorname{tr} X^\dagger X}$.
- **Trace norm**: $||T||_{\text{tr}} := \frac{1}{2}||T||_1$, note in particular that for two states $\rho, \sigma \in S(H)$, the trace distance $||\rho - \sigma||_{\text{tr}}$ can be rewritten as an optimization problem:

$$||\rho - \sigma||_{\text{tr}} = \max_{0 \leq P \leq 1} \operatorname{tr}[P(\rho - \sigma)].$$

- **Entropy**: $H(\rho) = -\operatorname{tr}[\rho \log(\rho)]$ for state $\rho \in S(H)$.
- **Relative Entropy**: $D(\rho||\omega) = -\operatorname{tr}[\rho(\log(\rho) - \log(\omega))]$.
- **Mutual Information**: $I(X:B) = H(B) - H(B|X) = H(\sum_x P(x)\rho_B^x) - \sum_x P(x)H(\rho_B^x)$, for a classical-quantum state $\rho_{XB} = \sum_x P(x) |x\rangle \langle x| \otimes \rho_B^x$ (X is classical, B is quantum).
- **Pauli matrices**: $X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$, $Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

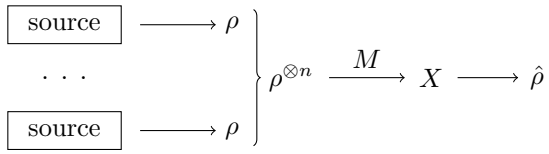
2 Quantum State Tomography

The basic problem of quantum state tomography is the reconstruction of a density matrix based on measurements. In more detail, we are given a source which produces (reliably) an unknown state $\rho \in S(H_d)$. We collect n of these states and perform a measurement on $\rho^{\otimes n}$, using a POVM[2] $M = \{M_x\}_{x \in X}$ ($M_x \geq 0$, $\sum_x M_x = \mathbb{1}_{d^n}$), which represents the most general form of measurement in quantum mechanics. X denotes the set of possible outcomes.

A specific outcome $x \in X$ is obtained with probability $\Pr[X = x] = \mathrm{tr}[\rho^{\otimes n} M_x]$ according to the *Born rule*. For such an outcome x, we then construct an estimate $\hat{\rho}(x)$ for ρ with the goal of achieving $\hat{\rho} \approx \rho$ with high probability. The measure of closeness is defined as follows:.

$$\Pr[||\hat{\rho}(X) - \rho||_{tr} \leq \epsilon] \geq 1 - \eta, \tag{1}$$

for η and ϵ small constants.



The above procedure is determined by the triple $\{X, M, \hat{\rho}\}$, which we call a *(quantum) tomography scheme*. Note that we could have chosen other criteria of closeness, for example we could require $\hat{\rho}(X)$ to be close in expectation, but this would later turn out to be too weak. To get an intuition for this definition, we look at a classical example, which also demonstrates why a perfect reconstruction is impossible (i.e. why ϵ and η cannot be just zero).

We consider Bernoulli trials, namely coin tosses: a coin shows either heads (H) or tails (T), and can be described by the probability to obtain heads, $\Pr[H] = p$ (note: $\Pr[T] = 1 - \Pr[H]$). Assume we have a coin for which p is unknown, but in n tosses, we observe only heads. In this case $\hat{p} = 1$ is the natural guess. However, it is possible that the actual p is quite different. Take for example a fair coin ($p = 1/2$): in this case the probability to see n successive heads is small, namely $p^n = 1/2^n$, but still finite. In this case it is thus not possible to guess p up to a precision less than $1/2$ with certainty, independent of how large n is chosen (as long as it is finite).

The ball $\{\sigma \in S(H_d) : ||\hat{\rho} - \sigma|| \leq \epsilon\}$ is called the *confidence region*. Equation 1 implies that ρ is in this region with probability $1 - \eta$. In the following, we are interested in the fundamental relationships between n, ϵ, and the dimension d. We assume η to be constant and will mostly deal with qubits for which we have $d = 2$. There are two possible approaches: either we provide a specific protocol and analyze the dependence of n, ϵ, and d, or we ask for fundamental limits.

2.1 Classical Warm Up

Consider i.i.d.[3] random variables $X_i \in \{-1, 1\}$, distributed according to $\Pr[X_i = -1] = p$, $\Pr[X_i = 1] = 1 - p$ for all i, where p is unknown. This problem is equivalent to the coin toss example from before (H, T correspond to $-1, 1$). Observing a sequence $X_1, X_2, ..., X_n$, we want to estimate the distribution $\hat{P}_x$, i.e., find $\hat{p}$. A good estimator should clearly be the *empirical frequency*,

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = -1\}.$$

An estimator is called *unbiased* if $\mathbb{E}[\hat{\rho}(X)] = \rho$, which is the case for the empirical frequency:

$$\mathbb{E}[\hat{p}(x)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{X_i = -1\}] = p.$$

[2] Positive Operator-Valued Measure
[3] Independent and Identically Distributed

To find the desired relationship between n, ϵ, and d, we make use of a concentration bound, namely the *Hoeffding inequality*: for $Y = \sum_{i=1}^{n} Y_i$ a sum of n i.i.d. random variables with $a \leq Y_i \leq b$, it holds that

$$\Pr[|Y - \mathbb{E}[Y]| > \delta] \leq 2e^{-\frac{2\delta^2}{n(b-a)^2}}. \tag{2}$$

In our case, $Y = \hat{p}(x)$ for $Y_i = \mathbb{1}\{X_i = -1\}/n$, which means $0 \leq Y_i \leq 1/n$ for all i. Hence

$$\Pr\Big[||\hat{P}(X) - P||_{\text{tr}} > \epsilon\Big] = \Pr[|\hat{p}(x) - p| > \epsilon] \leq 2e^{-2n\epsilon^2}.$$

Comparing this result to Equation 1, we find $\eta(n, \epsilon) = 2e^{-2n\epsilon^2}$.

2.2 Quantum Case

We will now consider the quantum case, more precisely reconstructing the density matrix of a qubit. The classical example before had only a single parameter, but now there exist multiple: a good parametrization needs to chosen. The idea is to measure Pauli operators and reconstruct the state from this. In the following we use the abbreviations $r_x = \text{tr}[\rho X]$, $r_y = \text{tr}[\rho Y]$, $r_z = \text{tr}[\rho Z]$ and $\Delta i = r_i - r_i'$. Recall the Bloch representation:

$$\rho = \frac{1}{2}(\mathbb{1} + r_x X + r_y Y + r_z Z) \qquad \text{with } |r|^2 = r_x^2 + r_y^2 + r_z^2 \leq 1.$$

For a qubit we want to minimize

$$\begin{aligned} ||\rho - \rho'||_{\text{tr}} &= \frac{1}{2}||(r_x - r_x')X + (r_y - r_y')Y + (r_z - r_z')Z||_{\text{tr}} \\ &= \frac{1}{2}\left|\left|\begin{pmatrix} \Delta z & \Delta x - i\Delta y \\ \Delta x + i\Delta y & -\Delta z \end{pmatrix}\right|\right|_{\text{tr}} \\ &= \frac{1}{2}||r - r'||_2. \end{aligned} \tag{3}$$

The last step requires the calculation of the eigenvalues, which for the case of 2×2 matrices, is easily done by using identities from linear algebra for trace and determinant including the eigenvalues λ_i,

$$\text{tr}[D] = 0 = \lambda_1 + \lambda_2, \qquad \det D = -\Delta z^2 - \Delta x^2 - \Delta y^2 = \lambda_1 \lambda_2.$$

We find $\lambda_{1,2} = \pm\sqrt{\Delta z^2 + \Delta x^2 + \Delta y^2}$.

We will now consider the following tomography procedure: (assume for simplicity $n \bmod 3 = 0$)

1. Measure $m = n/3$ states using X and record $x_1, x_2, ..., x_m$. Do the same for Y and Z.
2. Compute $\hat{r}_x = \frac{1}{m}\sum_{i=1}^{m} x_m$, and obtain $\hat{r}_y$ and $\hat{r}_z$ analogously.
3. Calculate $\tilde{\rho} = \frac{1}{2}(\mathbb{1} + \hat{r}_x X + \hat{r}_y Y + \hat{r}_z Z)$, use it to estimate $\hat{\rho}$.

Note that we cannot use $\tilde{\rho}$ directly as $\hat{\rho}$, as it might not actually be a state: $\hat{r}_x^2 + \hat{r}_y^2 + \hat{r}_z^2 \leq 1$ can be violated! However, we can project it onto the Bloch sphere to obtain a possible $\hat{\rho}$. From a statistical point of view, this operation can only improve the estimate. The downside is that such projections create pure states (which correspond to the surface of the Bloch sphere). It is important to consider this artifact caused by the projection to prevent wrong physical conclusions. Another option is to simply consider the full confidence region, not only the estimate itself.

Equation 3 allows us to calculate the probability with which we fail to get close to the true state. Note by

comparing to Equation 1 that the following probability is equal to our definition of η:

$$\begin{aligned}\Pr[||\rho-\hat{\rho}||_{\mathrm{tr}}>\epsilon] &= \Pr[||r-\hat{r}||_2>2\epsilon] \\ &= \Pr\left[\sum_{i\in\{x,y,z\}}|r_i-\hat{r}_i|^2>4\epsilon^2\right] \\ &\leq \Pr\left[\bigvee_{i\in\{x,y,z\}}\left(|r_i-\hat{r}_i|^2>\frac{4}{3}\epsilon^2\right)\right] \\ &\leq \sum_{i\in\{x,y,z\}}\Pr\left[|r_i-\hat{r}_i|>\frac{2}{\sqrt{3}}\epsilon\right].\end{aligned}$$

The last inequality follows from using the *union bound* and applying the square root on both sides of the inequality. Considering a single summand, say $i=x$, we note that $r_x=\mathbb{E}[\hat{r}_x]$. This allows us to apply the *Hoeffding bound* for $Y=r_x$, $Y_i=x_i/m$ and $-1/m\leq Y_i\leq 1/m$ since $-1\leq x_i\leq 1$ for all i. After applying Hoeffding for all summands we obtain

$$Pr[||\rho-\hat{\rho}||_{tr}>\epsilon]\leq 3\cdot 2e^{-\frac{2\left(\frac{2\epsilon}{\sqrt{3}}\right)^2}{\sum_{i=1}^{m}(-2/m)^2}}=6e^{-\frac{m}{2}\left(\frac{2\epsilon}{\sqrt{3}}\right)^2}=6e^{-\frac{2}{9}n\epsilon^2}=\eta$$

Solving for n tells us how to choose n as a function of η and ϵ:

$$6e^{-\frac{2}{9}n\epsilon^2}=\eta\Rightarrow\log\left(\frac{\eta}{6}\right)=-\frac{2}{9}n\epsilon^2\Rightarrow n=\frac{9}{2}\log\left(\frac{6}{\eta}\right)\frac{1}{\epsilon^2}.$$

Considering ϵ, we need $n=O(1/\epsilon^2)$ many samples to achieve a confidence region of ϵ. In the following section we will show that this asymptotic bound is indeed optimal.

2.3 Sample-Optimal Quantum State Tomography

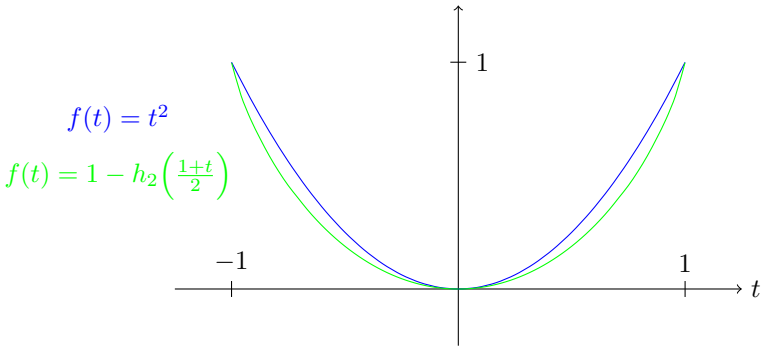
Consider a communication protocol over a classical-quantum channel $i\to\tau_{t,i}$, $i\in\{1,2,3\}\mapsto S(H_2)$, where

$$\tau_{t,1}=\frac{1}{2}(\mathbb{1}+tX),\quad \tau_{t,2}=\frac{1}{2}(\mathbb{1}+tY),\quad \tau_{t,3}=\frac{1}{2}(\mathbb{1}+tZ).$$

Note that due to symmetry, $H(\tau_{t,i})$ is the same for all i. We can show that for t small, the *capacity* C of this channel is low. Intuitively, the different τ become more similar the smaller t is, as the 'signal' (X,Y,Z) gets smaller with respect to the 'noise' $(\mathbb{1})$.

$$\begin{aligned}C=\max_{P_x}I(X:B)&=\max_{P_x}H\Big(\sum_x P_x(x)\tau_{t,x}\Big)-\sum_x P_x(x)H(\tau_{t,x}) \\ &\leq 1-H(\tau_{t,x}) \\ &=1-h_2\Big(\frac{1+t}{2}\Big) \\ &\leq t^2\end{aligned}$$

Here we used the *binary entropy*, which is defined as $h_2(p)=-p\log_2(p)-(1-p)\log_2(1-p)$. The last inequality can easily be verified using the following plot:



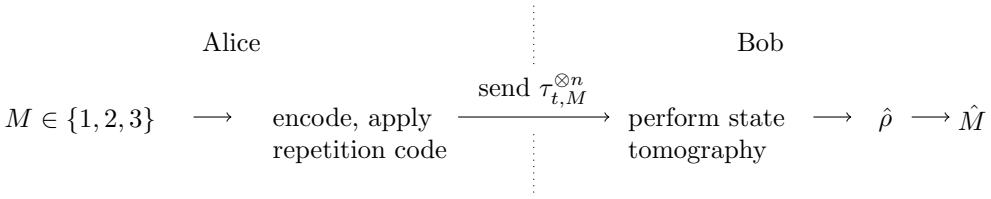
Alternatively, we can also give an analytical proof: define $\Delta(t) = t^2 - (1 - h_2((1+t)/2))$, which is 0 only for $t \in \{-1, 0, 1\}$. $\Delta(t)$ is also continuously differentiable and, for $t \in (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, convex since we have $\frac{\partial \Delta^2}{\partial t^2} = 2 - \frac{1}{1-t^2} > 0$ for such t. In this interval it follows that $\Delta(t) \geq 0$, due to convexity and $\Delta(0) = 0$ being a minimum. But since $\Delta(t)$ is zero only for $t \in \{-1, 0, 1\}$, we conclude that $\Delta(t) \geq 0$ has to hold for all $t \in [-1, 1]$.

Later we will also need that the different states τ_i are not too close together. In the following we show this for the case $\tau_{t,1}, \tau_{t,3}$, but due to symmetry the other cases follow similarly:

$$\begin{aligned} ||\tau_{t,1} - \tau_{t,3}||_{\text{tr}} &= \frac{t}{2}||X - Z||_{\text{tr}} \\ &= \frac{t}{2}||(2\,|+\rangle\langle+| - \mathbb{1}) - (2\,|0\rangle\langle 0| - \mathbb{1})||_{\text{tr}} \\ &= t||\,|+\rangle\langle+| - |0\rangle\langle 0|\,||_{\text{tr}} \\ &= t\sqrt{1 - |\langle +|0\rangle|^2} = \frac{t}{\sqrt{2}}. \end{aligned}$$

Note that while the capacity scales quadratically in t, the distance between states depends only linearly on t! Now set $t = 4\epsilon$ so that $||\tau_{t,i} - \tau_{t,j}||_{tr} = 2\sqrt{2}\epsilon > 2\epsilon$.

With these preliminaries, we can now derive a lower bound for n using a protocol for transmitting messages from Alice to Bob, which works as follows:



Alice picks a message and encodes it to the respective state τ. Then she applies the repetition code and sends the resulting n identical states to Bob. He performs state tomography where we assume the following equation to hold for n of a certain minimal size:

$$\Pr[||\tau_{4\epsilon,M} - \hat{\rho}||_{\text{tr}} \leq \epsilon] \geq 1 - \eta. \tag{4}$$

Finally, Bob can read off $\hat{M}$ based on his result for $\hat{\rho}$. The theory of channel coding can give us a lower bound on n for reconstructing M correctly, dependent on ϵ and η. On the other hand, if state tomography as assumed in Equation 4 could be performed with less samples than this bound predicts, we would clearly be able to send more information through the channel than is allowed by capacity. We thus arrive at a contradiction and have shown the desired lower bound.

Since the states are at least 2ϵ apart, at most one $\tau_{4\epsilon,i}$ lies in an ϵ-ball around $\hat{\rho}$. With Equation 4 it follows that with probability $1 - \eta$, exactly one $\tau_{4\epsilon,i}$ lies inside the ball. When decoding, we pick $\hat{M} = i$

where i is such that $||\hat{\rho} - \tau_{4\epsilon,i}||_{tr} \leq \epsilon$, if such an i exists. Otherwise we pick an arbitrary i. Note that $\Pr[\text{error}] = \Pr\Big[M \neq \hat{M}\Big] \leq \eta$.

Next we want to find a bound for n. *Fano's inequality* and the channel coding converse ensure that

$$\log |M| \leq \frac{nC + 1}{1 - \eta}. \tag{5}$$

We will show this inequality below, but for now we directly use it ($|M| = 3$, $C \leq 16\epsilon^2$) and solve for n:

$$n \geq \frac{1}{16\epsilon^2}((1 - \eta)\log(3) - 1).$$

We note that if η is small enough then $(1 - \eta)\log(3) > 1$ and, thus, we can conclude that $n = \Omega(1/\epsilon^2)$.

Now we prove Equation 4: Given that $\Pr\Big[M \neq \hat{M}\Big] \leq \eta$, Fano's inequality yields

$$H(M|\hat{M}) \leq h_2(\eta) + \eta \log(|M| - 1),$$

and using $h_2 \leq 1$, we can bound it by $H(M|\hat{M}) \leq 1 + \eta \log(|M|)$.

$$I(M : \hat{M}) = H(M) - H(M|\hat{M}) \geq \log(|M|) - 1 - \eta \log(|M|) = (1 - \eta)\log(|M|) - 1. \tag{6}$$

The inequality in the second step is obtained by using Fano and the fact that $H(M) = log(3)$ when we try to send the three messages uniformly at random. Note that $M \to X^n \to B^n \to \hat{M}$ forms a Markov chain, as every step is dependent solely on its predecessor and is conditionally independent of states further behind. This insight allows us to make use of the *data-processing inequality*:

$$I(M : \hat{M}) \leq I(X^n : B^n),$$

which states essentially that data processing can never increase information. Our channels use a specific code (the repetition code), but now we bound the mutual information by an optimization over a distribution of inputs: $I(X^n : B^n) \leq \max_{P^n} I(X^n : B^n)$. It can be shown that this is equal to nC and we arrive at

$$I(M : \hat{M}) \leq nC. \tag{7}$$

Combining Equations 6 and 7 directly results in Equation 5.

In the homework we consider the case $d \gg 2$ and construct a set of $|M| = e^{d^2 k}$ $(k > 0)$ states which are all at least 2ϵ apart and which constitute a classical-quantum channel with capacity $C' = O(\epsilon^2)$.

$$\log(|M|) \leq \frac{nC' + 1}{1 - \eta} \quad \text{now yields} \quad n = \Omega\Big(\frac{d^2}{\epsilon^2}\Big).$$

To achieve this, we actually need entangled measurements. Note that it is not sufficient to estimate every matrix element up to accuracy ϵ, since this does not ensure $||...||_{\text{tr}} \leq \epsilon$.

2.4 An Online Algorithm for State Tomography

In this setting, we perform a sequence of measurements. After each step, we update our estimate of the state using only the new information obtained by the current measurement. We denote the measurements and their outcomes by σ_i and y_i respectively. We assume no initial knowledge about the state, which is why we set $\hat{\rho}_0 = \mathbb{1}/d$, the uniform state.

$$\hat{\rho}_0 = \frac{\mathbb{1}}{d} \underset{\sigma_1, y_1}{\longrightarrow} \hat{\rho}_1 \underset{\sigma_2, y_2}{\longrightarrow} \hat{\rho}_2 \longrightarrow \cdots \longrightarrow \hat{\rho} \approx \rho$$

In the following we assume $d = 2^n$ (a system of n qubits) and for all i, take σ_i to be a random chain of Pauli matrices of length n, $\sigma_i \in \{X \otimes Y \otimes X..., \mathbb{1} \otimes X \otimes Z..., \dots\}$. Excluding the all-identity chain, there are $4^n - 1 = d^2 - 1$ different measurements σ_i we can perform.

When thinking about how to update our estimate, it is clear that there is a trade-off between updating and maintaining existing knowledge, which we can formalize as follows:

- To not update the state too much, we want to keep the relative entropy $D(\hat{\rho}_{t+1}||\hat{\rho}_t)$ small.
- To improve our estimate, we use the following loss function:

$$L_+(\hat{\rho}_+) = (\mathrm{tr}(\hat{\rho}_+\sigma_+) - \mathrm{tr}(\rho\sigma_t))^2.$$

  As a simplification, we assume that we can observe $\hat{y}_t = \mathrm{tr}(\rho\sigma_t)$ directly, which means without noise. Otherwise we would only have $\mathbb{E}[\hat{y}_t] = \mathrm{tr}(\rho\sigma_t)$. In practice, this assumption could only be realized by using an infinite amount of samples.

Now we combine the two targets into the following online loss function optimization:

$$\text{find } \hat{\rho}_{t+1} \text{ such that } P(\hat{\rho}_{t+1}||\hat{\rho}_t) + \eta L_+(\hat{\rho}_{t+1}) \text{ is minimal,}$$

where η is a parameter called the *learning rate*, responsible to realize the trade-off mentioned before.
This yields (up to some approximation) the update rule

$$\hat{\rho}_{t+1} = \frac{1}{C} e^{\log(\hat{\rho}_t) - \eta \nabla L_t(\hat{\rho}_t)},$$

where C is a normalization constant and the gradient of the loss function is given by $\nabla L_t(\hat{\rho}_t) = 2(\mathrm{tr}(\hat{\rho}_t\sigma_t) - \hat{y}_t)\sigma_t$. By definition, $\hat{\rho}_{t+1}$ is always positive semi-definite as it is the exponential of some quantity. However, we need to normalize it by hand. Again, the update rule only uses data obtained by the last measurement to update $\hat{\rho}_t$. Advantages compared to the case where all gathered information is used in each step are:

- Efficiency: amount of stored data and computation in each step is significantly reduced.
- This approach can follow changes in the state over time (to some extent). This is important since in actual experiments it is usually not possible to keep the conditions constant for the duration of the experiment.

2.5 Sketch of Convergence Proof

(Further details of this proof are part of the exercises.) We want to show convergence, namely that

$$\lim_{t\to\infty} \Pr[||\rho_t - \rho||_{\mathrm{tr}} < \delta] = 1 \quad \text{for all } \delta > 0.$$

This is a probabilistic statement since the σ_i's are picked at random in each step. If we would loosen our assumption on $\hat{y}_t$, noise would introduce additional randomness.

We start with the following bound[4]

$$\eta L_t(\hat{\rho}_t) \leq D(\rho||\hat{\rho}_t) - D(\rho||\hat{\rho}_{t+1}) \qquad \text{for } \eta > \frac{1}{2}.$$

Since the left-hand side is always positive, the update brings the state closer to ρ in relative entropy. The values of L and η influence by how much.
Now we sum this inequality over t. On the right hand side, most terms of the relative entropy cancel:

$$\eta \sum_{t=0}^{T} L_t(\hat{\rho}_t) \leq D(\rho||\hat{\rho}_0) - D(\rho||\hat{\rho}_{T+1}) \leq D(\rho||\hat{\rho}_0) \leq \log(d).$$

The last inequality follows from $\rho_0 = \mathbb{1}/d$:

$$D(\rho||\hat{\rho}_0) = \mathrm{tr}(\rho\log(\rho) - \rho\log(\hat{\rho}_0)) \leq -\mathrm{tr}(\rho\log(\hat{\rho}_0)) = \log(d)\,\mathrm{tr}(\rho) = \log(d).$$

[4] for details see arXiv:1807.01852

We now take the limit $T \to \infty$ and obtain $\eta \sum_{t=0}^{\infty} L_t(\hat{\rho}_t) \leq \log(d)$. Note that this is an infinite sum bounded by a constant. Taking the expectation of this bound yields $\eta \sum_{t=0}^{\infty} \mathbb{E}[L_t(\hat{\rho}_t)] \leq \log(d)$. We now analyze an individual summand:

$$\begin{aligned}
\mathbb{E}_{\sigma_t}[L_t(\hat{\rho}_t)] &= \mathbb{E}_{\sigma_t}[(\mathrm{tr}(\sigma_t \hat{\rho}_t) - \mathrm{tr}(\sigma_t \rho))^2] \\
&= \mathbb{E}_{\sigma_t}[(\sigma_t \otimes \sigma_t)(\hat{\rho}_t - \rho) \otimes (\hat{\rho}_t - \rho)] \\
&= \frac{d}{d^2-1} \mathrm{tr}(P_{12}(\hat{\rho}_t - \rho) \otimes (\hat{\rho}_t - \rho)) - \frac{1}{d^2-1} \mathrm{tr}((\hat{\rho}_t - \rho) \otimes (\hat{\rho}_t - \rho)) \\
&= \frac{d}{d^2-1} \mathrm{tr}\big((\hat{\rho}_t - \rho)^2\big) \\
&= \frac{d}{d^2-1} ||\hat{\rho}_t - \rho||_2^2.
\end{aligned} \tag{8}$$

The third step uses Exercise 3b with the *swap operator* $P_{12} = \sum_{i,j} |i\rangle \langle j| \otimes |j\rangle \langle i|$. The next step follows from the fact that $(\hat{\rho}_t - \rho) \otimes (\hat{\rho}_t - \rho)$ is traceless and Exercise 3c. Intuitively, we took the expectation over measurements which are evenly distributed over the space of all measurements (remember how we defined the Pauli chains). Intuitively speaking, we "checked all directions". As a result, in expectation our loss function corresponds to the Schatten 2-norm.
Applying Equation 8 to $\eta \sum_{t=0}^{\infty} \mathbb{E}[L_t(\hat{\rho}_t)] \leq \log(d)$ gives

$$\eta \sum_{t=0}^{\infty} \frac{d}{d^2-1} \mathbb{E}[||\hat{\rho}_t - \rho||_2^2] \leq \log(d) \quad \Rightarrow \quad \sum_{t=0}^{\infty} \mathbb{E}[||\hat{\rho}_t - \rho||_2^2] \leq \frac{d^2-1}{d\eta} \log(d),$$

from which it follows that

$$\lim_{t \to \infty} \mathbb{E}[||\hat{\rho}_t - \rho||_2^2] = 0. \tag{9}$$

Otherwise, we would have an infinite number of elements bigger than some δ and the sum would be unbounded.

Next we use *Markov's inequality*: for X a nonnegative random variable and $a > 0$ a positive constant, it holds that $\Pr[X \geq a] \leq \mathbb{E}[X]/a$. For $a = \delta$ and $X = ||\hat{\rho}_t - \rho||$ together with $Pr[X \geq a] = 1 - Pr[X < a]$, we find

$$\Pr\big[||\hat{\rho}_t - \rho||_2^2 < \delta\big] \geq 1 - \frac{\mathbb{E}[||\hat{\rho}_t - \rho||_2^2]}{\delta}.$$

Taking the limit $t \to \infty$ and using Equation 9 it follows that

$$\lim_{t \to \infty} \Pr\big[||\hat{\rho}_t - \rho||_2^2 < \delta\big] = 1 \quad \text{for all } \delta > 0.$$

Finally, inequalities between different norms can be used to show that any norm distance needs to tend towards zero in this limit, which then proves the desired convergence.

3 Quantum PAC Learning

In this chapter we consider one of the fundamental models in machine learning, *Probably Approximately Correct* (PAC) learning. This is similar to what we did for tomography: there is a confidence interval ('approximately') which we want to hit with high probability ('probably').

3.1 Classical PAC Learning

The problem setting is as follows:

- The *instance space* X is a set, for example $X = \{0, 1\}$, or $[0, 1] \subset \mathbb{R}$ etc
- A *concept* c is a map $X \to \{0, 1\}$, or alternatively a subset of X containing all $x \in X$ that map to 1.
- A *concept class* C is a collection of concepts.

These terms will be used in both the classical and quantum context. We also have an *oracle*, which works differently dependent on the case we look at:

- classical: the oracle gives us a sample $(x, c(x))$ upon request, where $x \leftarrow D$ for some distribution D on X.
- quantum: the oracle gives superposition access, meaning it gives us the state

$$|\Psi\rangle = \sum_{x \in X} \sqrt{D(x)}\, |x\rangle \otimes |c(x)\rangle\,. \tag{10}$$

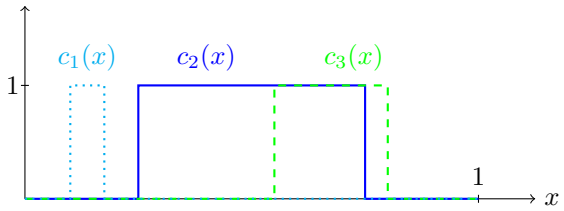
  It seems like the quantum oracle is much stronger, but note that we need to access the information contained in $|\Psi\rangle$ somehow. For example, measuring the first register of $|\Psi\rangle$ corresponds to sampling from the classical oracle. But as a quantum learner, we might also use $|\Psi\rangle$ in a different way (see the Grover algorithm).

The goal of the learner is to find with probability at least $1 - \delta$ a *hypothesis* h such that

$$\Pr_{x \leftarrow D}[h(x) \neq c(x)] \leq \epsilon.$$

The notation $\Pr_{x \leftarrow D}$ means that we want the inequality to hold for an x sampled randomly from D. The reasoning for why neither δ nor ϵ can be zero is similar to the tomography case. We will now explore how many samples/states we need to achieve a pair (ϵ, δ), and start with a classical example.

Consider $X = [0, 1]$, the concept class is formed by intervals in X (connected subsets of X).

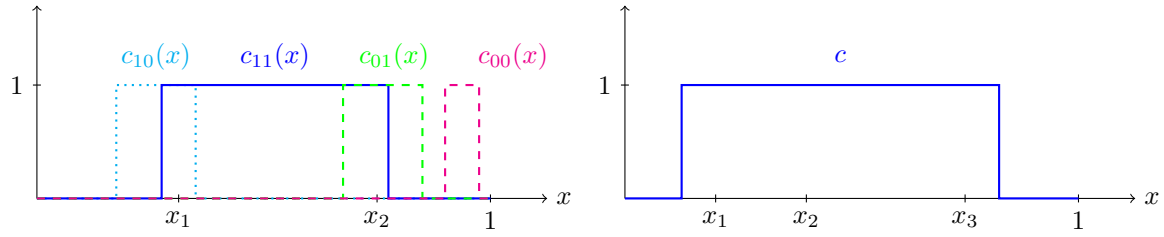


This is a very simple concept class. The difficulty of a concept class is generally characterized by the *VC-dimension* (Vapnik, Chervonenkis):

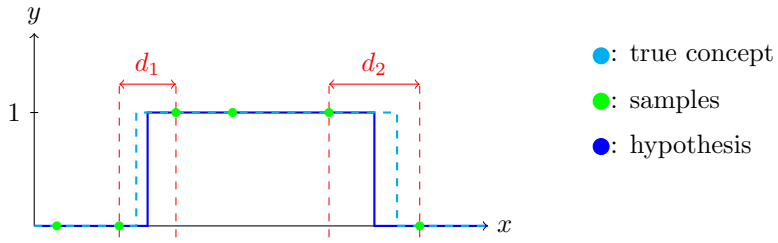
A set $S = \{x_1, x_2, ...x_d\}$ is said to be *shattered* by C if for every $a \in \{0, 1\}^d$, there exists a concept $c_a \in C$ such that $a = (c_a(x_1), c_a(x_2), ...c_a(x_d))$. The *VC-dimension* of C is the size of the largest set shattered by C.

In our example, VC = 2: this is illustrated in the Figure below. The left picture shows that sets of size 2 can be shattered, with the subscript of the respective c denoting the corresponding bit string a.

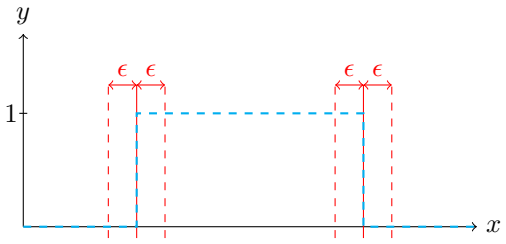
The right picture shows that it is not possible to shatter a set of size 3: a concept c_{101} cannot be achieved using a single interval.



Assume now that $x \in [0, 1]$ is uniform. The Figure below shows an example for a concept we want to learn, together with samples we obtained from it. We choose our classical learner to work as follows: the hypothesis h is determined by defining the bounds of the interval to lie in the middle between neighboring points which are evaluated differently by the oracle (their distances are denoted by d_1 and d_2 in the Figure).



We are interested in a sufficient condition for h to be an ϵ-approximation. For this, we define four regions of size ϵ each, placed at the bounds of the interval of the true concept. If we obtain a sample within each of these regions, then our choice for the boundaries of h can be off by at most $\epsilon/2$ each, and thus $\Pr[h(x) \neq c(x)] \leq \epsilon$.



Let the four regions be iterated by i, and define A_i to be an indicator variable for the event where we miss region i with each of the n samples. If we take n (uniformly sampled) samples, A_i occurs with probability $(1 - \epsilon)^n$. Using union bound, we find that the probability p of missing at least one of the four regions can be bounded as follows:

$$p = \Pr\left[\bigvee_i A_i\right] \leq \sum_i \Pr\left[A_i\right] = 4(1 - \epsilon)^n.$$

If we hit all regions, we are certainly within the desired bound. This means the probability δ to fail being ϵ close to the true concept is upper bounded by p, $\delta \leq p$. Hence we obtain $\delta \leq 4(1 - \epsilon)^n$, which we can solve for n:

$$n = \frac{\log(\delta/4)}{\log(1 - \epsilon)} \approx \frac{\log(4/\delta)}{\epsilon} \quad \Rightarrow \quad n = O\left(\frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)\right).$$

Testing the behavior of n for different numbers helps to obtain an intuition:

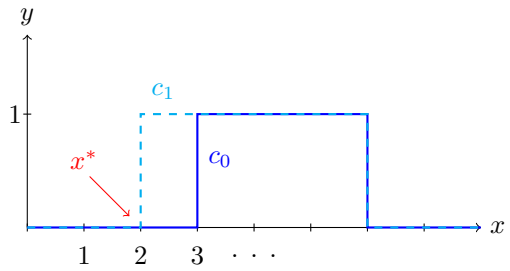
- if we change δ from 10^{-5} to 10^{-10} we need to increase n by a factor of 2.

- if we change ϵ from 0.02 to 0.01, we need to increase n by a factor of 2.

It is cheap (in terms of number of samples) to obtain a high confidence, but expensive to improve the approximation parameter ϵ.

3.2 A Quantum Advantage?

Now we are interested in whether we can obtain an advantage using quantum algorithms. To simplify things, we discretize X into bins and obtain $\tilde{X} = \{1, 2, ...\}$ ($\tilde{X}$ contains the indices of the bins) and denote the resulting distribution by D. We make this discretization such that each bin has probability $2\epsilon > D(x) > \epsilon$. We want to show a lower bound on n for a quantum learner. For this consider two concepts c_0 and c_1 which differ in exactly one point, x^*.



No hypothesis h satisfies both $\Pr[h(x) \neq c_0(x)] \leq \epsilon$ and $\Pr[h(x) \neq c_1(x)] \leq \epsilon$: due to our choice for D, $\Pr[h(x) \neq c_i(x)] \leq \epsilon$ implies $h = c_i$, hence c_0 and c_1 would need to be equal. The consequence is that a (ϵ, δ)-learner has to be able to distinguish between 0 and 1 with probability at least $1 - \delta$, meaning he has to learn the correct hypothesis exactly with high probability.

Recall Equation 10 defining the states the quantum oracle produces for c_0 and c_1. We now rewrite $|\Psi_0\rangle$ and $|\Psi_1\rangle$, making use of the fact that they differ in x^*, where we assume without loss of generality that $c_0(x^*) = 0$ and $c_1(x^*) = 1$:

$$|\Psi_i\rangle = \sum_{x \in X} \sqrt{D(x)}\,|x\rangle \otimes |c_i(x)\rangle = \sqrt{D(x^*)}\,|x^*\rangle \otimes |i\rangle + \sqrt{1 - D(x^*)}\,|\phi\rangle\,, \quad \text{for } i \in \{0, 1\}$$

and $|\phi\rangle$ normalized. We have $|\langle\Psi_0|\Psi_1\rangle| = 1 - D(x^*) \geq 1 - 2\epsilon$. As argued before, the PAC-learner needs to distinguish these two states. The number of samples required to achieve this is then a lower bound for the number of samples the learner needs in general.

The probability to correctly identify $i \in \{0, 1\}$ given either $|\Psi_0\rangle^{\otimes n}$ or $|\Psi_1\rangle^{\otimes n}$ (note that we are now looking at n copies of the respective state) is given by *Helstrom's Theorem* as

$$\begin{aligned} p =& \frac{1}{2}(1 + ||\,|\Psi_0\rangle\langle\Psi_0|^{\otimes n} - |\Psi_1\rangle\langle\Psi_1|^{\otimes n}\,||_{tr}) \\ =& \frac{1}{2}\left(1 + \sqrt{1 - F(|\Psi_0\rangle\langle\Psi_0|^{\otimes n}, |\Psi_1\rangle\langle\Psi_1|^{\otimes n})}\right). \end{aligned}$$

The second step is the special case of the *Fuchs-van de Graaf inequality* for pure states, and $F(\rho, \sigma) = \left[\mathrm{tr}\left(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)\right]^2$ is the *fidelity*. The fidelity is additive and in case of the pure states equal to the overlap:

$$\begin{aligned} p =& \frac{1}{2}\left(1 + \sqrt{1 - |\langle\psi_0|\psi_1\rangle|^{2n}}\right) \\ \leq& \frac{1}{2}\left(1 + \sqrt{1 - (1 - 2\epsilon)^{2n}}\right). \end{aligned}$$

Since our (ϵ, δ)-PAC learner needs to distinguish $|\Psi_0\rangle$ and $|\Psi_1\rangle$ with probability $\geq 1-\delta$, we must have

$$\begin{aligned} & 1-\delta && \leq \frac{1}{2}(1+\sqrt{1-(1-2\epsilon)^{2n}}) \\ \Rightarrow \quad & 2(1-\delta)-1 && \leq \sqrt{1-(1-2\epsilon)^{2n}} \\ \Rightarrow \quad & 1-(1-2\delta)^2 && \geq (1-2\epsilon)^{2n} \\ \Rightarrow \quad & \log(4\delta) && \geq 2n\log(1-2\epsilon). \end{aligned}$$

Finally (note that $\log(1-2\epsilon) < 0$),

$$n \geq \frac{1}{2}\frac{\log(1/4\delta)}{\log(1-2\epsilon)} = \Omega\Big(\frac{\log(1/\delta)}{\epsilon}\Big). \qquad \left(\text{classical: } n = O\left(\frac{\log(1/\delta)}{\epsilon}\right)\right)$$

This means no quantum advantage for this particular class of problems! More generally, if we consider a concept class with VC-dimension d, it can be shown that

$$\boxed{n = \Theta\Big(\frac{d}{\epsilon} + \frac{\log(1/\delta)}{\epsilon}\Big), \qquad \text{for both classical and quantum algorithms.}}$$

There can be a quantum advantage though in terms of how efficiently a hypothesis can be computed (for example in terms of circuit depth). We can make a statement of the following form: if there is no polynomial-time algorithm for factoring primes, then there exists a concept class C which is efficiently quantum PAC learnable, but not efficiently classically PAC learnable. *Efficiently* means polynomial in time and depth. Also, in a more heuristic approach, there might be quantum algorithms which "behave better" in practice compared to classical counterparts, despite having the same asymptotic bounds.