Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states

Josep Lumbreras[1], Erkka Haapasalo[1], and Marco Tomamichel[1,2]

[1]Centre for Quantum Technologies, National University of Singapore, Singapore
[2]Department of Electrical and Computer Engineering, Faculty of Engineering, National University of Singapore, Singapore

We initiate the study of tradeoffs between exploration and exploitation in online learning of properties of quantum states. Given sequential oracle access to an unknown quantum state, in each round, we are tasked to choose an observable from a set of actions aiming to maximize its expectation value on the state (the reward). Information gained about the unknown state from previous rounds can be used to gradually improve the choice of action, thus reducing the gap between the reward and the maximal reward attainable with the given action set (the regret). We provide various information-theoretic lower bounds on the cumulative regret that an optimal learner must incur, and show that it scales at least as the square root of the number of rounds played. We also investigate the dependence of the cumulative regret on the number of available actions and the dimension of the underlying space. Moreover, we exhibit strategies that are optimal for bandits with a finite number of arms and general mixed states.

1 Introduction

Multi-armed bandits have been studied for nearly a century and it is nowadays a very active research field, in terms of both theory and applications [1, 2, 3, 4]. The multi-armed bandit problem is a simple model of decision making with uncertainty that lies in the class of classical reinforcement learning problems. Given a set of arms, a learner interacts sequentially with these arms sampling a reward at each round and the objective of the learner is to identify the arm with largest expected reward while maximizing the total cumulative reward. The problem that the learner faces is a trade-off between exploration and exploitation: one wants to explore all the arms to identify the best one but also wants to exploit those arms that give the best rewards. Bandit algorithms are online, which means that the strategy is adaptive at each round and is learned from previously observed events. This is one of the reasons why nowadays bandit algorithms are applied to online services such as advertisement recommendation [5] or dynamical pricing [6].

The problem was introduced by Thompson in 1933 [7] where the original motivation was to study medical trials. The problem was to decide which treatment to use in the next patient given a set of different treatments for a certain disease. The mathematical formalization and popularization of the field is due to the mathematician and statistician

Robbins [8], who described the problem as a set of arms such that the reward associated to each arm follows an independent unknown probability distribution. This model is usually referred to as multi-armed stochastic bandits.

The main technique to solve this problem is based on upper confidence bounds that were introduced by Lai and Robbins [9]. There are many different models of bandits, including adversarial models, but the main setting that we are interested for our work are the multi-armed stochastic linear bandits, first considered by Aurer in [10]. In this model, the arms can be viewed as vectors and the expected reward of each arm is given by the inner product of the vector associated to the arm and an unknown vector that is the same for all arms. Since the literature of bandits is very extensive we refer to the book by Lattimore and Szepesvári [1] for a comprehensive review of bandit algorithms.

Quantum algorithms for the classical multi-armed stochastic bandit problem have been proposed recently [11, 12]. A quantum version of the Hedging algorithm, which is related to the adversarial bandit model, has also been studied [13]. These algorithms investigate potential improvements on the respective classical bandit algorithms when a quantum learner is given superposition access to the oracle, i.e., it can probe rewards for several arms in superposition.

Our work treats multi-armed bandits from a different perspective, focusing instead on the learning theory for quantum states. In our model, which we call the *multi-armed quantum bandit* model, the arms correspond to different observables or measurements and the corresponding reward is distributed according to Born's rule for these measurements on an unknown quantum state. We are interested in the optimal tradeoff between exploration (i.e. learning more about the unknown quantum state) and exploitation (i.e. using acquired information about the unknown state to choose the most rewarding but not necessarily most informative measurement). More precisely, we consider a learner that at each round has access to a copy of an unknown quantum state ρ (the *environment*) and has to choose an observable from a given set $\mathcal{A}$ (the *action set*) in order to perform a measurement and receive a reward. The *reward* is sampled from the probability distribution associated with the measurement of ρ using the chosen observable. The figure of merit that we study is the *cumulative expected regret*, the sum over all rounds of the difference between the maximal expected reward over $\mathcal{A}$ and the expected actual reward associated to the chosen observable at each round. If the actions set is comprised of all rank-1 projectors, we arrive at the problem of finding the maximal eigenvector of the unknown state. The detailed model is introduced in Section 2.

This notion of regret has a natural interpretation in some physical settings. Consider for example a sparse source of single photons with fixed but unknown polarisation (i.e., the reference frame is unknown). In order to learn the unknown reference frame we can perform a phase shift and apply a polarisation filter, adjusting the phase (the action) until the photons consistently pass the filter (the reward). The regret would then be proportional to the energy absorbed in the filter during our learning process. As mentioned above, classical bandit algorithms find applications in recommendation systems, and it is a valid question whether quantum bandit models could be used similarly for recommendation systems for quantum devices. However, we think that our present model does not capture any interesting recommendation tasks, and that a further generalisation of our model to allow for contextual information would be necessary. This is outside the scope of the current work.

In this work we give lower and upper bounds on the minimal regret attainable by any strategy in terms of the number of rounds, n, the number of arms, k, and the dimension of the Hilbert space, d, for different sets of environments and actions. We discuss our results

in detail in Section 3 and give the proofs in Sections 5, 6 and 7. For the upper bounds in Section 7 we use an algorithmic approach, i.e, we try to find an algorithm that has low regret for all states and perform the regret analysis. Here our main technique is to apply known classical algorithms using a vector representation of our problem (see Lemma 16). For the lower bounds in Sections 5 and 6 we use information theoretical techniques, focusing on finding environments for which all strategies perform "poorly".

It is worth stressing here that our model falls within the class of (classical) multi-armed stochastic linear bandits. This is not surprising: most problems in quantum tomography, metrology or learning can be converted to a classical problem with additional structure, since, in the end, we are interested in learning a matrix of complex numbers constituting the density operator. The question of interest then is whether the structure imposed by the quantum problem is helpful to find more efficient algorithms. Indeed, the fact that our model is a subset of a more general class of problems for which classical algorithms are known does not imply that these algorithms yield the smallest possible regret for the subset of problems we consider. In fact, our results show that the classical algorithms are optimal in most cases, although we identify at least one setting (learning pure states using the set of all rank-1 projectors) where we expect bespoke algorithms to perform better. Moreover, while inspired by lower bounds on the classical multi-armed stochastic bandit problem, our bounds require novel constructions that are specific to the quantum state space. On the one hand, due to correlations in the reward distribution associated to each arm (observable), the standard multi-armed stochastic bandit lower bounds proofs do not apply to our case. On the other hand, lower bounds for linear stochastic bandits have been studied only for specific action sets like the hypercube or unit sphere (see [14] and [15]) and there is no combination of action sets and environments in the classical proofs that can be mapped to our particular class of problems. For that reason, known classical regret lower bounds do not apply to our case.

Beyond bandits, there are other reinforcement learning frameworks where one can model more complex environments where actions have long-term consequences. One example are Markov decision processes (MDPs) where the learner interacts with the environment by choosing an action, receiving a reward, and observing a state (or partial information). The reward received by the learner not only depends on the action, it also depends on the state and this state also evolves depending on the action taken by the learner. MDPs have been generalized to the quantum setting (see [16, 17] where the underlying states are quantum states and the evolution and rewards are generated following quantum processes. Our model fits as a restricted version of [17] since actions do not prompt state transformations in our model. The model presented in [17] also involves measurements providing additional partial information. However, our model, being more specific, allows us to reach more detailed results including lower and upper bounds for the cumulative regret.

Finally, we note that one can try to apply techniques from quantum state tomography [18] or shadow tomography [19] in this setting using the observables of the action set in order to estimate the unknown quantum state. If we learn the unknown quantum state (or some approximation thereof) we can choose the best action in order to minimize the regret. However this strategy is not optimal: quantum state tomography algorithms can be thought of as pure exploration strategies since the algorithm only cares about choosing the action that helps us to learn most about the unknown quantum state, which is not necessarily the action that minimizes the regret. In particular, our setting is thus different from the setting of online learning of quantum states [20]. There, the learner is tasked to produce an estimate ω_t of ρ in each round and the regret is related to the quality of
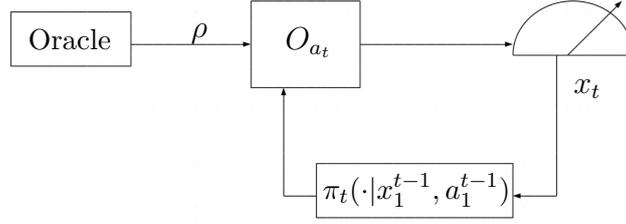
Figure 1: Scheme for the multi-armed quantum bandit setting.

this estimate. The main difference with our model is that we do not require our policy to produce an estimate of the unknown quantum state, just to choose an observable with large expectation value on the unknown quantum state.

2 Multi-armed quantum bandits

We assume familiarity with basic concepts and notations in probability theory and quantum information theory. Here we introduce notations needed to set up the model and present our results. Some more concepts will be introduced in the sections of proofs. We denote the natural logarithm as log. We define $[n] := \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$. We restrict our attention to finite-dimensional quantum systems. Let $\mathcal{S}_d = \{\rho \in \mathbb{C}^{d \times d} : \rho \geq 0 \wedge \mathrm{Tr}(\rho) = 1\}$ denote the set of positive semi-definite operators with unit trace, i.e *quantum states* that act on a d-dimensional Hilbert space $\mathbb{C}^d$. Pure states are rank-1 projectors in the set of states, given by $\mathcal{S}_d^* = \{\rho \in \mathcal{S}_d : \rho^2 = \rho\}$. Moreover, *observables* are Hermitian operators acting on $\mathbb{C}^d$, collected in the set $\mathcal{O}_d = \{O \in \mathbb{C}^{d \times d} : O^\dagger = O\}$. An observable O is called *traceless* if $\mathrm{Tr}(O) = 0$ and it is called *sub-normalised* if $\|O\| \leq 1$, where $\|\cdot\|$ denotes the operator norm. For two quantum states $\rho, \sigma \in \mathcal{S}_d$ we write their trace distance as $\frac{1}{2}\|\rho - \sigma\|_1$ where $\|X\|_1 = \mathrm{Tr}\,|X|$ is the Schatten 1-norm.

We will consider bandits with a finite set of actions, which we call discrete bandits, and bandits with a general, potentially continuous, set of actions.

2.1 Discrete bandits

Definition 1 (Multi-armed quantum bandit)**.** Let $d \in \mathbb{N}$. A d-dimensional *discrete multi-armed quantum bandit* is given by a finite set $\mathcal{A} \subseteq \mathcal{O}_d$ of observables that we call the *action set*. The bandit is in an *environment*, a quantum state $\rho \in \Gamma$, that is unknown but taken from a set of *potential environments* $\Gamma \subseteq \mathcal{S}_d$. The bandit problem is fully characterized by the tuple $(\mathcal{A}, \Gamma)$.

As potential environments in this work we consider either general states or only pure states. We will consider discrete bandits with (almost) arbitrary action sets and bandits with structured action sets, for example given by Pauli measurements.

Let us now fix a bandit with unknown state ρ and action set $\mathcal{A} = \{O_1, O_2, \ldots, O_k\}$ of cardinality $k = |\mathcal{A}|$. For each observable indexed by $a \in [k]$, we also introduce its spectral decomposition,

$$O_a = \sum_{i=1}^{d_a} \lambda_{a,i} \Pi_{a,i}, \tag{1}$$

where $\lambda_{a,i} \in \mathbb{R}$ denote the $d_a \leq d$ distinct eigenvalues of O_a and $\Pi_{a,i}$ are the orthogonal projectors on the respective eigenspaces. We also introduce the spectrum $\Lambda_a = \{\lambda_{a,1}, \ldots, \lambda_{a,d_a}\}$ of the observable O_a.

With this in hand we can now describe the learning process (see also Figure 1). In each round $t \in \mathbb{N}$, the learner (probabilistically or deterministically) chooses an action $A_t \in [k]$ and measures the observable O_{A_t} on a copy of the unknown quantum state ρ. The result of the measurement is the reward $X_t \in \Lambda_{A_t}$. More specifically, given the spectral decompositions in (1), the conditional probability distribution of the reward X_t is given by Born's rule:

$$\Pr[X_t = x | A_t = a] = P_\rho(x|a) = \begin{cases} \mathrm{Tr}(\rho \Pi_{a,i}) & \text{if } x = \lambda_{a,i} \\ 0 & \text{else} \end{cases}. \tag{2}$$

for any $a \in [k]$. The probability is thus nonzero only if $x \in \Lambda_a$. Note in particular that the conditional expectation of X_t is given by $\mathbb{E}_\rho[X_t|A_t] = \mathrm{Tr}(\rho O_{A_t})$. The learner's choice of action is determined by a policy.

Definition 2. A *policy* (or algorithm) for a multi-armed quantum bandit is a set of (conditional) probability distributions $\pi = \{\pi_t\}_{t \in \mathbb{N}}$ on the action index set $[k]$ of the form

$$\pi_t(\cdot|a_1, x_1, \ldots, a_{t-1}, x_{t-1}), \tag{3}$$

defined for all valid combinations of actions and rewards $(a_1, x_1), \ldots, (a_{t-1}, x_{t-1})$ up to time $t - 1$.

Then, if we run the policy π on the state ρ over $n \in \mathbb{N}$ rounds, we can define a joint probability distribution over the set of actions and rewards as

$$P_{\rho,\pi}(a_1, x_1, \ldots, a_n, x_n) := \prod_{t=1}^{n} \pi_t(a_t|a_1, x_1, \ldots, a_{t-1}, x_{t-1}) P_\rho(x_t|a_t). \tag{4}$$

This distribution fully describes the random learning process.

2.2 Regret

The objective of the learner is to find the action (observable) that in expectation maximizes the reward. This is equivalent to minimizing the expected cumulative regret.

Definition 3. Given a multi-armed quantum bandit problem $(\mathcal{A}, \Gamma)$, a state $\rho \in \Gamma$, a policy π and $n \in \mathbb{N}$, we define the *expected cumulative regret* as

$$R_n(\mathcal{A}, \rho, \pi) := \sum_{t=1}^{n} \max_{O \in \mathcal{A}} \mathrm{Tr}(\rho O) - \mathbb{E}_{\rho,\pi}[X_t], \tag{5}$$

where the expectation value is taken with respect to the probability density (4).

We note that $\mathbb{E}_{\rho,\pi}[X_t] = \mathrm{Tr}(\rho O_{A_t})$. We define the *sub-optimality gap* as,

$$\Delta_a := \max_{O \in \mathcal{A}} \mathrm{Tr}(\rho O) - \mathrm{Tr}(\rho O_a), \tag{6}$$

for $a \in [k]$. The sub-optimality gap represents the relative loss between the optimal measurement and the measurement induced by $O_a \in \mathcal{A}$. Note that $\Delta_a \geq 0$ and the equality is only achieved by the optimal observables (it does not need to be unique). In

order to analyse the expected cumulative regret it is often more convenient to work with the equivalent expression given by

$$R_n(\mathcal{A}, \rho, \pi) = \sum_{a=1}^{k} \sum_{t=1}^{n} \mathbb{E}_{\rho,\pi}[(\max_{O \in \mathcal{A}} \mathrm{Tr}(\rho O) - X_t)\mathbb{I}\{A_t = a\}] \tag{7}$$

$$= \sum_{a=1}^{k} \sum_{t=1}^{n} \Delta_a \, \mathbb{E}_{\rho,\pi}[\mathbb{I}\{A_t = a\}] \tag{8}$$

$$= \sum_{a=1}^{k} \Delta_a \, \mathbb{E}_{\rho,\pi}[T_a(n)], \tag{9}$$

where $T_a(n) = \sum_{t=1}^{n} \mathbb{I}\{A_t = a\}$ is a random variable that tells us how many times we have picked the action a over n rounds.

For a fixed policy π the cumulative expected regret depends on ρ. Thus, in order to determine how well a policy performs for a given set Γ of environments we need to find a state-independent metrics.

Definition 4. Given a multi-armed quantum bandit problem $(\mathcal{A}, \Gamma)$, a policy π and $n \in \mathbb{N}$, we define the *worst-case regret* as

$$R_n(\mathcal{A}, \Gamma, \pi) = \sup_{\rho \in \Gamma} R_n(\mathcal{A}, \rho, \pi). \tag{10}$$

Moreover, the *minimax regret* is defined as,

$$R_n(\mathcal{A}, \Gamma) = \inf_{\pi} R_n(\mathcal{A}, \Gamma, \pi), \tag{11}$$

where the infimum goes over all possible policies of the form in Definition 2.

The minimax regret is a measure of how difficult the multi-armed quantum bandit problem is and we will attempt to find upper and lower bounds on $R_n(\mathcal{A}, \Gamma)$ as a function of n and the Hilbert space dimension. A small value of $R_n(\mathcal{A}, \Gamma)$ means that the problem is less difficult to learn.

2.3 General bandits

Often it is natural to allow a continuous set of possible observables, for example all rank-1 projectors. This leads us to a more general definition of multi-armed quantum bandits that works for any measurable space of actions.

We will only mention the changes in the formalism in comparison to the discrete case, and build on the notions introduced above.

Definition 5. Let $d \in \mathbb{N}$. A general d-dimensional *multi-armed quantum bandit* is given by a measurable space $(\mathcal{A}, \Sigma)$, where Σ is a σ-algebra of subsets of $\mathcal{A}$. A *policy* for such a bandit is given by conditional probability measures $\pi_t(\cdot|a_1, x_1, \ldots, a_{t-1}, x_{t-1}) : \Sigma \to [0, 1]$.

The reward distribution $P_\rho(x|a)$ conditioned on playing arm $a \in \mathcal{A}$ when the environment state is ρ is still given by Born's rule in (2) and, importantly, remains discrete. To simplify our presentation we in the following also assume that the spectra of the observables satisfy $\Lambda_a \subset \mathcal{X}$ for some finite set $\mathcal{X}$, i.e., we only allow for a discrete set of possible rewards independently of the choice of action. This is for example trivially satisfied for the case of rank-1 projectors, where $\mathcal{X} = \{0, 1\}$ is trivial. We also need to assume that the function $a \mapsto P_\rho(x|a)$ is Σ-measurable for all environment states ρ and $x \in \mathcal{X}$.

The environment state ρ and the policy π define the probability measure $P_{\rho,\pi} : (\Sigma \times \mathcal{X})^{\times n} \to [0,1]$ through

$$P_{\rho,\pi}(A_1, x_1, \ldots, A_n, x_n)$$
$$= \int_{A_1} \cdots \int_{A_n} P_\rho(x_n|a_n)\pi_n(da_n|a_1, x_1, \ldots, a_{n-1}, x_{n-1}) \cdots P_\rho(x_1|a_1)\pi_1(da_1) \quad (12)$$

This map will be treated as a measure although it is not strictly speaking a set function. However, this should cause no confusion.

In order to present a formula for regret, we need the continuous counterpart for the function $a \mapsto \mathbb{E}_{\rho,\pi}(T_a(n))$ of the discrete case. To this end, we define, for all $t = 1, \ldots, n$, the margins $P_{\rho,\pi}^{(t)} : \Sigma \times \mathcal{X} \to [0,1]$ through

$$P_{\rho,\pi}^{(t)}(A, x) = \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_{t-1} \in \mathcal{X}} \sum_{x_{t+1} \in \mathcal{X}} \cdots \sum_{x_n \in \mathcal{X}} P_{\rho,\pi}(\mathcal{A}, x_1, \ldots, \mathcal{A}, x_{t-1}, A, x, \mathcal{A}, x_{t+1}, \ldots, \mathcal{A}, x_n).$$
$$(13)$$

We now define the measure $\gamma_{\rho,\pi} : \Sigma \to [0,n]$ through

$$\gamma_{\rho,\pi}(A) = \sum_{t=1}^n \sum_{x \in \mathcal{X}} P_{\rho,\pi}^{(t)}(A, x). \quad (14)$$

We denote by $E_{\rho|a}$ the expectation value of the conditional distribution $x \mapsto P_\rho(x|a)$ and by Δ_a the supremum of the function $b \mapsto E_{\rho|b} - E_{\rho|a}$ for any arm $a \in \mathcal{A}$; this is naturally the continuous analogue of the sub-optimality gap. Note that, in all the cases we study, $\mathcal{A}$ is a compact topological space, Σ is the associated Borel σ-algebra, and $a \mapsto E_{\rho|a}$ is continuous for every environment ρ, so this supremum (which is also a maximum) technically makes sense. Using the measure $\gamma_{\rho,\pi}$ and the above sub-optimality gap, we may define the expected cumulative regret analogously to the discrete case:

$$R_n(\rho, \pi, \mathcal{A}) := \sum_{t=1}^n \mathbb{E}_{\rho,\pi}(\Delta_{a_t}) \quad (15)$$

$$= \sum_{t=1}^n \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_n \in \mathcal{X}} \int_{\mathcal{A}^n} \Delta_{a_t} P_{\rho,\pi}(da_1, x_1, \ldots, da_n, x_n) \quad (16)$$

$$= \int_{\mathcal{A}} \Delta_a \, d\gamma_{\rho,\pi}(a). \quad (17)$$

Note that Equation (15) implies that, for the optimal strategy which always plays the best arm, the regret vanishes. The worst-case and minimax regrets are defined in the obvious way.

A generalisation of these definitions to environments comprised of states on infinite-dimensional Hilbert spaces is straight-forward, but outside the scope of this work. Especially, in the infinite-dimensional setting, it is natural that the reward set is also continuous.

3 Main results and discussion

We present our main results of lower and upper bounds on the minimax regret for different environments and action sets. The results are summarized in Table 1.

	Discrete		Continuous
	Arm-limited	Dimension-limited	(all rank-1 projectors)
$\Gamma = \mathcal{S}_d$	$R_n(\mathcal{A}, \Gamma) = \Omega(\sqrt{kn})\ (k < d^2)$ Result 2	$R_n(\mathcal{A}, \Gamma) = \Omega\left(d\sqrt{n}\right)$ Result 2	$R_n(\mathcal{A}, \Gamma) = \Omega\left(\sqrt{n}\right)$ Result 4
	$R_n(\mathcal{A}, \Gamma) = \widetilde{O}(\sqrt{kn})$ Result 6	$R_n(\mathcal{A}, \Gamma) = \widetilde{O}\left(d\sqrt{n}\right)$ Result 6	$R_n(\mathcal{A}, \Gamma) = \widetilde{O}(d^2\sqrt{n})$ Result 5 $R_n(\mathcal{A}, \Gamma) = \widetilde{O}(\sqrt{n}),\,(d = 2)$ Result 7
$\Gamma = \mathcal{S}_d^*$	$R_n(\mathcal{A}, \Gamma) = \Omega\left(\sqrt{n}\right)\ (d = 2, k = 3)$ Result 3		$R_n(\mathcal{A}, \Gamma) =?$

Table 1: Scaling of the minimax regret in terms of the number of rounds, n, dimension of the Hilbert space, d, and number of actions, k, for discrete bandits. We study general state environments, $\mathcal{S}_d$, and pure-states environments, $\mathcal{S}_d^*$. For discrete actions sets we differentiate between arm-limited (the number of arms is smaller than the degrees of freedom in the quantum state space) and dimension-limited (the number of arms can be arbitrary large).

3.1 Lower bounds

We start giving the lower bounds results, see Section 5 and Section 6 for the detailed statements and proofs. We first give a generic lower bound that works for (almost) all discrete models.

Result 1. *Consider a discrete multi-armed quantum bandit problem $(\mathcal{A}, \Gamma = \mathcal{S}_d)$ for some $d \geq 2$ where the action set $\mathcal{A}$ is comprised of traceless observables. Then, the minimax regret satisfies*

$$R_n(\mathcal{A}, \mathcal{S}_d) = \Omega\left(\sqrt{n}\right) \tag{18}$$

under some mild regularity conditions.

This is shown as Theorem 10 in Section 5.1. This result shows the dependence of the lower bound on the number of actions for a large class of action sets with mixed-state environments. The regularity condition serves to exclude sets where there is a dominant operator that performs optimally for all states, thus giving a trivial bound on the minimax regret. We also demand the observables to be traceless to aid the analysis.

For the special case of environments of one qubit quantum states and set of actions consisting of rank-1 projectors we give an exact form of the minimax regret lower bound. In this case, the minimax regret can be bounded as

$$R_n(\mathcal{A}, \mathcal{S}_2) \geq \frac{\sqrt{1-c} - (1-c)}{30}\sqrt{n}, \tag{19}$$

where $c = \mathrm{Tr}(\Pi_a \Pi_b)$ for the two observables that minimize the overlap $\max\{\mathrm{Tr}(\Pi_i \Pi_j), \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right)\}$. This is proved as Theorem 12 in Section 5.2.

Next we want to study the dependence of the lower bound on the number of available actions and the dimension for suitable action sets. For this purpose we consider action sets comprised of strings of Pauli observables (strings of single-qubit Pauli operators, excluding the overall identity).

Result 2. *Consider a discrete multi-armed quantum bandit problem $(\mathcal{A}, \Gamma = \mathcal{S}_d)$ for $d = 2^\ell$ and $\ell \in \mathbb{N}$, where the action set $\mathcal{A}$ is comprised of k distinct length-l strings of single-qubit Pauli operators. Then the minimax regret can be bounded as*

$$R_n(\mathcal{A}, \mathcal{S}_d) = \Omega\left(\sqrt{(k-1)n}\right). \tag{20}$$

Note in particular that we can have $d^2 - 1$ such strings; thus, we get a lower bound of $R_n(\mathcal{A}, \mathcal{S}_d) = \Omega(d\sqrt{n})$ for the most difficult of this type of quantum bandit problems. This result is stated as Theorem 13 in Section 5.3. As we will see when we discuss the generic algorithm, this dependence is optimal.

We also study the specific case of one qubit rank-1 environments and show that even under these restricted environments the scaling of the regret with the number of rounds is the same.

Result 3. *Consider a discrete multi-armed quantum bandit problem $(\mathcal{A}, \Gamma = \mathcal{S}_2^*)$, where the action set $\mathcal{A}$ is comprised of the one qubit Pauli observables. Then the minimax regret can be bounded as*

$$R_n(\mathcal{A}, \mathcal{S}_2^*) = \Omega\left(\sqrt{n}\right). \tag{21}$$

This result is stated as Theorem 14 in Section 5.4.

Now we move to general bandits with continuous action sets. In this case it is actually much more difficult to show lower bounds. One reason is that the sub-optimality gap is now not bounded away from zero when we choose a sub-optimal observable. We show that the scaling in the number of rounds is still the same also in this case.

Result 4. *Consider a general multi-armed quantum bandit problem $(\mathcal{A}, \Gamma = \mathcal{S}_d)$ where the action set $\mathcal{A}$ consists of all rank-1 projectors in d dimensions. Then, the minimax regret can be bounded as,*

$$R_n(\mathcal{A}, \mathcal{S}_d) = \Omega(\sqrt{n}). \tag{22}$$

This result is shown in Theorem 15 in Section 6.1. We would expect that the minimax regret necessarily also scales with the dimension, but we currently cannot show this. (See the discussion in Sections 6.1 and 8.)

3.2 Upper bounds

In order to provide upper bounds for the minimax regret we explore algorithms and analyse the regret. For the general case we adapt one of the main algorithms for linear stochastic bandits called LinUCB to our quantum setting, which leads to the following result.

Result 5. *For any general multi-armed quantum bandit problem $(\mathcal{A}, \Gamma)$ with $\Gamma = \mathcal{S}_d$ or $\Gamma = \mathcal{S}_d^*$, there exists a policy π such that the worst-case regret can be bounded as*

$$R_n(\mathcal{A}, \Gamma, \pi) = O\left(d^2\sqrt{n}\log n\right). \tag{23}$$

We give all the details in Section 7.1. Note that the scaling in terms of the numbers of rounds n is quasi-optimal since it matches the previous lower bound up to logarithmic terms. However there is a gap in the dimensional dependence in comparison to the lower bound $\Omega(\sqrt{n})$ of Results 1 and 4 or the lower bound $\Omega(d\sqrt{n})$ that we established for Pauli observables. The gap for Pauli observables can be closed if we consider a more generic algorithm called UCB that can be applied to any multi-armed stochastic bandit with a discrete number of arms, and thus can also treat linear bandits.

Result 6. *For any discrete multi-armed quantum bandit problem $(\mathcal{A}, \Gamma)$ with $\Gamma = \mathcal{S}_d$ or $\Gamma = \mathcal{S}_d^*$ such that $k = |\mathcal{A}|$, there exist policies π_1, π_2 such that the worst-case regret can be bounded as*

$$R_n(\mathcal{A}, \mathcal{S}_d, \pi_1) = O\left(\sqrt{kn}\right), \quad R_n(\mathcal{A}, \mathcal{S}_d, \pi_2) = O\left(d\sqrt{n \log(nk)}\right). \tag{24}$$

The details for the above results can be found in Section 7.2.

Finally, we consider potential environments that contain all pure states in $\mathcal{S}_d^*$, again with an action set $\mathcal{A}$ containing all rank-1 projectors. For this case, the cumulative regret can be expressed as, $R_n(\mathcal{A}, \rho, \pi) = \sum_{t=1}^{n} \mathbb{E}_{\rho, \pi}(\frac{1}{2}\|\rho - \Pi_{A_t}\|_1)^2$ where $\rho = |\psi\rangle\langle\psi|$ is the unknown pure quantum state and $\Pi_{A_t} \in \mathcal{A}$ is a rank-1 projector selected at time step t. This regret now has a resemblance with the regret in online learning of quantum states since it is a natural loss function for our estimate Π_{A_t} of the state ρ. However, the important difference is that in online learning the state estimate is provided separately from the choice of the next measurement, which is often chosen at random.

We are currently unable to provide a non-trivial lower bound on $R_n(\mathcal{A}, \mathcal{S}_d^*)$, and pose this as an open question. To get an upper bound we propose a policy that is based on a explore-then-commit strategy (see [1][Chapter 6]) where the exploration is done using a quantum state tomography algorithm. Using this policy we can show the following result.

Result 7. *For any general multi-armed quantum bandit problem $(\mathcal{A}, \Gamma = \mathcal{S}_2^*)$ there exists a policy π such that the worst-case regret can be bounded as*

$$R_n(\mathcal{A}, \Gamma, \pi) = O\left(\sqrt{n} \log n\right). \tag{25}$$

We will see in Section 7.3 that this policy is "simple" in the sense that it is not very adaptive, suggesting room for improvements. Nevertheless it achieves the same type of scaling in the number of rounds that Result 5.

4 Bretagnolle-Huber inequality and divergence decomposition lemma

We will need to introduce some additional notation. For two probability measures P, Q defined on the same probability space (Ω, Σ) we may introduce a (probability) measure μ that is dominating P and Q and define the Radon–Nikodym derivatives $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. The choice of μ is arbitrary for the following definitions. We define the Kullback–Leibler divergence as

$$D(P\|Q) := \int p(\log p - \log q)\, d\mu \tag{26}$$

where we use the convention that $0 \log 0 = 0$ and $D(P\|Q) = \infty$ whenever Q does not dominate P. We also define the (squared) Bhattacharyya coefficient as

$$F(P, Q) = \left(\int \sqrt{pq}\, d\mu\right)^2. \tag{27}$$

For two quantum states $\rho, \sigma \in \mathcal{S}_d$ we denote their quantum relative entropy as $D(\rho\|\sigma) = \mathrm{Tr}(\rho \log \rho) - \mathrm{Tr}(\rho \log \sigma)$ if $\mathrm{supp}(\rho) \subseteq \mathrm{supp}(\sigma)$ and $D(\rho\|\sigma) = \infty$ otherwise. Their fidelity is defined as $F(\rho, \sigma) = (\mathrm{Tr}|\sqrt{\rho}\sqrt{\sigma}|)^2$.

In this section we present some important technical results which we need in the subsequent proofs in order to prove the lower bounds summed up in the preceding section. One of the technical tools that we will use in order to bound the regret is the following lemma due to Bertagnolle and Huber [21].

Lemma 6 (Bretagnolle–Huber inequality). *Let P and Q be probability measures on the same measurable space (Ω, Σ), and let $A \in \Sigma$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D(P\|Q)), \tag{28}$$

where $A^c = \Omega \backslash A$ is the complement of A and $D(P\|Q)$ is the Kullback–Leibler divergence.

We will use also a stronger bound that replaces $D(P\|Q)$ with the Rényi divergence of order $\alpha = \frac{1}{2}$, defined as

$$D_{\frac{1}{2}}(P\|Q) = -\log F(P, Q). \tag{29}$$

We use the proof given in [1] for the Bretagnolle-Huber inequality in order to prove the alternative version.

Lemma 7. *Let P and Q be probability measures on the same measurable space (Ω, Σ), and let $A \in \Sigma$ be an arbitrary event with A^c its complement. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D_{\frac{1}{2}}(P\|Q)) = \frac{1}{2} F(P, Q). \tag{30}$$

Proof. Let us define the probability measure $\mu := \frac{1}{2}(P + Q)$ which dominates both P and Q and denote $dP/d\mu =: p$ and $dQ/d\mu =: q$. Note that,

$$P(A) + Q(A^c) = \int_A p \, d\mu + \int_{A^C} q \, d\mu \geq \int_A \min\{p, q\} \, d\mu + \int_{A^C} \min\{p, q\} \, d\mu \tag{31}$$

$$= \int \min\{p, q\} \, d\mu. \tag{32}$$

In order to find a lower bound on this expression we use the Cauchy–Schwarz inequality to show that

$$\frac{1}{2} \left(\int \sqrt{pq} \, d\mu \right)^2 = \frac{1}{2} \left(\int \sqrt{\max\{p, q\} \min\{p, q\}} \, d\mu \right)^2 \tag{33}$$

$$\leq \frac{1}{2} \left(\int \max\{p, q\} \, d\mu \right) \left(\int \min\{p, q\} \, d\mu \right) \leq \int \min\{p, q\} \, d\mu, \tag{34}$$

where in the final step we used that $\int \max\{p, q\} d\mu \leq \int (p + q) d\mu = 2$. $\square$

The other main result that we will need allows us to decompose the divergence computed for two joint distribution that result from the same policy applied to two different quantum states.

Lemma 8 (Divergence decomposition lemma). *Let $\mathcal{A} = \{O_1, ..., O_k\}$ be a set of actions and ρ and ρ' two quantum states defining two multi-armed quantum bandits with action set $\mathcal{A}$. Fix some policy π and let $P_{\rho,\pi}$ and $P_{\rho'\pi}$ be the probability distributions induced by the n-round interconnection of π and ρ described in (4). Then,*

$$D(P_{\rho,\pi} \| P_{\rho',\pi}) = \sum_{a=1}^{k} \mathbb{E}_{\rho,\pi}[T_a(n)] D(P_\rho(\cdot|a) \| P_{\rho'}(\cdot|a)). \tag{35}$$

The above lemma and the proof can be found in [1, Chapter 15] for the classical model of multi-armed stochastic bandits. We have restated the lemma for our quantum case but this statement and the proof follows trivially from the original one. The proof is

a consequence of the chain rule for the KL divergence. Unfortunately, there is no such decomposition for Rényi divergences but we will see in Section 5 that applying the data processing inequality and bounding it with the sandwiched quantum Rényi divergences (see, e.g., [22, 23]) between ρ and ρ' is sufficient to bound the regret.

For this purpose, we present result on the divergences between the probability distributions $P_{\rho,\pi}$. Note that this result also holds in the general bandit case where the set of arms can be continuous. What we mean below by 'quantum extension' of a classical relative entropy D is that, if quantum states ρ and σ commute, then $\tilde{D}(\rho\|\sigma) = D(p\|q)$ where p and, respectively, q are the vectors of eigenvalues of ρ and, respectively, σ, in a common eigenbasis.

Lemma 9. *Let $\alpha \in \mathbb{R}$ be such that the classical Rényi relative entropy D_α can be given a quantum extension (which we denote with the same symbol) which is additive and satisfies the data processing inequality. For any policy π and environment states ρ, ρ', we have*

$$D_\alpha(P_{\rho,\pi}\|P_{\rho',\pi}) \leq nD_\alpha(\rho\|\rho'). \tag{36}$$

In particular, for $\alpha = 1/2$ we can choose $D_{\frac{1}{2}}(\rho\|\rho') = -\log F(\rho, \rho')$ and, for $\alpha = 1$, we can let the quantum extension $D \equiv D_1$ of the Kullback-Leibler relative entropy be the quantum relative entropy.

Proof. Let us fix the policy π and use the notations and definitions of Section 2.3. We prove the claim by constructing a positive-operator-valued measure (POVM) E over the value space $\mathcal{C} := (\Sigma \times \mathcal{X})^n$ and operating in $\mathcal{H}^{\otimes n}$ such that

$$P_{\sigma,\pi}(A_1, x_1, \ldots, A_n, x_n) = \mathrm{Tr}\left(\sigma^{\otimes n}\mathsf{E}(A_1, x_1, \ldots, A_n, x_n)\right) \tag{37}$$

for all states σ, and $A_t \in \Sigma$ and $x|t \in \mathcal{X}$ for $t = 1, \ldots, n$. Note that, as E is not a set function, it is, strictly speaking, not a POVM in the same sense as $P_{\sigma,\pi}$ is not a measure. However, this should cause no confusion. Using the data-processing inequality and the additivity of D_α, we have, for all states ρ and ρ',

$$D_\alpha(P_{\rho,\pi}\|P_{\rho',\pi}) \leq D_\alpha(\rho^{\otimes n}\|\rho'^{\otimes n}) = nD_\alpha(\rho\|\rho'), \tag{38}$$

implying Inequality (36). (Note that the POVM E corresponds to a quantum-to-classical channel which maps $\sigma^{\otimes n}$ into $P_{\sigma,\pi}$). Thus all that remains is to write down E.

Recall the reward distributions $P_\sigma(x|a)$ in state σ conditioned by the arm $a \in \mathcal{A}$. Through linear extension, we may define these distributions $P_\tau(x|a)$ for any (trace-class) operators τ; these are, in general, complex distributions. We may define the POVM E through

$$\mathrm{Tr}\left((\tau_1 \otimes \cdots \otimes \tau_n)\mathsf{E}(A_1, x_1, \ldots, A_n, x_n)\right)$$
$$= \int_{A_1} \cdots \int_{A_n} \prod_{t=1}^n P_{\tau_t}(x_t|a_t)\pi_t(da_t|a_1, x_1, \ldots, a_{t-1}, x_{t-1}) \tag{39}$$

for all (trace-class) operators τ_t and $A_t \in \mathcal{B}$ and $x_t \in \mathcal{X}$, $t = 1, \ldots, n$. As $\tau \mapsto P_\tau(x|a)$ is a linear functional and $x \mapsto P_\sigma(x|a)$ is a probability distribution whenever σ is a state, it easily follows that there are POVMs P_a on $\mathcal{X}$ operating in $\mathcal{H}$ such that $P_\tau(x|a) = \mathrm{Tr}\left(\tau\mathsf{P}_a(x)\right)$. Equivalently (although slightly less formally) we may now write E in the differential form

$$\mathsf{E}(da_1, x_1, \ldots, da_n, x_n) = \bigotimes_{t=1}^n \pi_t(da_t|a_1, x_1, \ldots, a_{t-1}, x_{t-1})\mathsf{P}_{a_t}(x_t). \tag{40}$$

By substituting $\tau_t = \sigma$ for all $t = 1, \ldots, n$ in Equation (39), we immediately see that $P_{\sigma,\pi}(A_1, x_1 \ldots, A_n, x_n) = \operatorname{Tr}(\sigma^{\otimes n}\mathsf{E}(A_1, x_1 \ldots, A_n, x_n))$ for all $A_t \in \Sigma$ and $x_t \in \mathcal{X}$, $t = 1, \ldots, n$, and, thus, Equation (37) holds. $\qquad\square$

5 Regret lower bounds for discrete bandits

In this section we will focus on proving mini-max regret lower bounds for different sets of environments using information-theoretic tools.

5.1 General case

The first case that we study are quantum multi-armed bandits where the environment can be any state in a Hilbert space of dimension d, and the action set any set of observables. We will give a lower bound for the minimax regret, and in order to do it we construct a quantum state such that for every policy it achieves a non-trivial lower bound. Note that if we do not impose any condition on the action set the lower bound will vanish. The reason is that if there is an operator in the action set that dominates over the others, the policy that always chooses this operator at each round achieves 0 regret. More specifically, suppose that $\mathcal{A} = \{O_1, O_2\}$ with $O_1 \geq O_2$. Then we know that independently of the environment ρ, $\operatorname{Tr}(\rho O_1) \geq \operatorname{Tr}(\rho O_2)$ and the policy that always chooses O_1 will always pick the optimal action.

So, we will impose a condition on the action set that ensures that there is no such dominant action/operator. The condition is the following: in the action set $\mathcal{A}$ there exist at least two operators $O_a, O_b \in \mathcal{A}$ with maximal eigenvectors $|\psi_A\rangle, |\psi_B\rangle$ such that for any $i \neq a$ and $j \neq b$

$$\langle\psi_A| O_a |\psi_A\rangle > \langle\psi_A| O_i |\psi_A\rangle \quad \text{and} \quad \langle\psi_B| O_b |\psi_B\rangle > \langle\psi_B| O_j |\psi_B\rangle. \tag{41}$$

Theorem 10. *Let $n \in \mathbb{N}$. For any policy π and action set of traceless observables $\mathcal{A}$ that obeys condition (41) there exists an environment $\rho \in \mathcal{S}_d$ such that,*

$$R_n(\mathcal{A}, \rho, \pi) \geq C_{\mathcal{A}}\sqrt{n}, \tag{42}$$

where $C_{\mathcal{A}} > 0$ is a constant that depends on the action set.

Proof. Choose two operators $O_a, O_b \in \mathcal{A}$ that obey the condition in (41) with maximal eigenvectors $|\psi_A\rangle$ and $|\psi_B\rangle$, respectively. Define the following environments:

$$\rho := \frac{1-\Delta}{d}I + \Delta |\psi_A\rangle\langle\psi_A|, \quad \rho' := \frac{1-\Delta}{d}I + \Delta |\psi_B\rangle\langle\psi_B|, \tag{43}$$

for some constant $0 \leq \Delta \leq \frac{1}{2}$ to be defined later. Note that ρ and $\rho' \geq 0$. Define

$$c = \min\{\langle\psi_A| O_a |\psi_A\rangle - \langle\psi_A| O_i |\psi_A\rangle, \langle\psi_B| O_b |\psi_B\rangle - \langle\psi_B| O_j |\psi_B\rangle\}, \tag{44}$$

for $i \neq a$ and $j \neq b$. Using the expression for the regret in Equation (7), we can compute the regret for ρ as,

$$R_n(\mathcal{A}, \rho, \pi) = \sum_{i=1}^{k} \mathbb{E}_{\rho,\pi}(T_i(n))\Delta\left(\langle\psi_A| O_a |\psi_A\rangle - \langle\psi_A| O_i |\psi_A\rangle\right) \tag{45}$$

where we have used that the observables are traceless and

$$\max_{O_i \in \mathcal{A}} \mathrm{Tr}(\rho O_i) = \mathrm{Tr}(\rho O_a) = \Delta \langle \psi_A | O_a | \psi_A \rangle. \tag{46}$$

Note that for $i = a$ the sub-optimality gaps Equation (6) are zero. Thus, using Condition (41) we can bound the regret as,

$$R_n(\mathcal{A}, \rho, \pi) = \sum_{i \neq a} \mathbb{E}_{\rho,\pi}[T_i(n)] \Delta \left(\langle \psi_A | O_a | \psi_A \rangle - \langle \psi_A | O_i | \psi_A \rangle \right) \geq \Delta c \sum_{i \neq a} \mathbb{E}_{\rho,\pi}[T_i(n)]. \tag{47}$$

Using that $n = \sum_{i=1}^{k} \mathbb{E}_{\rho,\pi}[T_i(n)]$ and Markov inequality we have,

$$R_n(\mathcal{A}, \rho, \pi) \geq \Delta c \, \mathbb{E}_{\rho,\pi}\left[n - T_a(n) \right] \geq \frac{cn\Delta}{2} P_{\rho,\pi}\left(T_a(n) \leq \frac{n}{2} \right). \tag{48}$$

Similarly, the regret for ρ' can be bounded as,

$$R_n(\mathcal{A}, \rho', \pi) \geq \Delta \, \mathbb{E}_{\rho',\pi}[T_a(n)] \left(\langle \psi_B | O_b | \psi_B \rangle - \langle \psi_B | O_i | \psi_B \rangle \right), \tag{49}$$

where we have taken into account just the term with $i = a$. Using Condition (41) and Markov inequality we have,

$$R_n(\mathcal{A}, \rho', \pi) \geq \frac{cn\Delta}{2} P_{\rho',\pi}\left(T_a(n) > \frac{n}{2} \right). \tag{50}$$

Thus, combining Equations (48) and (50),

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \frac{cn\Delta}{2} \left(P_{\rho,\pi}\left(T_a(n) \leq \frac{n}{2} \right) + P_{\rho',\pi}\left(T_a(n) > \frac{n}{2} \right) \right). \tag{51}$$

Using Lemma 6 we can bound the above expression as,

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \frac{cn\Delta}{4} \exp\left(-D(P_{\rho,\pi} \| P_{\rho',\pi}) \right). \tag{52}$$

Using Lemma 8 combined with data-processing inequality we can bound the Kullback-divergence as,

$$D(P_{\rho,\pi} \| P_{\rho',\pi}) = \sum_{a=1}^{k} \mathbb{E}_{\rho,\pi}(T_a(n)) D(P_\rho(\cdot|a) \| P_{\rho'}(\cdot|a)) \leq D(\rho\|\rho') \sum_{a=1}^{k} \mathbb{E}_{\rho',\pi}[T_a(n)]$$
$$= nD(\rho\|\rho'). \tag{53}$$

where $D(\rho\|\rho') = \mathrm{Tr}\, \rho \log \rho - \mathrm{Tr}\, \rho \log \rho'$ is the relative entropy between ρ and ρ'. Now define $f(\Delta) = D(\rho\|\rho')$ for the corresponding expressions of ρ and ρ' and note that $f(0) = 0$ since for $\Delta = 0$, $\rho = \rho'$ and $f'(0) = 0$ using the convexity of the quantum relative entropy. Thus, using Taylor's theorem we can express $f(\Delta)$ as,

$$f(\Delta) = \frac{\Delta^2}{2} f''(\chi), \tag{54}$$

for some $\chi \in [0, \Delta]$. Define,

$$c_f = \max f''(\chi) \quad \text{for} \quad \chi \in [0, \tfrac{1}{2}]. \tag{55}$$

Note that $f''(\chi)$ is well defined since the states ρ, ρ' have full support and then $f(\chi)$ is a smooth function. Thus using that $f''(\chi) \le c_f$ for $\Delta \in [0, \frac{1}{2}]$ we can plug the above expression into Equation (53) and using Equation (52) we have,

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \ge \frac{cn\Delta}{4} \exp\left(-\frac{n\Delta^2}{2}c_f\right). \tag{56}$$

Finally, if we choose $\Delta = \frac{1}{2\sqrt{n}}$,

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \ge \frac{c}{8} \exp\left(-\frac{1}{8}c_f\right) \sqrt{n}, \tag{57}$$

and the result follows using $2 \max\{R_n(\mathcal{A}, \rho, \pi), R_n(\mathcal{A}, \rho', \pi)\} \ge R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi)$. $\quad\square$

5.2 Special case: one qubit and rank-1 projectors

Now we study the specific case where the environment belongs to the set of one-qubit quantum states and the set of actions is a set of rank-1 projectors, i.e $\mathcal{A} = \{\Pi_1, ..., \Pi_k\}$ where Π_i are rank-1 matrices such that $\Pi_i^2 = \Pi_i$. For this case we are able to give the exact constant of Theorem 10.

Given a quantum state ρ we will use the Bloch state representation,

$$\rho = \frac{I}{2} + \frac{1}{2}\mathbf{r} \cdot \sigma, \tag{58}$$

where $\mathbf{r} \in \mathbb{R}^3$ is the Bloch vector and $\sigma = (\sigma_x, \sigma_y, \sigma_z)$ is the Pauli vector. Recall that for mixed states $\|\mathbf{r}\|_2 < 1$ and for pure states $\|\mathbf{r}\|_2 = 1$. The projectors Π_i of the action set can be written as (58) for some unit vector $\mathbf{r}_i \in \mathbb{R}^3$. So, the action set is characterized by the set of measurement directions $\mathcal{A}_r = \{\mathbf{r}_1, ..., \mathbf{r}_k\}$ and the mean of each action is,

$$\text{Tr}(\rho\Pi_t) = \frac{1}{2} + \frac{1}{2}\mathbf{r} \cdot \mathbf{r}_t. \tag{59}$$

In order to prove the main result we will need the computation of the relative entropy between two quantum states given by the following Lemma.

Lemma 11. *Let $\rho = \frac{I}{2} + \frac{\Delta}{2}\sigma_x$ and $\rho' = \frac{I}{2} + \frac{\Delta}{2}\sigma_z$ be two one-qubit density matrices. Then, their quantum relative entropy can be computed as,*

$$D(\rho\|\rho') = \frac{\Delta}{2} \log\left(\frac{1+\Delta}{1-\Delta}\right). \tag{60}$$

Proof. Let Π_i^+, Π_i^- be the projectors for the $i = x, z$ Pauli matrix into the subspaces of eigenvalue $+1$ and -1 respectively. Then we can express the density matrices as,

$$\rho = \frac{1+\Delta}{2}\Pi_x^+ + \frac{1-\Delta}{2}\Pi_x^-, \quad \rho' = \frac{1+\Delta}{2}\Pi_z^+ + \frac{1-\Delta}{2}\Pi_z^-, \tag{61}$$

where we have used that $\sigma_i = \Pi_i^+ - \Pi_i^-$. Then using that $I = \Pi_i^+ + \Pi_i^-$, $\text{Tr}(\sigma_i) = 0$ and $\text{Tr}\,\sigma_x\sigma_z = 0$ we can compute the following identity

$$\text{Tr}(\Pi_i^+\Pi_j^-) = \begin{cases} \frac{1}{2} & \text{if } i \ne j \\ 0 & \text{if } i = j. \end{cases} \tag{62}$$

Then, using Equations (61) and (62),

$$\mathrm{Tr}\,\rho \log \rho = \frac{1+\Delta}{2} \log\left(\frac{1+\Delta}{2}\right) + \frac{1-\Delta}{2} \log\left(\frac{1-\Delta}{2}\right), \tag{63}$$

$$\mathrm{Tr}\,\rho \log \rho' = \frac{1}{2} \log\left(\frac{1+\Delta}{2}\right) + \frac{1}{2} \log\left(\frac{1-\Delta}{2}\right). \tag{64}$$

The result follows from the definition of the relative entropy and rearranging the above terms.

$\square$

Theorem 12. *Let $n \in \mathbb{N}$. For any policy π and finite action set of observables $\mathcal{A}$ of rank-1 projectors, there exists a 1-qubit environment $\rho \in \mathcal{S}_2$ such that,*

$$R_n(\mathcal{A}, \rho, \pi) \geq \frac{\sqrt{1-c} - (1-c)}{30} \sqrt{n}, \tag{65}$$

where $c = \mathrm{Tr}(\Pi_a \Pi_b)$ and $(\Pi_a, \Pi_b) = \mathrm{argmin}_{\Pi_i, \Pi_j \in \mathcal{A}} \max\left\{\mathrm{Tr}(\Pi_i \Pi_j), \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right)\right\}$.

Note that that $\sqrt{1-c} \geq 1 - c$ since $c \in [0,1]$, and the constant is thus positive and only vanishes for $c \in \{0, 1\}$.

Proof. Let $\mathcal{A}_r = \{\mathbf{r}_1, ..., \mathbf{r}_k\}$ be the set of vectors associated to the k elements of $\mathcal{A}$. Let $\Delta \in [0, \frac{1}{2}]$ be a constant to be chosen later. Pick $\mathbf{r}_a, \mathbf{r}_b \in \mathcal{A}_r$ such that they are the two directions with smallest inner product, that is

$$(\mathbf{r}_a, \mathbf{r}_b) = \underset{\mathbf{r}_i, \mathbf{r}_j \in \mathcal{A}_r}{\mathrm{argmin}} |\mathbf{r}_i \cdot \mathbf{r}_j| = \underset{\Pi_i, \Pi_j \in \mathcal{A}}{\mathrm{argmin}} \left|\mathrm{Tr}(\Pi_i \Pi_j) - \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right)\right| \tag{66}$$

$$= \underset{\Pi_i, \Pi_j \in \mathcal{A}}{\mathrm{argmin}} \max\left\{\mathrm{Tr}(\Pi_i \Pi_j) - \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right), \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right) - \mathrm{Tr}(\Pi_i \Pi_j)\right\} \tag{67}$$

$$= \underset{\Pi_i, \Pi_j \in \mathcal{A}}{\mathrm{argmin}} \max\left\{\mathrm{Tr}(\Pi_i \Pi_j), \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right)\right\}, \tag{68}$$

where in the last step we used that $\mathrm{Tr}(\Pi_i \Pi_j) + \mathrm{Tr}\left(\Pi_i(I - \Pi_j)\right) = 1$. Since $\mathcal{A}$ contains at least two independent directions, $\mathbf{r}_a, \mathbf{r}_b$ determine a plane. Now rotate $\mathbf{r}_a, \mathbf{r}_b$ symmetrically in this plane until we find two orthogonal directions $\mathbf{r}'_a$ and $\mathbf{r}'_b$ that obey the following conditions:

- $|\mathbf{r}'_a| = |\mathbf{r}'_b| = 1$ and $\mathbf{r}'_a \cdot \mathbf{r}'_b = 0$ (unit orthogonal vectors).

- $\mathbf{r}_a \cdot \mathbf{r}'_a = \mathbf{r}_b \cdot \mathbf{r}'_b$ (symmetrical rotation).

- $\mathbf{r}_a \times \mathbf{r}_b = \mathbf{r}'_a \times \mathbf{r}'_b$ (they remain in the same plane).

- $\max_{\mathbf{r}_i \in \mathcal{A}} \mathbf{r}'_a \cdot \mathbf{r}_i = \mathbf{r}'_a \cdot \mathbf{r}_a$, $\max_{\mathbf{r}_i \in \mathcal{A}} \mathbf{r}'_b \cdot \mathbf{r}_i = \mathbf{r}'_b \cdot \mathbf{r}_b$ (closest directions that obey the above conditions).

Using the directions $\mathbf{r}'_a$ and $\mathbf{r}'_b$ define the following environments ρ_a, ρ_b

$$\rho_a := \frac{I}{2} + \frac{\Delta}{2} \mathbf{r}'_a \cdot \sigma \quad \rho_b := \frac{I}{2} + \frac{\Delta}{2} \mathbf{r}'_b \cdot \sigma. \tag{69}$$
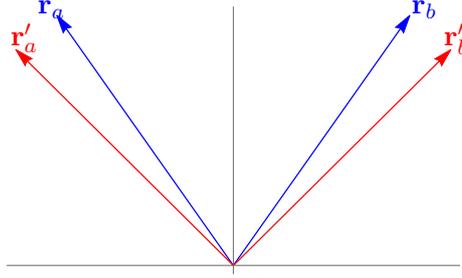
Figure 2: Scheme for the choice of the vectors $\mathbf{r}'_a$ and $\mathbf{r}'_b$.

Note that $\rho \geq 0$ and $\rho_b \geq 0$. In order to calculate the sub-optimality gaps (6) for ρ and ρ_b we will use Equation (59) and an orthonormal system of coordinates defined by the directions $\mathcal{C}_{a,b} = \{\mathbf{r}'_a, \mathbf{r}'_b, \mathbf{r}'_a \times \mathbf{r}'_b\}$. Using the above conditions for $\mathbf{r}'_a$ and $\mathbf{r}'_b$ we have

$$\Delta_i^a = \text{Tr}(\rho \Pi_i) = \Delta(p - r_{ia}), \quad \Delta_i^b = \text{Tr}(\rho_b \Pi_b) = \Delta(p - r_{ib}) \quad \text{for } i = 1, ..., k, \tag{70}$$

where $p = \mathbf{r}_a \cdot \mathbf{r}'_a$ and r_{ia}, r_{ib} are the first and second coordinates of $\mathbf{r}_i \in \mathcal{A}$ in the coordinate system $\mathcal{C}_{a,b}$ respectively. This last fact is because we defined $\mathcal{C}_{a,b}$ to be a orthonormal set of coordinates. Note that by the construction of $\mathbf{r}'_a$ and $\mathbf{r}'_b$ we have

$$p = \cos\left(\frac{\pi}{4} - \frac{\theta}{2}\right), \tag{71}$$

where θ is defined via $\cos\theta = \mathbf{r_a} \cdot \mathbf{r_b}$. Let Π_a, Π_b be the projectors associated to $\mathbf{r}_a, \mathbf{r}_b$, then using the trigonometric identity for $\cos\left(\frac{a}{2}\right)$ we have $\text{Tr}(\Pi_a \Pi_b) = \cos^2\left(\frac{\theta}{2}\right)$. Thus, using the trigonometric identity for $\cos(a + b)$ we have

$$p = \frac{1}{\sqrt{2}}\left(\sqrt{\text{Tr}(\Pi_a \Pi_b)} + \sqrt{1 - \text{Tr}(\Pi_a \Pi_b)}\right). \tag{72}$$

Now define the following subsets of indices of $[k]$,

$$\mathcal{N}_a = \left\{i \colon r_{ia}^2 \geq \frac{1}{2} \text{ for } \mathbf{r}_i \in \mathcal{A}_r\right\} \quad \mathcal{N}_a^C = \left\{i \colon r_{ia}^2 < \frac{1}{2} \text{ for } \mathbf{r}_i \in \mathcal{A}_r\right\}. \tag{73}$$

Note that this sets are complementary, $\mathcal{N}_a \cap \mathcal{N}_a^C = \emptyset$ and $\mathcal{N}_a \cup \mathcal{N}_a^C = [k]$.

Using the expression for the regret (7) and the sub-optimality gap (70), the regret for ρ can be bounded as,

$$R_n(\mathcal{A}, \rho_a, \pi) = \Delta \sum_{i=1}^k (p - r_{ia}) \, \mathbb{E}_{\rho_a, \pi}[T_i(n)] \tag{74}$$

$$\geq \Delta \sum_{i \in \mathcal{N}_a^C} (p - r_{ia}) \, \mathbb{E}_{\rho_a, \pi}[T_i(n)] \tag{75}$$

$$\geq \Delta \left(p - \frac{1}{\sqrt{2}}\right) \sum_{i \in \mathcal{N}_a^C} \mathbb{E}_{\rho_a, \pi}[T_i(n)], \tag{76}$$

where we have used that for $i \in \mathcal{N}_a^C$, we have $r_{ia} \leq \frac{1}{\sqrt{2}}$. Using that $n = \sum_{i=1}^k \mathbb{E}_{\pi, \rho_a}(T_i(n))$,

$$R_n(\mathcal{A}, \rho_a, \pi) \geq \Delta \left(p - \frac{1}{\sqrt{2}}\right) \mathbb{E}_{\pi, \rho_a}\left[n - \sum_{i \in \mathcal{N}_a} T_i(n)\right]. \tag{77}$$

Applying Markov's inequality to the above expression we have,

$$R_n(\mathcal{A}, \rho_a, \pi) \geq \frac{n\Delta}{2}\left(p - \frac{1}{\sqrt{2}}\right)P_{\pi,\rho_a}\left(n - \sum_{i \in \mathcal{N}_a} T_i(n) \geq \frac{n}{2}\right). \tag{78}$$

Rearranging all the terms,

$$R_n(\mathcal{A}, \rho_a, \pi) \geq \frac{n\Delta}{2}\left(p - \frac{1}{\sqrt{2}}\right)P_{\pi,\rho_a}\left(\sum_{i \in \mathcal{N}_a} T_i(n) \leq \frac{n}{2}\right). \tag{79}$$

In order to bound the regret for ρ_b note that if $i \in \mathcal{N}_a$ then $r_{ib}^2 \leq \frac{1}{2}$ since $\|\mathbf{r}_i\| = 1$. Using the same tricks as before,

$$R_n(\mathcal{A}, \rho_b, \pi) \geq \Delta \sum_{i \in \mathcal{N}_a}(p - r_{ib})\,\mathbb{E}_{\pi,\rho_b}[T_i(n)] \tag{80}$$

$$\geq \Delta\left(p - \frac{1}{\sqrt{2}}\right)\sum_{i \in \mathcal{N}_a}\mathbb{E}_{\pi,\rho_b}[T_i(n)]. \tag{81}$$

Using again Markov's inequality,

$$R_n(\mathcal{A}, \rho_b, \pi) \geq \frac{n\Delta}{2}\left(p - \frac{1}{\sqrt{2}}\right)P_{\pi,\rho_b}\left(\sum_{i \in \mathcal{N}_a} T_i(n) > \frac{n}{2}\right). \tag{82}$$

Thus, combining Equations (79) and (82),

$$R_n(\mathcal{A}, \rho_a, \pi) + R_n(\mathcal{A}, \rho_b, \pi) \geq$$

$$\frac{n\Delta}{2}\left(p - \frac{1}{\sqrt{2}}\right)\left(P_{\pi,\rho_a}\left(\sum_{i \in \mathcal{N}_a} T_i(n) \leq \frac{n}{2}\right) + P_{\pi,\rho_b}\left(\sum_{i \in \mathcal{N}_a} T_i(n) > \frac{n}{2}\right)\right). \tag{83}$$

Using Lemma 6 we can bound the above expression as,

$$R_n(\rho_a, \pi, \mathcal{A}) + R_n(\mathcal{A}, \rho_b, \pi) \geq \frac{n\Delta}{4}\left(p - \frac{1}{\sqrt{2}}\right)\exp\left(-D(P_{\pi,\rho_a}\|P_{\pi,\rho_b})\right). \tag{84}$$

Using Lemma 8 combined with data-processing inequality we can bound the Kullback–Leibler divergence as,

$$D(P_{\pi,\rho_a}\|P_{\pi,\rho_b}) = \sum_{i=1}^{k}\mathbb{E}_{\pi,\rho_a}[T_i(n)]D\big(P_{\rho_a}(\cdot|i)\|P_{\rho_b}(\cdot|i)\big) \tag{85}$$

$$\leq D(\rho_a\|\rho_b)\sum_{i=1}^{k}\mathbb{E}_{\rho,\pi}[T_i(n)] = nD(\rho_a\|\rho_b). \tag{86}$$

Using that the relative entropy $D(\rho_a\|\rho_b)$ is unitarily invariant and $\mathbf{r}_a', \mathbf{r}_b'$ are orthogonal we can use the computation of the relative entropy given by Lemma 11. Thus we can lower bound Equation (84) as,

$$R_n(\mathcal{A}, \rho_a, \pi) + R_n(\mathcal{A}, \rho_b, \pi) \geq \frac{n\Delta}{4}\left(p - \frac{1}{\sqrt{2}}\right)\exp\left(-\frac{n\Delta}{2}\log\left(\frac{1+\Delta}{1-\Delta}\right)\right). \tag{87}$$

Choosing $\Delta = \frac{1}{2\sqrt{n}}$ we have the following upper bound,

$$\log\left(\frac{1+\Delta}{1-\Delta}\right) \leq 2\log(3)\Delta \quad \text{for } \Delta \in [0, 1/2], \tag{88}$$

where we have used that $\log\left(\frac{1+x}{1-x}\right)$ is convex and finite for $x \in [0, \frac{1}{2}]$. Thus, choosing $\Delta = \frac{1}{2\sqrt{n}}$ we have

$$R_n(\mathcal{A}, \rho_a, \pi) + R_n(\mathcal{A}, \rho_b, \pi) \geq \frac{p - \frac{1}{\sqrt{2}}}{8 \cdot 3^{1/4}} \sqrt{n}, \tag{89}$$

and the result follows using the expression for p (72) and $15 > 8 \cdot 3^{1/4}\sqrt{2}$. From Equation (72) is easy to check that $p \geq \frac{1}{\sqrt{2}}$, thus the above lower bound is positive. □

5.3 Pauli observables

The next set of actions that we study are Pauli observables. If we consider a $d = 2^m$−dimensional (m qubits) Hilbert space, there are d^2 different Pauli observables and they can be expressed as the m-fold tensor product of the 2×2 Pauli matrices. Let $\sigma_1, ..., \sigma_{d^2}$ denote all the possible Pauli observables. Each σ_i can be expressed as $\sigma_i = \Pi_i^+ - \Pi_i^-$, where Π_i^+, Π_i^- are projectors associated to the $+1$ and -1 subspaces and they describe the 2 possible outcomes when we perform measurements using Pauli observables.

Theorem 13. *Let $n \in \mathbb{N}$. For any policy π and action set of observables $\mathcal{A}$ comprised of k distinct length-m strings of single-qubit Pauli observables for $d = 2^m$ with $m \in \mathbb{N}$, there exists an environment $\rho \in \mathcal{S}_d$ such that*

$$R_n(\mathcal{A}, \rho, \pi) \geq \frac{3}{100}\sqrt{(k-1)n}, \tag{90}$$

for $n \geq 2(k-1)$.

Proof. The case $k = 1$ is trivial since $R_n = 0$ always. Suppose $k > 1$ and let $0 \leq \Delta \leq \frac{1}{3}$ be a constant to be chosen later. Pick $\sigma_1 \in \mathcal{A}$ and define the following environment,

$$\rho := \frac{I}{d} + \frac{\Delta}{d}\sigma_1. \tag{91}$$

Define

$$l := \underset{j > 1}{\operatorname{argmin}} \, \mathbb{E}_{\rho, \pi}[T_j(n)] \tag{92}$$

as the index for the least expected picked observable different from σ_1. Define a second environment as,

$$\rho' := \frac{I}{d} + \frac{\Delta}{d}\sigma_1 + \frac{2\Delta}{d}\sigma_l. \tag{93}$$

Note that $\rho \geq 0$ and $\rho' \geq 0$ since $0 \leq \Delta \leq \frac{1}{3}$. In order to compute the sub-optimality gaps (6) we compute the following quantities for $\sigma_i \in \mathcal{A}$,

$$\operatorname{Tr}(\rho\sigma_i) = \frac{\Delta}{d}\operatorname{Tr}(\sigma_1\sigma_i) = \begin{cases} \Delta & \text{if } i = 1. \\ 0 & \text{otherwise.} \end{cases} \tag{94}$$

$$\text{Tr}(\rho'\sigma_i) = \frac{\Delta}{d}\left(\text{Tr}(\sigma_1\sigma_i) + 2\,\text{Tr}(\sigma_l\sigma_i)\right) = \begin{cases} 2\Delta & i = l. \\ \Delta & i = 1. \\ 0 & \text{otherwise.} \end{cases} \tag{95}$$

For ρ the sub-optimality (6) gaps are $\Delta_i = \Delta$ for $i \neq 1$ and $\Delta_i = 0$ for $i = 1$. Thus we can compute the regret as,

$$R_n(\mathcal{A}, \rho, \pi) = \sum_{a=1}^{k} \mathbb{E}_{\rho,\pi}[T_a(n)]\Delta_a = \Delta \sum_{a \neq 1} \mathbb{E}_{\rho,\pi}[T_a(n)] \tag{96}$$

Using Markov's inequality and $n = \sum_{i=a}^{k} \mathbb{E}_{\rho,\pi}[T_a(n)]$,

$$R_n(\mathcal{A}, \rho, \pi) = \Delta\,\mathbb{E}_{\rho,\pi}\left[n - T_1(n)\right] \geq \frac{n\Delta}{2} P_{\rho,\pi}\left(T_1(n) \leq \frac{n}{2}\right). \tag{97}$$

For ρ' we bound the regret for the term $i = 1$, using that the sub-optimality gap is $\Delta_1 = \Delta$ and Markov's inequality we have,

$$R_n(\mathcal{A}, \rho', \pi) = \sum_{a=1}^{k} \mathbb{E}_{\rho',\pi}[T_a(n)]\Delta_a \geq \Delta\,\mathbb{E}_{\rho',\pi}[T_1(n)] \geq \frac{n\Delta}{2} P_{\rho',\pi}\left(T_1(n) > \frac{n}{2}\right). \tag{98}$$

Combining Equations (97),(98) and the Bretagnolle-Huber inequality (6) we have,

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \frac{n\Delta}{4} \exp\left(-D(P_{\rho,\pi}\|P_{\rho',\pi})\right). \tag{99}$$

Note that the probabilities of the rewards are just Bernoulli distributions since the observables $\sigma_i \in \mathcal{A}$ have two outcomes, +1 and -1. Using Lemma 8 we can express

$$D(P_{\rho,\pi}\|P_{\rho',\pi}) = \sum_{a=1}^{k} \mathbb{E}_{\rho,\pi}[T_a(n)]D(P_\rho(\cdot|a)\|P_{\rho'}(\cdot|a)). \tag{100}$$

Note that $P_\rho(\cdot|a), P_{\rho'}(\cdot|a)$ are not equal only when $a = l$, so $D(P_\rho(\cdot|a)\|P_{\rho'}(\cdot|a)) = 0$ for $a \neq l$ and,

$$D(P_{\rho,\pi}, P_{\rho',\pi}) = \mathbb{E}_{\rho,\pi}[T_l(n)]D(P_\rho(\cdot|l)\|P_{\rho'}(\cdot|l)). \tag{101}$$

Using Equations (94) and (95) we have

$$P_\rho(1|l) = \frac{1}{2}, \quad P_{\rho'}(1|l) = \frac{1}{2} + \Delta. \tag{102}$$

Then we can compute the Kullback–Leibler divergence as,

$$D(P_\rho(\cdot|l)\|P_{\rho'}(\cdot|l)) = \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}+\Delta} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}-\Delta} = \frac{1}{2}\log\frac{1}{1-4\Delta^2}. \tag{103}$$

Note that using the definition of l (92) it follows that $n = \sum_{i=1}^{k} E_{\rho,\pi}[T_i(n)] \geq \sum_{i=2}^{k} E_{\rho,\pi}[T_i(n)] \geq (k-1)E_{\rho,\pi}[T_l(n)]$. Thus it holds that,

$$E_{\rho,\pi}[T_l(n)] \leq \frac{n}{k-1}. \tag{104}$$

Combining Equations (99),(101),(103) and (104) we have,

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \frac{n\Delta}{4} \exp\left(-\frac{n}{2(k-1)} \log \frac{1}{1-4\Delta^2}\right). \tag{105}$$

Finally choosing $\Delta = \frac{1}{2}\sqrt{1 - e^{-\frac{k-1}{n}}}$ and using that $e^{-x} \leq (e^{-1} - 1)x + 1$ for $0 \leq x \leq 1$ we arrive to the result,

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \frac{e^{-1/2}\sqrt{1-e^{-1}}}{8}\sqrt{(k-1)n}. \tag{106}$$

The condition $n \geq 2(k-1)$ suffices to have $0 \leq \Delta \leq \frac{1}{3}$. The theorem follows using $\frac{e^{-1/2}\sqrt{1-e^{-1}}}{8} > \frac{3}{50}$. $\qquad\square$

5.4 Pure-states environments

The last setting that we study for the discrete multi-armed quantum bandits is for pure-states environments. Specifically we restrict the problem to one-qubit pure-state environments with actions set comprised of the Pauli observables.

Theorem 14. *Let $n \in \mathbb{N}$. For any policy π and action set $\mathcal{A}$ containing the Pauli observables for 1-qubit, there exists an environment $|\psi\rangle\langle\psi| \in \mathcal{S}_2^*$ such that*

$$R_n(\mathcal{A}, \psi, \pi) \geq \frac{3}{200}\sqrt{n}.$$

Proof. Let $\Delta \in [0,1]$ and define the following the following two pure-states environments,

$$|\psi_\pm\rangle := c_\pm \left(1 + \frac{1 \pm \Delta}{\sqrt{2}}|0\rangle + \frac{1+\Delta}{\sqrt{2}}|1\rangle\right), \tag{107}$$

where $c_\pm = \frac{1}{\sqrt{\left(1 + \frac{1\pm\Delta}{\sqrt{2}}\right)^2 + \left(\frac{1\pm\Delta}{\sqrt{2}}\right)^2}}$. Let $|0\rangle, |+\rangle, |\psi_y\rangle$ be the eigenvector of the Pauli observables $\sigma_z, \sigma_x, \sigma_y$ respectively and define $x_\pm = \frac{1\pm\Delta}{\sqrt{2}}$. Then, compute the following quantities

$$p_\pm^x = |\langle+|\psi_\pm\rangle|^2 = \frac{(1+2x_\pm)^2}{2\left((1+x_\pm)^2 + x_\pm^2\right)}, \quad p_\pm^z = |\langle0|\psi_\pm\rangle|^2 = \frac{(1+x_\pm)^2}{(1+x_\pm)^2 + x_\pm^2}, \tag{108}$$

$$p_\pm^y = |\langle\psi_y|\psi_\pm\rangle|^2 = \frac{1}{2}.$$

Note that the expectation values for each arm on the environments can be computed as,

$$\mathrm{Tr}\left(\sigma_i|\psi_\pm\rangle\langle\psi_\pm|\right) = 2p_\pm^i - 1, \tag{109}$$

for $i = x, y, z$. Note that $\mathrm{Tr}\left(\sigma_y|\psi_\pm\rangle\langle\psi_\pm|\right) = 0$, $\mathrm{Tr}\left(\sigma_x|\psi_+\rangle\langle\psi_+|\right) \geq \mathrm{Tr}\left(\sigma_z|\psi_+\rangle\langle\psi_+|\right) \geq 0$ and $\mathrm{Tr}\left(\sigma_z|\psi_-\rangle\langle\psi_-|\right) \geq \mathrm{Tr}\left(\sigma_x|\psi_-\rangle\langle\psi_-|\right) \geq 0$ for $\Delta \in [0,1]$. Thus, for the environment $|\psi_+\rangle$ the optimal action is given by σ_x and for $|\psi_-\rangle$ by σ_z. Let $\Delta_i^\pm$ denote the sub-optimality gap for the i-th action in the environment $\psi_\pm$ respectively. Using Equations (108) and (109) we can compute the sub-optimality gaps as,

$$\Delta_x^+ = 0, \quad \Delta_z^+ = \frac{2+\Delta}{(1+\Delta)^2 + \sqrt{2}(1+\Delta) + 1}\Delta, \quad \Delta_y^+ = \frac{(\Delta+1)(1+\sqrt{2}+\Delta)}{\Delta^2 + (2+\sqrt{2})\Delta + 2 + \sqrt{2}} \tag{110}$$

$$\Delta_x^- = \frac{2-\Delta}{(1-\Delta)^2 + \sqrt{2}(1-\Delta) + 1}\Delta, \quad \Delta_z^- = 0, \quad \Delta_y^- = \frac{-\sqrt{2}\Delta + 1 + \sqrt{2}}{\Delta^2 - (\sqrt{2}+2)\Delta + \sqrt{2} + 2}. \tag{111}$$
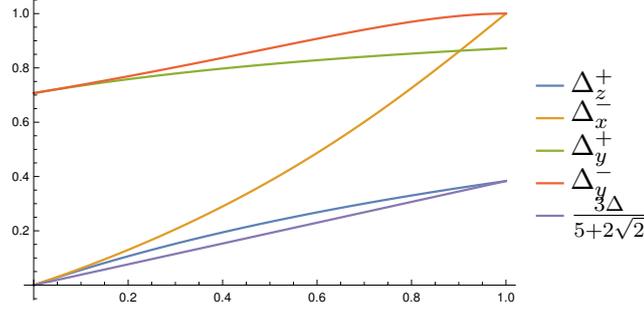
Figure 3: Sub-optimality gaps for the environments ψ_+, ψ_- and the lower bound $\frac{3}{5+2\sqrt{(2)}}\Delta$ in the range $\Delta \in [0, 1]$.

Analysing the behaviour of $\Delta_z^+, \Delta_y^+, \Delta_x^-, \Delta_y^-$ for $\Delta \in [0, 1]$ we see that all these quantities are lower bounded by $\frac{3}{5+2\sqrt{(2)}}\Delta$. In order to simplify the expression we will use that $\frac{3}{5+2\sqrt{(2)}}\Delta > \frac{3}{10}\Delta$ and take $\frac{3}{10}\Delta$ as the lower bound.

We start analysing the regret for ψ_+. Using the Equations (110) for the sub-optimality gaps and the bound $\Delta_z^+, \Delta_y^+ \geq \frac{3}{10}\Delta$ we have

$$R_n(\mathcal{A}, \psi_+, \pi) = \sum_{i=x,y,z} \mathbb{E}_{\psi_+,\pi}(T_i(n))\Delta_i^+ \geq \frac{3}{10}\Delta \sum_{i=y,z} \mathbb{E}_{\psi_+,\pi}(T_i(n)). \qquad (112)$$

Now using $n = \sum_{i=x,y,z} \mathbb{E}_{\psi_+,\pi}(T_i(n))$ and applying Markov's inequality we have

$$R_n(\mathcal{A}, \psi_+, \pi) \geq \frac{3}{10}\Delta\, \mathbb{E}_{\psi_+,\pi}\left(n - T_x(n)\right) \geq \frac{3n\Delta}{20} P_{\psi_+,\pi}\left(T_x(n) \leq \frac{n}{2}\right). \qquad (113)$$

In order to bound the regret for the environment ψ_- we bound just the first term, use Markov inequality and the bound Δ_x^- in order to obtain

$$R_n(\mathcal{A}, \psi_-, \pi) = \sum_{i=x,y,z} \mathbb{E}_{\psi_-,\pi}(T_i(n))\Delta_i^- \geq \Delta_x^-\, \mathbb{E}_{\psi_-,\pi}(T_i(n)) \geq \frac{3n\Delta}{20} P_{\psi_-,\pi}\left(T_x(n) > \frac{n}{2}\right). \qquad (114)$$

Combining Equations (113) and (114)

$$R_n(\mathcal{A}, \psi_+, \pi) + R_n(\mathcal{A}, \psi_-, \pi) \geq \frac{3n\Delta}{20}\left(P_{\psi_+,\pi}\left(T_x(n) \leq \frac{n}{2}\right) + P_{\psi_-,\pi}\left(T_x(n) > \frac{n}{2}\right)\right). \qquad (115)$$

Applying Lemma 7 together with Lemma 9 we obtain

$$R_n(\mathcal{A}, \psi_+, \pi) + R_n(\mathcal{A}, \psi_-, \pi) \geq \frac{3n\Delta}{40} \exp\left(-n D_{\frac{1}{2}}(\psi_+\|\psi_-)\right), \qquad (116)$$

where $D_{\frac{1}{2}}(\psi_+\|\psi_-) = -\log|\langle\psi_+|\psi_-\rangle|^2$. The overlap between the two environments can be computed as

$$\langle\psi_+|\psi_-\rangle = 2 + \sqrt{2} - \Delta^2, \qquad (117)$$

and we use it to give the following upper bound

$$-\log|\langle\psi_+|\psi_-\rangle| = \log\frac{1}{2+\sqrt{2}-\Delta^2} \le \frac{\Delta^2-1-\sqrt{2}}{2+\sqrt{2}-\Delta^2} \le \frac{\Delta^2}{2+\sqrt{2}-\Delta^2} \le \frac{\Delta^2}{1+\sqrt{2}}, \quad (118)$$

where the last inequality follows from $\frac{1}{2+\sqrt{2}-\Delta^2} \le \frac{1}{1+\sqrt{2}}$ for $\Delta \in [0,1]$.

Thus, plugging Equation (118) into Equation (116) we have

$$R_n(\mathcal{A},\psi_+,\pi) + R_n(\mathcal{A},\psi_-,\pi) \ge \frac{3n\Delta}{40}\exp\left(-\frac{2}{1+\sqrt{2}}n\Delta^2\right). \quad (119)$$

Finally, if we choose $\Delta = \frac{1}{\sqrt{n}}$,

$$R_n(\mathcal{A},\psi_+,\pi) + R_n(\mathcal{A},\psi_-,\pi) \ge \frac{3}{40}\exp\left(-\frac{2}{1+\sqrt{2}}\right)\sqrt{n}, \quad (120)$$

and the result follows using $\frac{3}{40}\exp\left(-\frac{2}{1+\sqrt{2}}\right) \ge \frac{3}{100}$. $\qquad\square$

6 Regret lower bounds for general bandits

6.1 Rank-1 projectors for general environments

In the following theorem, we consider a bandit whose action set is the set of all rank-1 projections, i.e., $\mathcal{A} = \mathcal{S}_d^*$ where $d \in \mathbb{N}$ is the dimension of the Hilbert space. When we play the arm $|\varphi\rangle\langle\varphi| \in S_d^*$ on the environment state ρ, the probability of reward 1 is $\langle\varphi|\rho|\varphi\rangle$ and the probability of reward 0 is $1 - \langle\varphi|\rho|\varphi\rangle$. Denoting by $\lambda_{\max}(\rho)$ the highest eigenvalue of any environment state ρ, the sub-optimality gap is $\Delta_\varphi := \lambda_{\max}(\rho) - \langle\varphi|\rho|\varphi\rangle$ for the environment state ρ upon playing the arm $|\varphi\rangle\langle\varphi|$.

Theorem 15. *Let $n, d \in \mathbb{N}$. For any policy π and action set of observables $\mathcal{A}$ containing all rank-1 projections, i.e., $\mathcal{A} = \mathcal{S}_d^*$, there exists an environment $\rho \in \mathcal{S}_d$ such that*

$$R_n(\mathcal{A},\rho,\pi) \ge C_\mathcal{A}\sqrt{n} \quad (121)$$

for some constant $C_\mathcal{A} > 0$ that depends on the action set.

Proof. Let us fix a policy π and an orthonormal basis $\{|n\rangle\}_{n=0}^{d-1}$ for $\mathbb{C}^d$ and define $|\psi\rangle := d^{-1/2}(|0\rangle + \cdots + |d-1\rangle)$ and the sets

$$\mathcal{N}_1 := \left\{ |\eta\rangle\langle\eta| \in \mathcal{S}_d^* \,\middle|\, |\langle 0|\eta\rangle|^2 < \frac{3}{4} + \frac{1}{4d} \right\}, \quad (122)$$

$$\mathcal{N}_2 := \left\{ |\eta\rangle\langle\eta| \in \mathcal{S}_d^* \,\middle|\, |\langle\psi|\eta\rangle|^2 < \frac{3}{4} + \frac{1}{4d} \right\}. \quad (123)$$

Let us show that these sets are disjoint. First note that $P(|0\rangle\langle 0|, |\psi\rangle\langle\psi|) = \sqrt{1-1/d}$ where $P = \sqrt{1-F^2}$ is the purified distance where, in turn, F is the fidelity. Assume that $|\eta\rangle\langle\eta| \in \mathcal{N}_1$, so that, using the triangle inequality for the purified distance,

$$\sqrt{1-\frac{1}{d}} = P(|0\rangle\langle 0|, |\psi\rangle\langle\psi|) \le P(|0\rangle\langle 0|, |\eta\rangle\langle\eta|) + P(|\eta\rangle\langle\eta|, |\psi\rangle\langle\psi|) \quad (124)$$

$$< \frac{1}{2}\sqrt{1-\frac{1}{d}} + P(|\eta\rangle\langle\eta|, |\psi\rangle\langle\psi|) \quad (125)$$

where we have used the definition of $\mathcal{N}_1$ in the final inequality. Thus, $P(|\eta\rangle\langle\eta|, |\psi\rangle\langle\psi|) > (1/2)\sqrt{1 - 1/d}$ which is easily seen to imply $|\eta\rangle\langle\eta| \in \mathcal{N}_2^c$. Thus, $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$.

Let us now define

$$\rho := \frac{1 - \Delta}{d} I + \Delta |0\rangle\langle 0|, \tag{126}$$

where $\Delta \in [0, 1]$ is a constant to be determined later. For this environment, the sub-optimality gap is

$$\Delta_\eta = \Delta(1 - |\langle 0|\eta\rangle|^2). \tag{127}$$

We may now evaluate

$$R_n(\mathcal{A}, \rho, \pi) \geq \Delta \int_{\mathcal{N}_1^c} (1 - |\langle 0|\eta\rangle|^2) \, d\gamma_{\rho,\pi}(|\eta\rangle\langle\eta|) \geq \Delta \frac{d-1}{4d} \gamma_{\rho,\pi}(\mathcal{N}_1^c). \tag{128}$$

Let us define another state

$$\rho' := \frac{1 - \Delta}{d} I + \Delta |\psi\rangle\langle\psi|, \tag{129}$$

where $|\psi\rangle$ is the unit vector defined earlier. Similarly as above, we find

$$R_n(\mathcal{A}, \rho', \pi) \geq \Delta \int_{\mathcal{N}_2^c} (1 - |\langle\psi|\eta\rangle|^2) \, d\gamma_{\rho',\pi}(|\eta\rangle\langle\eta|) \geq \Delta \frac{d-1}{4d} \gamma_{\rho',\pi}(\mathcal{N}_2^c) \tag{130}$$

$$\geq \Delta \frac{d-1}{4d} \gamma_{\rho',\pi}(\mathcal{N}_1) \tag{131}$$

where the final inequality follows from $\mathcal{N}_1 \subseteq \mathcal{N}_2^c$; recall that $\mathcal{N}_1$ and $\mathcal{N}_2$ are disjoint. Recalling that $(1/n)\gamma_{\rho,\pi}$ and $(1/n)\gamma_{\rho',\pi}$ are probability measures and using Lemma 6, we now obtain

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \Delta \frac{d-1}{4d} n \left(\frac{1}{n} \gamma_{\rho,\pi}(\mathcal{N}_1^c) + \frac{1}{n} \gamma_{\rho',\pi}(\mathcal{N}_1) \right) \tag{132}$$

$$\geq \Delta \frac{d-1}{4d} n \exp\left(-D\left(\frac{1}{n} \gamma_{\rho,\pi} \middle\| \frac{1}{n} \gamma_{\rho',\pi} \right) \right). \tag{133}$$

Recall that, e.g., $\gamma_{\rho,\pi} = \sum_{t=1}^n \left(P_{\rho,\pi}^{(t)}(\cdot, 0) + P_{\rho,\pi}^{(t)}(\cdot, 1) \right)$. Defining, for all $t = 1, \ldots, n$, the (measurable) function $f_t : (\mathcal{A} \times \{0, 1\})^n \to \mathcal{A}$ through

$$f_t((|\eta_s\rangle\langle\eta_s|, x_s)_{s=1}^n) = |\eta_t\rangle\langle\eta_t|, \tag{134}$$

we now see that $\gamma_{\rho,\pi} = \sum_{t=1}^n P_{\rho,\pi} \circ f_t^{-1}$ where we view $P_{\rho,\pi}$ as a measure. Using the joint convexity of the Kullback-Leibler divergence and the data processing inequality, we finally get

$$D\left(\frac{1}{n} \gamma_{\rho,\pi} \middle\| \frac{1}{n} \gamma_{\rho',\pi} \right) \leq \frac{1}{n} \sum_{t=1}^n D(P_{\rho,\pi} \circ f_t^{-1} \| P_{\rho',\pi} \circ f_t^{-1}) \tag{135}$$

$$\leq \frac{1}{n} \sum_{t=1}^n D(P_{\rho,\pi} \| P_{\rho',\pi}) = D(P_{\rho,\pi} \| P_{\rho',\pi}) \tag{136}$$

$$\leq n D(\rho \| \rho') \tag{137}$$

where the final inequality follows from Lemma 9 (for $\alpha = 1$). Combining this with Equation (132) gives us

$$R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi) \geq \Delta \frac{d-1}{4d} n \exp\left(-nD(\rho\|\rho')\right). \tag{138}$$

Whenever $\Delta \leq 1/2$, we may follow the end of the proof of Theorem 10 and obtain $c > 0$ such that $D(\rho\|\rho') \leq (c/2)\Delta^2$. Fixing $\Delta = 1/\sqrt{n}$, we now have

$$\max\{R_n(\mathcal{A}, \rho, \pi), R_n(\mathcal{A}, \rho', \pi)\} \geq \frac{1}{2}\left(R_n(\mathcal{A}, \rho, \pi) + R_n(\mathcal{A}, \rho', \pi)\right) \geq \frac{d-1}{8d}\sqrt{n}e^{-c/2}. \tag{139}$$

$\square$

We would like to mention that for continuous action sets there are different methods for proving minimax regret lower bounds for classical stochastic linear bandit (see [1][Chapter 24] or [15]). While the generalization to quantum is far from trivial we expect that alternatives techniques would be able to extract a non-trivial dimensional dependence on our regret lower bound.

7 Algorithms and regret upper bounds

In this section we are going to review some of the multi-armed stochastic bandit algorithms and see how they can be implemented in the multi-armed quantum bandit case. First we are going to review the linear stochastic bandits where the expected reward has a linear structure in terms of the actions and explain the LinUCB (linear upper confidence bound) algorithm. We will see that the LinUCB algorithm can be applied to both discrete or continuous sets of actions for multi-armed quantum bandits. For the discrete case we are going to review the UCB algorithm and the Phased Elimination algorithm. The first one does not assume correlations between the arms whereas the second one does. For different regimes of the number of actions, these algorithms offer better scaling on the regret than LinUCB.

First of all we are going to give an expression for the rewards of the multi-armed quantum bandit model where there is a linear part that depends on the actions and a random part that comes from some subgaussian noise. We introduce the notion of subgaussianity. We say that a random variable X is R-subgaussian if for all $\mu \in \mathbb{R}$ we have,

$$\mathbb{E}\left[\exp(\mu X)\right] \leq \exp(R^2\mu^2/2). \tag{140}$$

The notion of subgaussianity implies that $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \leq R^2$.

The following Lemma will allow us to match the linear structure of the rewards of the multi-armed quantum bandit problem to the classical bandit problem.

Lemma 16. *Let $n, d \in \mathbb{N}$. Consider a general or discrete multi-armed quantum bandit problem $(\mathcal{A}, \Gamma)$ with environment $\rho \in \Gamma$ such that, for any $O \in \mathcal{A}$, $\|O\| \leq 1$. Then if $O_t \in \mathcal{A}$ is the observable selected at round $t \in [n]$, the rewards are of the form*

$$X_t = \boldsymbol{\theta} \cdot \mathbf{A}_t + \eta_t, \tag{141}$$

where $\boldsymbol{\theta}, \mathbf{A}_t \in \mathbb{R}^{d^2}$, $\mathrm{Tr}(\rho O_t) = \boldsymbol{\theta} \cdot \mathbf{A}_t$ and η_t is 1-subgaussian given $X_1, O_1, ..., X_{t-1}, O_{t-1}$.

Proof. Let $\{\sigma_i\}_{i=1}^{d^2}$ be a set of orthonormal $(\mathrm{Tr}\,(\sigma_i\sigma_j)=\delta_{ij})$ Hermitian matrices $(\sigma_i^\dagger = \sigma_i)$. Then we have that for any $\rho \in \Gamma$ and action $O_a \in \mathcal{O}_d$

$$\rho = \sum_{i=1}^{d^2} \theta_i\sigma_i, \quad O_a = \sum_{i=1}^{d^2} A_{a,i}\sigma_i, \tag{142}$$

where $\theta_i = \mathrm{Tr}(\rho\sigma_i)$ and $A_{a,i} = \mathrm{Tr}(O_a\sigma_i)$. We will denote $\boldsymbol{\theta} \in \mathbb{R}^{d^2}$ the vector associated to ρ with components θ_i and $\mathbf{A}_a \in \mathbb{R}^{d^2}$ the vector associated to O_a with components $A_{a,i}$. Note that $\mathrm{Tr}(\rho O_a) = \boldsymbol{\theta} \cdot \mathbf{A}_a$ since the set $\{\sigma_i\}_{i=1}^{d^2}$ is orthonormal.

At round t define $\eta_t = X_t - \mathrm{Tr}(\rho O_t)$ where X_t is the reward and O_t the observable selected by the learner. Recall that $\mathbb{E}[X_t] = \mathrm{Tr}(\rho O_t)$ and using the assumption $\|O_a\| \leq 1$ we have $|X_t| \leq 1$. Thus, if he apply Hoeffding Lemma [24, Equation 4.16] to η_t we have that for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda\eta_t)] \leq \exp\left(\frac{\lambda^2}{2}\right). \tag{143}$$

By definition of subgaussian (140) it follows that η_t is 1-subgaussian, and the result follows. □

7.1 LinUCB algorithm for general multi-armed quantum bandits

In this section we will review one algorithm that has been used for classical bandits named LinUCB (linear upper confidence bound) or LinRel (linear reinforcement learning). The classical bandits are analogous to our quantum case with the main difference that instead of performing a measurement at each round we sample a reward from a set of probability distributions. In order to introduce the LinUCB algorithm we quickly review stochastic linear bandits.

The linear bandit model is described as follows. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ be an unknown vector and $\mathcal{A} \subset \mathbb{R}^d$ be a set of vectors that we call the action set. Then we have a learner, that during a sequence of n rounds, at each round $t \in \{1,...,n\}$ selects a vector $\mathbf{A}_t \in \mathcal{A}$ and samples a reward

$$X_t = \boldsymbol{\theta} \cdot \mathbf{A}_t + \eta_t \tag{144}$$

where η_t is R-subgaussian given $\mathbf{A}_1, X_1, ..., \mathbf{A}_{t-1}, X_{t-1}$ and we call it the *gaussian noise*. The regret is defined as

$$R_n(\boldsymbol{\theta}, \pi, \mathcal{A}) = \sum_{t=1}^n \max_{\mathbf{A}_i \in \mathcal{A}} \boldsymbol{\theta} \cdot \mathbf{A}_i - \mathbb{E}_{\boldsymbol{\theta},\pi}[\boldsymbol{\theta} \cdot \mathbf{A}_t]. \tag{145}$$

The expectation value above is taken with respect to the probability distribution $P_{\boldsymbol{\theta},\pi}$ determined by the policy π analogously to Equation (4) where the conditional probabilities $P_\rho(\cdot|\cdot) = P_{\boldsymbol{\theta}}(\cdot|\cdot)$ are now some subgaussian conditional distributions which depend on the specific problem we are studying. In order to analyse the regret of an algorithm it is convenient to define the *pseudo-regret* as

$$\hat{R}_n(\boldsymbol{\theta}, \pi, \mathcal{A}) = \sum_{t=1}^n \max_{\mathbf{A}_i \in \mathcal{A}} \boldsymbol{\theta} \cdot \mathbf{A}_i - \boldsymbol{\theta} \cdot \mathbf{A}_t. \tag{146}$$

Note that $\mathbb{E}_{\boldsymbol{\theta},\pi}[\hat{R}_n] = R_n$. Now we introduce the LinUCB algorithm and explain how it fits to our quantum model. The first paper that studied an algorithm for linear stochastic bandits was [25] where they considered a finite action set. The LinUCB algorithm that we shall now introduce is based on [26, 27, 28] where they allow an infinite action set and are based on statistical techniques in order to estimate the unknown parameter $\boldsymbol{\theta}$. The constructions and proofs that we will explain can be found in these references.

The LinUCB is an algorithm that minimizes the regret for linear stochastic bandits and it works constructing a confidence ellipsoid region for the unknown parameter $\boldsymbol{\theta}$ and selecting the best action on $\mathcal{A}$ that maximizes the reward in the confidence region. The confidence region is constructed via the least-square estimator,

$$\hat{\boldsymbol{\theta}}_t = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(\sum_{s=1}^{t} (X_s - \boldsymbol{\theta} \cdot \mathbf{A}_s)^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \right), \tag{147}$$

where X_s is the reward received at round s, $\mathbf{A}_s$ the action selected from $\mathcal{A}$ at round s and $\lambda \geq 0$ is a regulator parameter that ensures that the function has a unique minimum. The above equation can be solved for $\boldsymbol{\theta}$, which has the following closed form:

$$\hat{\boldsymbol{\theta}}_t = V_t^{-1} \sum_{s=1}^{t} \mathbf{A}_s X_s \tag{148}$$

where V_t is a $d \times d$ matrix,

$$V_t = \lambda I + \sum_{s=1}^{t} \mathbf{A}_s \mathbf{A}_s^T. \tag{149}$$

Note that V_t is positive definite by construction and induces the norm $\|\mathbf{x}\|_{V_t}^2 = \mathbf{x}^T V_t \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^d$. Using the estimator (148) we can build at each round t a confidence region $\mathcal{C}_t \subset \mathbb{R}^d$. Assume that the gaussian noise η_t for each X_t is 1-subgaussian and let m be an upper bound on $\|\boldsymbol{\theta}\|_2$, L an upper bound for any action, i.e for any $\boldsymbol{A} \in \mathcal{A}$, $\|\boldsymbol{A}\|_2 \leq L$, and $\delta \in (0, 1)$. Then the following theorem gives us the exact form of the confidence region:

Theorem 17 (Theorem 20.5 in [1]). *Let $d, n \in \mathbb{N}$, $t \in [n]$ and $\delta \in (0, 1)$. Define a classical linear bandit problem with action set $\mathcal{A} \subset \mathbb{R}^d$ and hidden parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ such that for any $\mathbf{A} \in \mathcal{A}$, $\|\mathbf{A}\|_2 \leq L$ and $\|\boldsymbol{\theta}\|_2 \leq m$ for some $L, m \in \mathbb{R}$. Assume that the rewards are of the form (144) with 1-subgaussian noise. Then, $P_{\boldsymbol{\theta},\pi}(\text{exists } t \in [n] : \boldsymbol{\theta} \notin \mathcal{C}_t) \leq \delta$ with*

$$\mathcal{C}_t = \left\{ \boldsymbol{\theta}^* \in \mathbb{R}^d : \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}\|_{V_{t-1}} \leq m\sqrt{\lambda} + \sqrt{2 \log\left(\frac{1}{\delta} + d \log\left(\frac{d\lambda + nL^2}{d\lambda} \right) \right)} \right\}. \tag{150}$$

Algorithm 1 LinUCB

1: **for** $t = 1, 2, \ldots$ **do**
2:      $(\mathbf{A}_t, \tilde{\boldsymbol{\theta}}_t) = \operatorname{argmax}_{\mathbf{A} \in \mathcal{A}, \boldsymbol{\theta} \in \mathcal{C}_t} \langle \boldsymbol{\theta}, \mathbf{A} \rangle$;
3:      Select $\mathbf{A}_t$ and observe reward X_t;
4:      Update $\mathcal{C}_t$;
5: **end for**

The LinUCB algorithm works as follows: at each round $t \in \{1, ..., n\}$ it constructs the confidence interval $\mathcal{C}_t$ and chooses and estimator $\tilde{\boldsymbol{\theta}}_t$ and action that maximizes the reward, i.e $(\mathbf{A}_t, \tilde{\boldsymbol{\theta}}_t) = \operatorname{argmax}_{\mathbf{A} \in \mathcal{A}, \boldsymbol{\theta} \in \mathcal{C}_t} \langle \boldsymbol{\theta}, \mathbf{A} \rangle$, and selects $\mathbf{A}_t$ in order to sample the reward. The pseudo-code can be found in Algorithm 1.

Theorem 18 (Theorem 19.2 in [1]). *Under the assumptions of Theorem 17 for $\boldsymbol{\theta} \in \mathbb{R}^d, \mathcal{A} \subseteq \mathbb{R}^d$ the pseudo-regret* (146) *of LinUCB satisfies*

$$\hat{R}_n(\boldsymbol{\theta}, \pi, \mathcal{A}) \leq \sqrt{8dn\beta_n \log\left(\frac{d\lambda + nL^2}{d\lambda}\right)} \tag{151}$$

where

$$\sqrt{\beta_n} = \sqrt{\lambda}m + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(\frac{d\lambda + nL^2}{d\lambda}.\right)} \tag{152}$$

In particular the regret (145) *of LinUCB with $\delta = 1/n$ is bounded by*

$$R_n(\boldsymbol{\theta}, \pi, \mathcal{A}) \leq Cd\sqrt{n}\log(n), \tag{153}$$

where $n \in \mathbb{N}$ and $C > 0$ is a suitably large universal constant.

The following theorem shows how the LinUCB can be applied to the multiarmed quantum bandit problem and the scaling of the regret.

Theorem 19. *Let $d, n \in \mathbb{N}$, and consider a general or discrete multi-armed quantum bandit with action set $\mathcal{A}$ such that, for all $O \in \mathcal{A}$, $\|O\| \leq 1$. Then the LinUCB algorithm, associated to policy π, can be applied to the general or discrete multi-armed bandit problem and for some universal constant $C > 0$ and for any $\rho \in \mathcal{S}_d$ the regret is bounded by*

$$R_n(\mathcal{A}, \rho, \pi) \leq Cd^2\sqrt{n}log(n). \tag{154}$$

Proof. Considering a set $\{\sigma_i\}_{i=1}^{d^2}$ of orthonormal Hermitian matrices, we can parametrize $\rho \in \mathcal{S}_d$ and any $O_a \in \mathcal{A}$ as $\rho = \sum_{i=1}^{d^2} \theta_i\sigma_i$, $O_a = \sum_{i=1}^{d^2} A_{a,i}\sigma_i$ where $\theta_i = \text{Tr}(\rho\sigma_i)$ and $A_{a,i} = \text{Tr}(O_a\sigma_i)$. We denote $\boldsymbol{\theta}, \mathbf{A}_a \in \mathbb{R}^{d^2}$ as the vectors of components θ_i and $A_{a,i}$ respectively. If O_t is the action selected at time step $t \in [n]$ we can use Lemma 9 and express the rewards of a multi-armed quantum bandit as

$$X_t = \boldsymbol{\theta} \cdot \mathbf{A}_t + \eta_t, \tag{155}$$

where η_t is 1-subgaussian. Thus, this rewards is of the form (144). Using that $\text{Tr}(\rho O_a) = \boldsymbol{\theta} \cdot \mathbf{A}_a$ and $\mathbb{E}_{\rho,\pi}(\eta_t) = 0$, we can express the regret for the multi-armed quantum bandit as

$$R_n(\mathcal{A}, \rho, \pi) = \sum_{t=1}^{n} \max_{\mathbf{A}_i} \boldsymbol{\theta} \cdot \mathbf{A}_i - \mathbb{E}_{\rho,\pi}[\boldsymbol{\theta} \cdot \mathbf{A_t}], \tag{156}$$

which is of the form of the linear stochastic bandit (145). With these observations we see how the multi-armed quantum bandit is mapped to the linear bandit and the LinUCB algorithm can be used. In order to bound the regret we need to check the conditions of Theorem 17. It remains to check an upper bound for $\|\boldsymbol{\theta}\|_2$. Using that the set $\{\sigma_i\}_{i=1}^{d^2}$ is orthonormal we have that $\text{Tr}(\sigma_i^2) = 1$, so $\|\sigma\| \leq 1$. Thus,

$$\|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^{d^2} \text{Tr}(\rho\sigma_i)^2} \leq d. \tag{157}$$

Inserting the above bound into the analysis of Theorem 18 and taking into account that $\boldsymbol{\theta}, \mathbf{A}_a \in \mathbb{R}^{d^2}$ the result follows.

$\square$

7.2 UCB and Phased Elimination algorithms for discrete multi-armed quantum bandits

The previous algorithm can be applied to both discrete and general multi-armed quantum bandits. For the discrete case we can apply the upper confidence bound (UCB) algorithm that was first introduced in [29]. We will see that UCB can offer a tighter bound of the regret under certain condition on the number of actions and the dimension of the Hilbert space. First we are going to review the UCB algorithm for multi-armed stochastic bandits.

A *multi-armed stochastic bandit* is defined with a set of probability distributions (environment) $\nu = (P_a : a \in \mathcal{A})$ where $\mathcal{A}$ is the set of actions of cardinality $k = |\mathcal{A}|$ and we denote by μ_i for $i \in [k]$ the mean of each action. At each round $t \in [n]$ a learner selects an action $A_t \in \mathcal{A}$ and samples a reward $X_t \sim P_{A_t}$. The regret is defined as

$$R_n(\nu, \pi) = \sum_{t=1}^{n} \max_{i \in [k]} \mu_i - \mathbb{E}[X_t], \tag{158}$$

where π is the policy and for each action $a \in [k]$ the sub-optimality gap is defined as $\Delta_a = \max_{i \in [k]} \mu_i - \mu_a$. Note that this model is more general than the linear bandits where the rewards where of the form (144). Moreover, the discrete multi-armed quantum bandit falls trivially in this class of bandits.

The basic idea of the UCB algorithm is to overestimate the mean of each action with high probability. The version of UCB that we consider assumes that the rewards are 1-subgaussian and it is based on the following bound of the empirical mean. If $(Z_t)_{t=1}^{n}$ is a sequence of independent 1-subgaussian random variables with mean μ and empirical mean $\hat{\mu}$, then for $\delta \in (0, 1)$,

$$\Pr\left(\mu \geq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta. \tag{159}$$

Let n be the number of rounds and $(X_t)_{t=1}^{n}$ be the sequence of rewards for a multi-armed stochastic bandit. For each action indexed by $a \in [k]$ we define its empirical mean as

$$\hat{\mu}_a(n) = \frac{1}{T_a(n)} \sum_{t=1}^{n} X_t \mathbb{I}\{A_t = a\}, \tag{160}$$

where A_t is the index of the action picked at round t and $T_a(n) = \sum_{t=1}^{n} \mathbb{I}\{A_t = a\}$ the number of times that we have picked the action indexed by $a \in [k]$. Then at each round $t \in [n]$ for each action $a \in [k]$ the UCB algorithm assigns a possible mean $\tilde{\mu}(t)_a$ as

$$\tilde{\mu}(t)_a = \begin{cases} \infty & \text{if } T_a(t-1) = 0. \\ \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} & \text{otherwise} \end{cases} \tag{161}$$

and selects the action $A_t = \text{argmax}_{a \in [k]} \tilde{\mu}(t)_a$. Note that the assignment of ∞ to the actions that have not been played ensures that each is action is played at least once during the first k rounds. The pseudocode of the UCB algorithm can be found in Algorithm 2.

Algorithm 2 UCB

1: **for** $t = 1, 2, \ldots, n$ **do**

2:     Choose action $A_t = \mathrm{argmax}_{i \in [k]} \begin{cases} \infty & \text{if} \quad T_i(t-1) = 0. \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise.} \end{cases}$

3:     Observe reward X_t and update confidence bounds;

4: **end for**

The following result on the scaling of the regret of UCB applied to discrete multi-armed quantum bandits follows from Lemma 16 and the analysis of the UCB regret for stochastic bandits from Theorem 7.2 in [1]. Recall that UCB can be applied to the multi-armed quantum bandit problem since they fall in the class of multi-armed stochastic bandits.

Theorem 20. *Let $n \in \mathbb{N}$ and consider discrete multi-armed quantum bandit with action set $\mathcal{A}$ of cardinality $k = |\mathcal{A}|$ such that for all $O \in \mathcal{A}$, $\|O\| \leq 1$. Then, the UCB algorithm associated to policy π, can be applied to the discrete multi-armed bandit problem with and for $\delta = \frac{1}{n^2}$ and any $\rho \in \mathcal{S}_d$ the regret can be bounded by*

$$R_n(\mathcal{A}, \rho, \pi) \leq 8\sqrt{nk\log(n)} + \sum_{a=1}^{k} \Delta_a. \tag{162}$$

Now we are going to compare the scaling of the UCB regret with the LinUCB regret for the discrete multi-armed quantum bandits. Let k be the number of observables and d the dimension of the Hilbert space of a discrete multi-armed bandit problem. The above Theorem shows that if $k \leq d^4$ the regret scaling of UCB is better than LinUCB (Theorem (19)). Note that if we apply UCB to the discrete multi-armed quantum bandits the algorithm does not take advantage of the linear structure of the rewards.

We mention that there is a variant of LinUCB (see [1][Chapter 22]) that assumes a finite number of actions and can be applied to the discrete multi-armed quantum bandits. This variant is called Phased Elimination and it is stated in Algorithm 3. This algorithm requires the assumptions of Theorem 17 and the action set $\mathcal{A} \subseteq \mathbb{R}^d$ to be finite. The analysis of Theorem 22.1 in [1] along with the observations of the proof of Theorem 19 lead immediately to the following result:

Theorem 21. *Let $d, n \in \mathbb{N}$, and consider a discrete multi-armed quantum bandit problem with action set $\mathcal{A}$ of cardinality $k = |\mathcal{A}|$ such that for all $O \in \mathcal{A}$, $\|O\| \leq 1$. Then, the Phased Elimination algorithm associated to policy π, with $\delta = \frac{1}{n}$, can be applied to the discrete multi-armed bandit problem and for some universal constant $C > 0$ and for all $\rho \in \mathcal{S}_d$ the regret is bounded by*

$$R_n(\mathcal{A}, \rho, \pi) \leq Cd\sqrt{n\log(nk)}. \tag{163}$$

Note that the scaling is better than LinUCB as long as the number of observables k is not exponential in the dimension of the Hilbert space.

Algorithm 3 Phased Elimination

1: Set $l = 1$ and $\mathcal{A}_1 = \mathcal{A}$.
2: Let $t_l = t$ be the current time step t and find $\pi_l : \mathcal{A}_l \to [0, 1]$ with $\texttt{support}(\pi_l) \leq \frac{d(d+1)}{2}$ that maximises

$$\log \det \left(\sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a}) \mathbf{a} \mathbf{a}^T \right) \text{ subject to } \sum_{\mathbf{a} \in \mathcal{A}_l} \pi(\mathbf{a}) = 1$$

3: Let $\epsilon_l = \frac{1}{2^l}$ and $T_l(\mathbf{a}) = \left\lceil \frac{2 d \pi_l(\mathbf{a})}{\epsilon_l^2} \log\left(\frac{kl(l+1)}{\delta} \right) \right\rceil$ and $T_l = \sum_{\mathbf{a} \in \mathcal{A}_l} T_l(\mathbf{a})$
4: Choose each action $\mathbf{a} \in \mathcal{A}_l$ exactly $T_l(\mathbf{a})$ times
5: Calculate the empirical estimate:

$$\hat{\boldsymbol{\theta}}_l = V_l^{-1} \sum_{t=t_l}^{t_l + T_l} \mathbf{A}_t X_t \text{ with } V_l = \sum_{\mathbf{a} \in \mathcal{A}_l} T_l(a) \mathbf{a} \mathbf{a}^T$$

6: Eliminate low rewarding arms:

$$\mathcal{A}_{l+1} = \{ \mathbf{a} \in \mathcal{A}_l : \max_{\mathbf{b} \in \mathcal{A}_l} \hat{\boldsymbol{\theta}}_l \cdot (\mathbf{b} - \mathbf{a}) \leq 2\epsilon_l \}$$

7: $l = l + 1$ and **Goto Step 1**

7.3 Tomography algorithm for pure-states environments

The last algorithm that we study is for the general model of multi-armed quantum bandits with pure-states environments. Recall that in this setting the regret can be given as,

$$R_n(\mathcal{A}, \psi, \pi) := \frac{1}{4} \sum_{t=1}^{n} \mathbb{E}_{\rho, \pi} \left\| \rho - \Pi_{A_t} \right\|_1^2, \tag{164}$$

where $\rho = |\psi\rangle\langle\psi|$ is the unknown pure quantum state and Π_{A_t} is the rank-1 projector from our action set $\mathcal{A}$ that contains all rank-1 projectors.

The algorithm that we propose is based on the projected least squares (PLS) method given in [30]. Now, we briefly review this method and explain how to apply it for the regret analysis. Let $\mathcal{H}_d$ be a $d-$dimensional Hilbert space, $\rho \in \mathcal{D}(\mathcal{H}_d)$ be an unknown quantum state that we want to estimate and $\{M_i\}_{i=1}^{m}$ the elements of a POVM that we will use to measure ρ. Suppose that we perform measurements over n rounds and n_i is the number of times that we have observed the outcome i associated to the element M_i from the POVM. Then knowing that the probabilities associated to each outcome i are given by the Born's rule $\text{Tr}(\rho M_i)$, the least square estimator for ρ is defined as,

$$L_n = \operatorname*{argmin}_{X \in \mathcal{O}_d} \sum_{i=1}^{m} \left(\frac{n_i}{n} - \text{Tr}\left(M_i X \right) \right)^2. \tag{165}$$

The estimator L_n is not guaranteed to be a physical state, so the next step is to project it to the physical state space as follows,

$$\hat{\rho}_n = \operatorname*{argmin}_{\tilde{\rho} \in \mathcal{S}_d} \| L_n - \tilde{\rho} \|_2, \tag{166}$$

where $\| \cdot \mathring{u} \|_2$ denotes de Frobenius norm. In [30] the above estimator is studied for structured POVMs, Pauli observables and Pauli basis measurements. For our problem we will use Pauli observables. Fix $d = 2^k$, and let $\{\sigma_i\}_{i=1}^{d^2}$ be the set of all possible tensor products of the 2×2 Pauli matrices. Note that they form a basis of the form (142). For Pauli observables the least square estimator (165) has the following form

$$L_n = \frac{1}{d} \sum_{i=1}^{d^2} \left(\frac{n_i^+ - n_i^-}{n/d^2} \right) \sigma_i, \tag{167}$$

where $n_i^\pm$ are the empirical frequencies associated to the 2-outcomes POVM $\Pi_i^\pm = \frac{1}{2}(I \pm \sigma_i)$ and each observable has been measured n/d^2 times. The next step is to take the projection (166). The convergence analysis in [30] shows the following theorem.

Theorem 22 (Theorem 1 in [30]). *Let ρ be a quantum state in a d-dimensional Hilbert space and fix a number of samples $n \in \mathbb{N}$. Then, using Pauli measurements, the PLS estimator $\hat{\rho}_n$ (166) obeys,*

$$\Pr\left(\|\hat{\rho}_n - \rho\|_1 \le \epsilon\right) \ge 1 - e^{-\frac{n\epsilon^2}{43d^2r^2}} \quad \text{for} \quad \epsilon \in [0,1],$$

where $r = \min\{\mathrm{rank}(\rho), \mathrm{rank}(\hat{\rho}_n)\}$.

After reviewing the PLS estimation method we explain how we use it to construct an algorithm for our pure state problem and how to analyse the regret. Recall that we have considered an action set that contains all rank-1 projectors. For that reason we consider the algorithm for one qubit states ($d = 2$) since Pauli observables for one qubit can be measured using rank-1 projectors. The pseudo-code for the algorithm can be found in Algorithm 2. If n is the number of rounds, fix the error $\epsilon^2 = \frac{172 \log(n)}{\sqrt{n}}$ and assume $\frac{172 \log(n)}{\sqrt{n}} \le 1$ in order to have $\epsilon \in [0,1]$. The condition $\frac{172 \log(n)}{\sqrt{n}} \le 1$ is achieved by $n \ge 8 \cdot 10^6$. Then the algorithm (policy π) works as follows:

- During the first $\sqrt{n}$ rounds we perform the Pauli observables measurements, and calculate the estimator $\hat{\rho}_{\sqrt{n}}$ (166) for ρ. Then using Theorem 22 we have,

$$P_{\rho,\pi}\left(\|\hat{\rho}_{\sqrt{n}} - \rho\|_1 \le \epsilon\right) \ge 1 - \frac{1}{n}, \tag{168}$$

  where we have used $r = 1$ since $\mathrm{rank}(\rho) = 1$ and $\epsilon^2 = \frac{172 \log(n)}{\sqrt{n}}$. In order to ensure that the estimator is pure we project the estimator into the rank-1 subspace in a ϵ−ball as follows,

$$\hat{\rho} = \operatorname*{argmin}_{\rho \in \mathcal{S}_1^*} \|\hat{\rho}_{\sqrt{n}} - \rho\|_1 \quad \text{such that} \quad \|\hat{\rho}_{\sqrt{n}} - \rho\|_1 \le \epsilon. \tag{169}$$

  Recall that for the above equation there is at least one solution with probability greater than $1 - 1/n$ since ρ is rank-1 by assumption. Note that using (168) we have,

$$P_{\rho,\pi}\left(\|\hat{\rho} - \rho\|_1 \le 2\epsilon\right) \ge 1 - \frac{1}{n}. \tag{170}$$

- For the remaining rounds we perform the measurement using $\hat{\rho}$ as the rank-1 projector from our action set.

Algorithm 4 Bandit PLS

1: **for** $i = 1, 2, \ldots, d^2$ **do**
2:      **for** $j = 1, 2, \ldots, \lceil \sqrt{n} \rceil / d^2$ **do**
3:          Measure ρ with σ_i;
4:          Update outcomes $n_i^{\pm}$;
5:      **end for**
6: **end for**
7: Compute $L_n = \frac{1}{d} \sum_{i=1}^{d^2} \left(\frac{n_i^+ - n_i^-}{n/d^2} \right) \sigma_i$;
8: Compute $\hat{\rho}_{\sqrt{n}} = \operatorname{argmin}_{\tilde{\rho} \in \mathcal{S}_d} \|L_n - \tilde{\rho}\|_2$;
9: Fix $\epsilon^2 = \frac{43 d^2 \log(n)}{\sqrt{n}}$
10: Compute $\hat{\rho} = \operatorname{argmin}_{\rho \in \mathcal{S}_d^*} \|\hat{\rho}_{\sqrt{n}} - \rho\|_1 \quad$ such that $\quad \|\hat{\rho}_{\sqrt{n}} - \rho\|_1 \leq \epsilon$
11: **for** $t = \lceil \sqrt{n} \rceil + 1, \ldots, n$ **do**
12:      Measure ρ with $\hat{\rho}$;
13: **end for**

Theorem 23. *Let $|\psi\rangle$ be some unknown one-qubit pure state. Fix π to be the policy associated to Algorithm 4 with $d = 2$ applied to the unknown state $|\psi\rangle$. Let $n \in \mathbb{N}$ denote the number of rounds, assume $\frac{172 \log(n)}{\sqrt{n}} \leq 1$ $(n \geq 8 \cdot 10^6)$ and let $\mathcal{A}$ be a set containing all one-qubit rank-1 projectors. Then the regret can be bounded as,*

$$R_n(\mathcal{A}, \psi, \pi) = O\left(\sqrt{n} \log(n) \right). \tag{171}$$

Proof. First we will bound the random regret defined as,

$$R_n^*(\mathcal{A}, \psi, \pi) := \frac{1}{4} \sum_{t=1}^{n} \left\| |\psi\rangle\langle\psi| - \Pi_{A_t} \right\|_1^2, \tag{172}$$

and then we will take the expectation value. During the first $\sqrt{n}$ rounds the algorithm performs the Pauli measurements using rank-1 projectors and builds the estimator $\hat{\rho}$ using Equation (169). Using Equation (170) we have the probabilistic error bound

$$P_{\psi,\pi} \left(\left\| |\psi\rangle\langle\psi| - \hat{\rho} \right\|_1 \leq 2\epsilon \right) \geq 1 - \frac{1}{n}, \tag{173}$$

where $\epsilon^2 = \frac{172 \log(n)}{\sqrt{n}}$. Then, with probability greater than $1 - \frac{1}{n}$, the random regret can be bounded as,

$$R_n^*(\mathcal{A}, \psi, \pi) \leq \sqrt{n} + (n - \sqrt{n}) \frac{172 \log(n)}{\sqrt{n}} = O(\sqrt{n} \log(n)), \tag{174}$$

where the first term comes from the trivial bound $\frac{1}{2} \| |\psi\rangle\langle\psi| - \Pi_t \|_1 \leq 1$ for the first $\sqrt{n}$ rounds and the second term from the bound of Equation (173) for the remaining rounds where Π_t is the estimator (169). Let C be the hidden constant in $O(\sqrt{n} \log(n))$, and G the probabilistic event where $R_n^*(\mathcal{A}, \psi, \pi) \leq C\sqrt{n} \log(n)$ and G^C its complement. Note that $P_{\psi,\pi}(G^C) \leq 1/n$. Then we can calculate the regret as,

$$R_n(\mathcal{A}, \psi, \pi) = \mathbb{E}_{\psi,\pi}[R_n^*(\mathcal{A}, \psi, \pi)] = \mathbb{E}_{\psi,\pi}[\mathbb{I}\{G\} R_n^*(\mathcal{A}, \psi, \pi)] + \mathbb{E}_{\psi,\pi}[\mathbb{I}\{G^C\} R_n^*(\mathcal{A}, \psi, \pi)]. \tag{175}$$

For the first term we use the bound (174) given by the event G. For the second term we use the trivial bound $R_n(\mathcal{A}, \psi, \pi) \leq n$ combined with $P_{\psi,\pi}(G_i^C) \leq 1/n$. Thus,

$$R_n(\mathcal{A}, \psi, \pi) \leq C\sqrt{n} \log(n) + n P_{\psi,\pi}(G_i^C) \leq C\sqrt{n} \log(n) + 1 = O(\sqrt{n} \log(n)). \tag{176}$$

$\square$

8   Conclusions

We have proposed a new quantum learning framework generalizing multi-armed stochastic bandits to the quantum setting. We have given fundamental bounds on the scaling of the regret for different environments and action sets that show the difficulty of the associated learning tasks. For the case of mixed-state environments and discrete bandits our upper and lower bounds match. For the case of general bandits with action set containing all rank-1 projectors our upper and lower bound match in terms of the number of rounds n but there is a gap in the dimension of the system of d^2. However, in the case of pure-state environments, we are not yet able to show lower bounds that scale with the number of arms or the dimension of the system and we leave this as an open question. We showed that an algorithm based in the explore-then-commit strategy and the PLS estimator achieves a regret upper bound $R_n = O(\sqrt{n}\log(n))$ which scales as good as LinUCB in this setting. For the qubit case this problem can be mapped to stochastic linear bandits where the actions and the environments are vectors that lie in the unit sphere $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2^2 = 1\}$. In this setting, classical techniques fail to show a lower bound $R_n = \Omega(\sqrt{n})$, since although the unit sphere action set has been considered (see Chapter 24 in [1]) the lower bound is proven for environments where the 2-norm of the vector scales as $1/\sqrt{n}$. The problem can also be connected to phase retrieval bandits [31] but we encounter the same issue with the lower bounds. In [15] they showed a lower bound $R_n = \Omega(\sqrt{n})$ for actions in the unit sphere using a randomization technique that consists on computing the expectation of the regret over a distribution of possible environments. This result relies heavily in some properties of the linear least square estimator that do not hold if one consider a distribution of environments over the unit sphere. Thus, we identify that there is also a gap in the classical bandit literature if one considers a stochastic linear bandit with actions and environments in the unit sphere.

References

[1] T. Lattimore and C. Szepesvári. "Bandit algorithms". Cambridge University Press. (2020).

[2] A. Slivkins. "Introduction to multi-armed bandits". Foundations and Trends in Machine Learning **12**, 1–286 (2019).

[3] S. Bubeck and N. Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". Foundations and Trends in Machine Learning **5**, 1–122 (2012).

[4] D. Bouneffouf, I. Rish, and C. Aggarwal. "Survey on applications of multi-armed and contextual bandits". In 2020 IEEE Congress on Evolutionary Computation (CEC). Pages 1–8.   (2020).

[5] L. Tang, R. Rosales, A. Singh, and D. Agarwal. "Automatic ad format selection via contextual bandits". In Proceedings of the 22nd ACM International Conference on In-

formation and Knowledge Management. Page 1587–1594. Association for Computing Machinery (2013).

[6] M. Cohen, I. Lobel, and R. Paes Leme. "Feature-based dynamic pricing". Management Science **66**, 4921–4943 (2020).

[7] W. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika **25**, 285–294 (1933).

[8] H. Robbins. "Some Aspects of the Sequential Design of Experiments". Bulletin of the American Mathematical Society **58**, 527–535 (1952).

[9] T.L Lai and H. Robbins. "Asymptotically efficient adaptive allocation rules". Advances in Applied Mathematics **6**, 4–22 (1985).

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-time analysis of the multiarmed bandit problem". Mach. Learn. **47**, 235–256 (2002).

[11] B. Casalé, G. Di Molfetta, H. Kadri, , and L. Ralaivola. "Quantum bandits". Quantum Mach. Intell.**2** (2020).

[12] D. Wang, X. You, T. Li, and A. Childs. "Quantum exploration algorithms for multi-armed bandits". In Proceedings of the AAAI Conference on Artificial Intelligence. Volume 35, pages 10102–10110. (2021).

[13] P. Rebentrost, Y. Hamoudi, M. Ray, X. Wang, S. Yang, and M. Santha. "Quantum algorithms for hedging and the learning of ising models". Phys. Rev. A **103**, 012418 (2021).

[14] O. Shamir. "On the complexity of bandit linear optimization". In Proceedings of The 28th Conference on Learning Theory. Volume 40 of Proceedings of Machine Learning Research, pages 1523–1551. PMLR (2015).

[15] P. Rusmevichientong and J. Tsitsiklis. "Linearly parameterized bandits". Mathematics of Operations Research**35** (2008).

[16] J. Barry, D.T. Barry, and S. Aaronson. "Quantum partially observable markov decision processes". Phys. Rev. A **90**, 032311 (2014).

[17] M. Ying, Y. Feng, and S. Ying. "Optimal policies for quantum markov decision processes". International Journal of Automation and Computing **18**, 410–421 (2021).

[18] M. Paris and J. Rehacek. "Quantum state estimation". Springer Publishing Company, Incorporated. (2010). 1st edition.

[19] S. Aaronson. "Shadow tomography of quantum states". In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. Page 325–338. STOC 2018. Association for Computing Machinery (2018).

[20] S. Aaronson, X. Chen, E. Hazan, S. Kale, and A. Nayak. "Online learning of quantum states". Journal of Statistical Mechanics: Theory and Experiment**2019** (2018).

[21] J. Bretagnolle and C. Huber. "Estimation des densités: risque minimax". Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **47**, 119–137 (1979).

[22] M. Müller-Lennert, F. Dupuis, O. Szehr, S. Fehr, and M. Tomamichel. "On Quantum Rényi Entropies: A New Generalization and Some Properties". Journal of Mathematical Physics **54**, 122203 (2013).

[23] M. Wilde, A. Winter, and D. Yang. "Strong Converse for the Classical Capacity of Entanglement-Breaking and Hadamard Channels via a Sandwiched Rényi Relative Entropy". Communications in Mathematical Physics **331**, 593–622 (2014).

[24] W. Hoeffding. "Probability inequalities for sums of bounded random variables". Journal of the American Statistical Association **58**, 13–30 (1963).

[25] P. Auer. "Using confidence bounds for exploitation-exploration trade-offs". J. Mach. Learn. Res. **3**, 397–422 (2003).

[26] D. Varsha, T. Hayes, and S.Kakade. "Stochastic linear optimization under bandit feedback.". In Proceedings of the 21st Conference on Learning Theory. Pages 355–366. (2008).

[27] P. Rusmevichientong and J. N. Tsitsiklis. "Linearly parameterized bandits". Mathematics of Operations Research **35**, 395–411 (2010).

[28] Y. Abbasi-Yadkori, D. Pál, and Cs. Szepesvári. "Improved algorithms for linear stochastic bandits". In Advances in Neural Information Processing Systems. Volume 24. Curran Associates, Inc. (2011).

[29] T. L. Lai. "Adaptive Treatment Allocation and the Multi-Armed Bandit Problem". The Annals of Statistics **15**, 1091 – 1114 (1987).

[30] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp. "Fast state tomography with optimal error bounds". Journal of Physics A: Mathematical and Theoretical **53**, 204001 (2020).

[31] T. Lattimore and B. Hao. "Bandit phase retrieval". In Advances in Neural Information Processing Systems. Volume 34, pages 18801–18811. Curran Associates, Inc. (2021).