

Multi-Armed Bandits and Quantum Channel Oracles

Simon Buchholz, Jonas M. Kübler, Bernhard Schölkopf

Max Planck Institute for Intelligent Systems, Tübingen, Germany

Multi-armed bandits are one of the theoretical pillars of reinforcement learning. Recently, the investigation of quantum algorithms for multi-armed bandit problems was started, and it was found that a quadratic speedup (in query complexity) is possible when the arms and the randomness of the rewards of the arms can be queried in superposition. Here we introduce further bandit models where we only have limited access to the randomness of the rewards, but we can still query the arms in superposition. We show that then the query complexity is the same as for classical algorithms. This generalizes the prior result that no speedup is possible for unstructured search when the oracle has positive failure probability.

1 Introduction

Quantum computing is a model of computation that is based on quantum properties of matter. By using superposition and entanglement, it offers potentially large speedups when compared to classical algorithms. For some problems exponential speedups have been shown under the assumption of widely believed hardness results for classical computing, the two most prominent examples being Shor’s algorithm for factoring integers [1] and the HHL algorithm for sampling from the solution of sparse linear equations [2]. On the other hand, it was shown that for many problems only polynomial speedups are possible, in particular Grover’s algorithm [3] for the unstructured search problem only offers a quadratic speedup and no greater improvement is possible.

Recently, quantum machine learning has emerged as one potential area of application for quantum computers [4]. It was suggested to use quantum computers for linear algebra subroutines but also complete implementations of well-known classical algorithms for supervised learning were designed, e.g., quantum support vector machines [5], quantum principal component analysis [6], and recommender systems [7]. There has also been some work on unsupervised learning and reinforcement learning [8, 9, 10]. For a recent review, we refer to [11].

Multi-armed bandit problems are a class of problems from machine learning where a decision maker in each time step selects an action from a fixed set of options and then receives a corresponding reward. The goal of the decision maker is to identify the best possible action. The two main problems considered in this field are regret minimization where the goal is to maximize the total reward received, and best arm identification where we want to identify the best arm in the fewest rounds possible. Bandit problems have a long history, starting among others from the works [12, 13, 14] and a long list of works established different algorithms and (almost) optimal results for a wide range of settings, and we refer to the next section for additional details and the literature for a complete overview [15, 16].

Recently, the investigation of the best arm identification problem with quantum computers was initialized in [17]. They show that a quadratic speedup compared to the classical algorithm is possible and optimal in their setting. Their implementation of the bandit arms is, however, not directly comparable to the classical setting because they assume that the algorithm has control over the internal randomness of the bandit arms. In particular, the rewards are obtained by the action of a unitary oracle.

In this work, we will discuss when this multi-armed bandit model is applicable, and we introduce further models of multi-armed bandits that are more suitable in different settings. Those models

Simon Buchholz, Jonas M. Kübler, Bernhard Schölkopf: {sbuchholz, jmkuebler, bs}@tue.mpg.de, JMK is now with Amazon.

differ in the degree of control that we have over the internal randomness of the rewards of each arm. When we have no access to the randomness of the rewards, a pull of the arms can no longer be described by a unitary map, but instead they are modelled by a non-unitary quantum channel. This provides a link to the channel discrimination problem, however, there the focus has been on rather different types of quantum channels that are either generic or more related to the transmission of information [18]. Here we show that in this more restricted setting no polynomial speedup compared to classical algorithms is possible, even though the bandits can be queried in superposition. Previously, it was shown in [19] that no speedup for unstructured search is possible if the oracle has a fixed probability of error. Our results are an extension to a substantially more general setting. Thus, this work highlights that quantum speedups can be impeded by (small) amounts of classical randomness present in the quantum channel used in the algorithm, underscoring again that having parts of the computation routine without error correction poses challenges. From a technical side, we connect classical methods from quantum information theory with coupling techniques from probability theory.

The rest of this paper is structured as follows: In the next section, we introduce different settings of quantum bandits and give an overview of their query complexities. Then we discuss the main strategies and ingredients used in the proofs of the main results in Section 3. Full proofs of our results are delegated to the appendices.

2 Setting and Main Results

We now discuss the setting for our main results. In this section, we review the relevant background on (classical) bandits and then state our main results for quantum bandits.

2.1 Classical bandits

To set the scene, we briefly review the multi-armed bandit problem, which should also provide sufficient background for readers from the quantum side who are less familiar with the setting. In the multi-armed bandit problem, an agent can choose in every round t an arm (an action from a finite set) $i_t \in \{1, \dots, N\}$ and receives a probabilistic reward r_t depending on the chosen action. There are a vast number of variants and generalizations of this problem in the classical setting, but here we focus on the simplest model where the number of arms is finite, and the rewards upon choosing arm i are drawn independently from some distribution ν_i where $\nu_i \in \mathcal{P}$ is in some class of probability distributions known to the agent. Assume that ν_i has mean μ_i and we denote the highest mean reward by $\mu^* = \max_i \mu_i$.

The goal in bandit theory is generally to identify the action that leads to the highest mean reward μ^* . There are two main problems of interest. The first is regret minimization, where we try to minimize the regret which is defined after T rounds with rewards $(r_t)_{1 \leq t \leq T}$ by

$$\rho_T = T\mu^* - \sum_{t=1}^T r_t. \quad (1)$$

Equivalently, we want to maximize the total reward received, and the regret measures the reward difference to the optimal course of actions. Often, a key goal is to achieve sublinear regret, i.e., $\rho_T \in o(T)$ which means that asymptotically the average reward received converges to the mean of the best arm. In the context of multi-armed bandits, a much finer understanding of optimal regret bounds is known, namely it was shown that the uniform confidence bound algorithm is asymptotically optimal and corresponding bounds were derived (see, e.g., [14, 20, 21]). The regret minimization problem captures the trade-off between exploration, i.e., finding good, unexplored options, and exploitation, i.e., choosing favorable options, which is essential for reinforcement learning.

A second problem investigated in the bandit setting is the best arm identification problem. Here the goal is to identify the best arm with a fixed confidence (i.e., up to some given error probability at most δ) in the fewest rounds possible. Another variant is to maximize the probability of finding the best arm for a fixed time horizon. Note that the setting only features the exploration component and not the exploitation component of the regret minimization problem. Even though the setting

of best arm identification is simpler than regret minimization, it is often a challenging problem and in some settings optimal results were found later than for the regret minimization problem. In this paper, we focus on the best arm identification problem because this readily generalizes to the quantum setting, while it is not directly obvious how the regret should be defined and interpreted in the quantum setting.

Let us now first briefly review the main results and provide some intuition for the best arm identification problem in the classical multi-armed bandit setting introduced above. We restrict our attention to Bernoulli distributions $\nu_i \sim \text{Ber}(p_i)$ which have mass p on 1 and mass $1 - p$ on 0 because this can be easily generalized to the quantum setting where the reward can then be encoded by a single qubit. In this case, the reward distribution is fully characterized by the means p_i of the arms. For concreteness, we fix a mean reward vectors $\mathbf{p} = (p_0, \dots, p_{N-1}) \in \mathbb{R}^N$, indicating that arm i has mean reward p_i . Note that unconventional indexing is used here to ensure consistency with our setup later on. We usually assume that the rewards are ordered, i.e., $p_0 > p_1 \geq \dots \geq p_{N-1}$. We will use the shorthand $\Delta_i = p_0 - p_i$ for the difference in mean rewards between the best arm and arm i . We define the important quantity

$$H(\mathbf{p}) = \sum_{i \geq 1} \frac{1}{(p_0 - p_i)^2} = \sum_{i \geq 1} \Delta_i^{-2}. \quad (2)$$

It can be shown that $H(\mathbf{p})$ governs the optimal query complexity to identify the best arm (up to logarithmic terms) in a fixed confidence setting. Moreover, this is optimal when $p_i \in [\eta, 1 - \eta]$ for all i and some $\eta > 0$. The following well-known theorem provides a formal statement of those results.

Theorem 1 (Thm. 2 in [22], Thm. 5 in [23]). *Consider an algorithm that identifies the best arm of a multi-armed bandit with probability at least $1 - \delta$ for Bernoulli distributed rewards with reward vector $\mathbf{p} \in [0, 1]^N$ such that $\mathbf{p}_i \in [\eta, 1 - \eta]$. Then this algorithm requires $\Omega(H(\mathbf{p}))$ rounds in the worst case. On the other hand, there exists such an algorithm requiring $\tilde{O}(H(\mathbf{p}))$ steps.*

Let us emphasize that identification of the best arm is not easier if we know the rewards up to a permutation, in fact, the following result holds.

Theorem 2 (Thm. 4 in [24]). *Let $\mathbf{p} \in [\eta, 1 - \eta]^N$ be a reward vector. Then any algorithm that identifies the best arm for Bernoulli distributed rewards with means $\mathbf{p}'$ where $\mathbf{p}'$ is any permutation of $\mathbf{p}$ requires at least $\Omega(H(\mathbf{p}))$ rounds and such an algorithm exists (up to logarithmic terms).*

Let us add several remarks to these results. There is a long list of results improving upon the two results above by deriving bounds on the logarithmic correction, considering more general reward distributions, and analyzing various algorithms, see e.g., [25, 26, 27, 28].

Let us explain for the readers not so familiar with the literature on bandit problems the intuition underlying the results mentioned above. The key statistical problem is essentially to decide which of two arms has a higher reward. Assume that we know the mean reward p_0 of the best arm exactly. Let us first investigate how many pulls t_i of the arm with true mean reward p_i are sufficient to verify with high probability that $p_i < p_1$. We can bound the difference between the empirical mean $\hat{p}_i$ and p_i using Hoeffding's inequality (see Lemma 17) by

$$\mathbb{P}(\hat{p}_i - p_i \geq (p_0 - p_i)) \leq e^{-2t_i(p_0 - p_i)^2}. \quad (3)$$

This shows that after

$$t_i \gtrsim (p_0 - p_i)^{-2} = \Delta_i^{-2} \quad (4)$$

pulls we can rule out with high probability that an arm with true reward p_i has a return higher than p_0 . On the other hand, t_i of order Δ_i^{-2} is necessary when $p_i \in [\eta, 1 - \eta]$ for some constant $\eta > 0$. Indeed, by the central limit theorem $\hat{p}_i - p_i$ will be typically of order $\sqrt{\text{Var}(\text{Ber}(p_i))/t_i}$ and the variance is lower bounded by $\eta(1 - \eta)$ if $p_i \in [\eta, 1 - \eta]$. We conclude that (4) is necessary to conclude that $p_i < p_0$ with high probability. Applying this reasoning to all arms i suggests that $H(\mathbf{p})$ governs the query complexity of the best arm identification problem. Note that here we

ignored the fact that the highest reward p^* is unknown, and algorithms overcome this problem by using the optimism principle.

Let us emphasize that the argument for the lower bound uses crucially that all rewards are away from 1 and 0. Otherwise, we cannot consider the variance term $p_i(1 - p_i)$ to be a fixed constant in the analysis. To illustrate this, we consider the particular case that $p_1 = p > 0$ and $p_i = 0$ for $i > 1$. Then $H(\mathbf{p}) = N/p^2$ but only $\mathcal{O}(N/p)$ pulls are required to identify the best arm. To see this, note that it takes with high probability $\mathcal{O}(p^{-1})$ pulls to get a single success on an arm with $\text{Ber}(p)$ distributed rewards. This is the setting considered in [19] and we will come back to it in Section 3.1.

2.2 Quantum bandits

In this section, we discuss how the classical bandit problem can be generalized to a quantum setting. This builds upon the recent works [29, 17] that studied the best arm identification problem for quantum bandits. Here, we want to review and extend prior results and definitions in the literature and in particular highlight that different assumptions for the available oracles are reasonable in different settings.

Let us first explain the general implementation of decision problems in the quantum setting, while we refer to standard textbooks (e.g., [30]) for a general introduction to quantum computing. Typically, we assume that we are given black box access to an oracle O that is a unitary map on some Hilbert space from a finite set $\mathcal{O}$ of unitary maps and our goal is to identify which oracle we are given using the fewest possible number of invocations of O . To give a concrete example, we consider the arguably most famous example of unstructured search where we are given one of the oracles O^i that mark an element $i \in \{1, \dots, N\}$, i.e., they act by

$$O^i |j\rangle |c\rangle = |j\rangle |c \oplus \delta_{ij}\rangle \quad (5)$$

where $\oplus$ denotes addition modulo 2 and $|j\rangle$ is a basis of an N dimensional Hilbert space and $|c\rangle$ is a qubit basis state. Our goal is now to identify the marked element i by a quantum algorithm. This means that starting from some initial state $|\Omega\rangle$, we can apply a sequence of arbitrary unitary maps U_t interleaved by invocations of the oracle O to obtain the state

$$|\varphi_t\rangle = U_T O U_{T-1} O \dots O U_1 O U_0 |\Omega\rangle. \quad (6)$$

Finally, we perform a measurement on the state $|\varphi_t\rangle$ and measurement outcomes are then mapped to an oracle O^i . We say that the algorithm succeeds if it outputs i when applied with $O = O^i$ with a fixed lower bounded probability for all i . This can then be lifted to a high probability guarantee by repetition. In the example of unstructured search introduced above, Grover's algorithm can be used to identify i with $\mathcal{O}(\sqrt{N})$ calls to O^i [3]. Note that this setting can be interpreted as a bandit problem where a single arm always returns reward 1 and all other arms always return reward 0.

We now introduce our definition of general bandit problems. In this case, we need to consider the broader class of decision problems where we can partition the set of oracles $\mathcal{O} = \mathcal{O}_1 \sqcup \mathcal{O}_2 \sqcup \dots \sqcup \mathcal{O}_N$ and we want to identify n such that $O \in \mathcal{O}_n$ with the fewest number of invocations of the given oracle O . Let us now describe how the oracle O that models a pull of the arms is implemented. As before, we assume that there are N arms. To query the arms, we consider a Hilbert space $\mathcal{H}_A$ for the arms with dimension $|\mathcal{H}_A| = N$ and we identify the arms with states $|i\rangle$ from a fixed basis. Moreover, we assume that the internal randomness of the reward is captured through an additional Hilbert space $\mathcal{H}_P$ and we fix a basis $|\omega\rangle$. The reward for a certain arm and state of internal randomness is collected in a two-dimensional Hilbert space $\mathcal{H}_R$ with basis states $|0\rangle$ and $|1\rangle$. We then assume that in each round we can query the arms through an oracle O acting on $\in \mathcal{H}_A \otimes \mathcal{H}_P \otimes \mathcal{H}_R$. It acts on a state a basis state $|i\rangle |\omega\rangle |c\rangle$ i.e., arm i , internal randomness ω , and initial reward state c by

$$O |i\rangle |\omega\rangle |c\rangle = |i\rangle |\omega\rangle |c + r_i(\omega)\rangle \quad (7)$$

where $r_i(\omega) \in \{0, 1\}$ denotes the reward for arm i with internal randomness ω and addition is again modulo 2 (this is very similar to the definition in [31]). Note that $r_i(\omega)$ are not random for

fixed i and ω . Thus, for each arm we receive for every ω and any arm i a reward that is either 0 or 1 and averaging over ω gives

$$p_i = |\mathcal{H}_P|^{-1} \sum_{\omega} r_i(\omega) \quad (8)$$

the mean reward for arm i . As before, we collect the mean rewards p_i in a vector $\mathbf{p} \in [0, 1]^N$ with $\mathbf{p}_i = p_i$. Quantum algorithms then consist of a sequence of unitary operations on $\mathcal{H}_A \otimes \mathcal{H}_P \otimes \mathcal{H}_R \otimes \mathcal{H}_W$ where $\mathcal{H}_W$ is a work space interleaved by calls to the oracle O . We emphasize that the indices are named to reflect the Hilbert spaces for the arms, the internal randomness (probability), the rewards, and the work space.

Let us now compare this setting to the classical setting and outline a crucial difference. In the classical setting the rewards are random variables, e.g., functions from some probability space Ω to $\mathbb{R}$ but all we observe is the received reward and explicit reference to the probability space is not necessary. In contrast, here the probability space is made explicit through the space $\mathcal{H}_P$ and a crucial ingredient of the model. Moreover, the rewards for all rounds are given by $r_i(\omega)$ so there is no independence of the reward distributions for different times. This is different from the classical setting. But note that if we consider a setting where we can in each round query an action (i, ω) and receive the reward $r_i(\omega)$ the best-arm identification problem is not simpler than in the standard formulation (if the dimension of $\mathcal{H}_P$ is sufficiently large). On the other hand, the regret minimization problem is not meaningful in this setting because it is sufficient to identify a single good realization such that $r_i(\omega) = 1$ which can then be selected in all future rounds. Note that this setup might be a reasonable model for certain classical problems, e.g., when the different arms correspond to different exercises and different ω to different students and the goal is to identify the hardest exercise.

We will discuss potential applications in the quantum setting below, but first we consider different variants of the problem above. Motivated by the differences from the classical setting, we investigate three different settings that are characterized by the degree of control that we have over the space $\mathcal{H}_P$ determining the internal randomness of the bandits.

1. We have full control over the space $\mathcal{H}_P$, i.e., we can apply arbitrary unitary operators to the system $\mathcal{H}_A \otimes \mathcal{H}_P \otimes \mathcal{H}_R$ and a potential work space.
2. We can prepare an arbitrary computational basis state $\omega \in \mathcal{H}_P$ but have no further access to $\mathcal{H}_P$ except through the oracle O .
3. For each pull of the arm, i.e., for each invocation of O the state of $\mathcal{H}_P$ is a uniformly random basis state $|\omega\rangle$ which is resampled in each step. Equivalently, each oracle invocation uses a different copy of the maximally mixed state on $\mathcal{H}_P$.

We now discuss how the restricted control over $\mathcal{H}_P$ can be related to settings without an explicit space $\mathcal{H}_P$ which are therefore closer to the standard setup of classical bandits. We define for $X \in \{0, 1\}^N$ (corresponding to a vector of binary rewards) the oracle O_X acting on $\mathcal{H}_A \otimes \mathcal{H}_R$ by

$$O_X |i\rangle |c\rangle \rightarrow |i\rangle |c + X^i\rangle. \quad (9)$$

Let us assume that $X_t \in \{0, 1\}^N$ is a sequence of i.i.d. random variables where the coordinates X_t^i are independent and $\text{Ber}(p_i)$ distributed. Then having access to the sequence of oracles

$$O_{X_t} |i\rangle |c\rangle = |i\rangle |c + X_t^i\rangle, \quad \text{for } t = 1, 2, \dots \quad (10)$$

is similar to the second scenario above (by identifying $r_i(\omega_t) = X_t^i$ where ω_t enumerates the basis of $\mathcal{H}_P$). Note that here we assume for simplicity that the coordinates X_t^i are independent, while the general oracle definition in (7) can also model correlated rewards for different arms. In the classical setting, this is not important because we can never observe the rewards of two different arms in the same round. In the quantum setting, it might matter because we can query the arms in superposition.

Oracle	Lower bound	Upper bound
Classical	$\Omega(H(\mathbf{p}))$	$\tilde{O}(H(\mathbf{p}))$
ERM (eq. (7))	$\Omega(\sqrt{H(\mathbf{p})})$ (Thm. 4)	$\tilde{O}(\sqrt{H(\mathbf{p})})$ (Thm. 1 in [17])
Reusable (eq. (10))	?	$\tilde{O}(\sqrt{\sum_i \Delta_i^{-4}})$ (Thm. 5)
One-time (eq. (11))	$\Omega(H(\mathbf{p}))$ (Thm. 6)	$\tilde{O}(H(\mathbf{p}))$

Table 1: Overview of query complexity bounds. The upper bound for the one-time oracle and the reusable oracle follow from the classical result. We conjecture (see Conjecture 1) that the upper bound for the reusable oracle are optimal.

We now similarly reformulate the third setting above. This setting is similar to having access to an oracle acting as a quantum channel on $\mathcal{H}_A \otimes \mathcal{H}_R$ by

$$\mathcal{E}(\rho) = \sum_{X \in \{0,1\}^n} \mathbb{P}(X) O_X \rho (O_X)^\dagger$$

$$\text{where } \mathbb{P}(X) = \prod_i p_i^{X_i} (1 - p_i)^{1-X_i}.$$
(11)

Again, the difference to the setting above is the assumed independence of different arms of the variables X^i .

Note that the second and the third setting are equivalent if we are allowed to use each of the oracles O_{X_t} only once. Multiple invocations can be useful to uncompute parts of the computation and thereby avoiding decoherence of the system.

2.3 Models and Main Results

We now motivate the three settings described above in more detail and discuss our main results. Let us emphasize that quantum bandits can only be useful when the reward is given by the observable of a quantum system or the evaluation of a computation on a quantum device, as the acquisition of the data is commonly seen as the expensive part. In other words, it appears unlikely that we collect rewards in some trial and then store this data in, e.g., a QRAM to query them in superposition to identify the best arm, as this will always be more expensive than classically evaluating the mean of the collected rewards. Thus, we will motivate all three settings from a quantum perspective. Note that the classical setting roughly corresponds to the case where only queries of the form $|i\rangle|\omega\rangle$ without superposition are allowed, i.e., we can query a single arm for a single reward.

2.3.1 Empirical risk minimisation

As already explained in [17], one setting where we have full access to an oracle as in (7) is empirical risk minimization. To make this concrete, assume we have a dataset $(x_j, y_j) \in \mathcal{X} \times \{1, \dots, K\}$ and a finite set of candidate functions f_i . We now want to find the index i_0 such that

$$i_0 = \arg \min_i \sum_j \mathbf{1}_{f_i(x_j) \neq y_j} = \arg \max_i \sum_j \mathbf{1}_{f_i(x_j) = y_j},$$
(12)

i.e., for simplicity we consider 0-1 loss in a classification problem or equivalently, we maximize the accuracy over the functions f_i . If we can access x_j and evaluate f_i this provides us with an oracle acting by

$$|i\rangle|\omega\rangle|c\rangle \rightarrow |i\rangle|\omega\rangle|c \oplus r_i(\omega)\rangle,$$
(13)

where the reward for $\omega = (x, y)$ is given by $r_i(\omega) = \mathbf{1}_{y=f_i(x)}$. Now, the problem of best arm identification with respect to this oracle is equivalent to empirical risk minimization. Moreover,

this oracle is exactly of the form introduced in (7). When considering this setting, it is a reasonable assumption that the dataset can be accessed in arbitrary superposition, i.e., is stored in our computing device, and we can also evaluate functions f_i and thus losses in superposition. Note that this setup can also arise in the investigation of quantum systems. Suppose that we have a quantum system of interest with corresponding Hilbert space $\mathcal{H}_P$ and observables A_i acting on this space. We assume for simplicity that $A_i |\omega\rangle = r_i(\omega) |\omega\rangle$ where $r_i(\omega) \in \{0, 1\}$, i.e., the observables have eigenvalues 0 and 1 and can be simultaneously diagonalized. We are interested in finding the observable with the largest expectation, i.e.,

$$\arg \max_i \text{tr}(A_i \rho) = \arg \max_i \sum_{\omega} \langle \omega | A_i | \omega \rangle, \quad (14)$$

where $\rho \propto \sum_{\omega} |\omega\rangle \langle \omega|$ is the completely mixed state. Then, this reduces to the best arm identification problem if we can construct an oracle acting as in (7) where $r_i(\omega)$ corresponds to the eigenvalue of A_i for the eigenvector ω .

Note that the problem of empirical risk minimization and the relation to bandit problems was considered before in [17], where they consider an oracle that acts as

$$|i\rangle |0\rangle \rightarrow |i\rangle (\sqrt{p_i} |1\rangle + \sqrt{1-p_i} |0\rangle). \quad (15)$$

The relation to our setting is that when applying our oracle to a uniform superposition over the ω register we obtain

$$\sum_{\omega} |i\rangle |\omega\rangle |0\rangle \rightarrow \sqrt{p_i} |i\rangle |v\rangle |1\rangle + \sqrt{1-p_i} |i\rangle |u\rangle |0\rangle \quad (16)$$

where u and v are suitable junk states which can be neglected as argued in [17] (at least when restricting attention to query complexity). Note that this superposition eliminates the statistical randomness of the bandits. Then the following result holds.

Theorem 3 (Theorem 1 in [17]). *There is a quantum algorithm that identifies the best arm of a quantum oracle as in (15) for any reward vector $\mathbf{p} \in [\eta, 1-\eta]^N$ with probability $1-\delta$ with query complexity*

$$T \leq \tilde{O}(\sqrt{H(\mathbf{p})}). \quad (17)$$

Moreover, this bound is optimal up to logarithmic terms.

Thus, a quadratic speedup compared to the classical setting is achievable. They prove the lower bound only for the oracle (15). For completeness, we prove the same lower bound when the more general oracle (7) is available, i.e., the ability to query arbitrary superpositions of the data points does not allow any speedups compared to always considering the uniform superposition.

Theorem 4 (Informal version). *Any algorithm that identifies the best arm for any reward vector $\mathbf{p} \in [\eta, 1-\eta]^N$ with confidence $1-\delta$ given access to an oracle as in (7) requires at least $\Omega(\sqrt{H(\mathbf{p})})$ calls to the oracle.*

The proof and a formal statement of this result are in Appendix I. Note that while it is intuitively clear that it is optimal to query over the uniform mixture of ω as in (16) a rigorous proof requires a careful tracking of the classical randomness of the oracle and its interaction with the quantum algorithm.

2.3.2 Reusable oracles

We now consider oracles as in (10), i.e., we can query arms in superposition, but we can only retrieve the reward for one chosen realization of the randomness. A similar type of oracle was considered in [32]. They show that hedging algorithms can be implemented using these oracles which have runtime $O(\sqrt{N})$ for N arms, thus offering a quadratic speedup compared to the classical algorithm. Their setting is not directly comparable to the best arm identification problem for multi-armed bandits considered here. We try to identify the best arm, while in their setting an ε -optimal arm is

sufficient. On the other hand, they want to control a suitable variant of the regret. One motivation is that as in Section 2.3.1 we want to identify the observable A_i from a collection of observables that has the maximal expectation. But in contrast to the previous setting, we cannot perform arbitrary manipulations on the experimental setup (because $\mathcal{H}_P$ corresponds to a quantum system we study and not part of the computing device). Instead, we can only apply maps that transition between different ω values from the fixed basis of $\mathcal{H}_P$ and probe the system through the oracle O . As stated above, this is equivalent to having access to oracles of the form O_{X_t} acting on $\mathcal{H}_A \otimes \mathcal{H}_R$. While this setup might be not the physically most relevant model, it is nevertheless helpful as an intermediate setting that helps us understand when different speedups arise. For this case we can only give partial results.

We have the following result.

Theorem 5. *For confidence $\delta \in (0, 1)$ there is a quantum algorithm that outputs the best arm with probability $1 - \delta$ using*

$$T \leq \tilde{O} \left(\sqrt{\sum_{i \geq 1} \Delta_i^{-4}} \right) \quad (18)$$

queries to an oracle as in (10) where $\sim$ indicates terms that are polynomial in $\ln(N/(\delta\Delta_2))$.

A sketch of the proof of this result is in Appendix K. It relies on a small modification of the algorithm in [17]. Their algorithm is based on a clever application of variable time algorithms [33] to count the number of arms with reward bigger than a given threshold and to rotate on the corresponding subspace of arms. As this is the only upper bound in this work, its proof has no direct relation to the remaining results. We conjecture that the bound in Theorem 5 is also optimal.

Conjecture 1. *Any quantum algorithm that identifies the best arm for any reward vector $\mathbf{p} \in [\eta, 1 - \eta]^n$ for some $\eta > 0$ with probability $1 - \delta$ for some $\delta < \frac{1}{2}$ requires at least*

$$T \geq c \sqrt{\sum_{i \geq 1} \Delta_i^{-4}} \quad (19)$$

calls to an oracle as in (10).

The main difficulty to prove a lower bound is that the applied oracles can be reused so that the fidelity between the quantum states obtained for different mean rewards $\mathbf{p}$ and $\mathbf{p}'$ is not necessarily monotonically decreasing. This makes it hard to extend our other proofs that rely on the loss of fidelity in a single step to this setting. In general, standard techniques to obtain lower bounds and the techniques used in this work do not appear to be sufficient to address this problem.

2.3.3 One-time oracles

Finally, we consider the oracle defined in (11). Let us first make the connection to the oracle O in (7) a bit more precise. We assume that we cannot act on the space $\mathcal{H}_P$ and only extract information from the system through the oracle O otherwise the state on $\mathcal{H}_P$ follows, e.g., a time evolution given through some Hamiltonian H . Then it is reasonable to assume that the initial state on $\mathcal{H}_P$ is the completely mixed state $\rho_P \propto \sum |\omega\rangle \langle \omega|$ and then the reduced state on the system $\mathcal{H}_A \times \mathcal{H}_R$ after application of O is given by

$$\text{tr}_P O(\rho_{AR} \otimes \rho_P) O^\dagger = \sum_{X \in \{0,1\}^N} p(X) O_X \rho_{AR} O_X^\dagger \quad (20)$$

where $p(X) = |\{\omega : r_i(\omega) = X^i\}|/|\mathcal{H}_P|$ and O_X is defined in (9). In particular, we recover indeed the expression in (11) if the rewards $\omega \rightarrow r_i(\omega)$ for different arms i are independent. If the dynamics H is sufficiently mixing also future invocations will approximately act as the quantum channel $\mathcal{E}$ on the system.

Assume that the oracle O corresponds to our running example of identifying the mean of a collection of observables A_i . Then this setup is connected to the theory of quantum sensing [34] which refers to the general use of quantum phenomena to measure quantum observables. Using quantum devices to learn properties of quantum systems has recently emerged as a promising direction to achieve quantum advantage. In particular, it was shown that learning with a quantum device interacting with the system of interest can have an exponential advantage over simply performing measurements on the system [35]. We show that, in contrast, in our setting no speedup compared to the classical setting is possible as stated (slightly informally) in the following result.

Theorem 6. *Any algorithm that identifies the best arm for any reward vector $\mathbf{p} \in [\eta, 1 - \eta]^N$ for some $\eta > 0$ with probability $1 - \delta$ for some $\delta < 1/2$ based on calls to an oracle as in (11) requires at least*

$$T \geq c(\delta, \eta) H(\mathbf{p}) \quad (21)$$

calls to the oracle.

In particular, in our setting it is not advantageous (up to constant factors) to query the system in superpositions, but we can also instead use a classical algorithm to decide which arm i to query, evaluate $\mathcal{E}(|i\rangle\langle 0| \langle 0| \langle i|)$, i.e., query arm i and then measure the reward register containing a $\text{Ber}(p_i)$ sample. In the context of our example, where we want to identify the observable A_i with maximum mean, we can just directly measure $\langle A_i \rangle$.

Again, this result is not a consequence of the generality of reward vectors $\mathbf{p}$ allowed, but even when the set of mean rewards is known no better result is possible. Note that the assumption that the rewards are Bernoulli distributed and independent might be unrealistic for applications. However, our main result shows that even in this simplified case no improvement over classical algorithms is possible in terms of query complexity. A more precise and slightly stronger statement of the result above is given in Theorem 10. Theorem 10 and its proof can be found in Appendix H. An overview of the proof techniques and related results will be given in Section 3.

We also remark that on a technical level, our setting is similar to [19], where they essentially consider the case $p_i = 0$ for all $i \neq i_0$ and $p_{i_0} > 0$. Their motivation is to study Grover search where the oracle has a certain failure probability $1 - p_{i_0}$ and they also find that this impeded any quantum speedup. We will review their results in more detail in Section 3.1 which also provides insight into the more general problem.

2.4 Comparison of settings

Let us further discuss these results to give a better intuition of the result. We remark that access to the oracle (7) is strictly more powerful than access to oracles as in (10) which in turn is more powerful than access to the channel oracle (11). This is reflected in the following chain of inequalities for the query complexities

$$\sqrt{\sum_{i \geq 1} \Delta_i^{-2}} \leq \sqrt{\sum_{i \geq 1} \Delta_i^{-4}} \leq \sum_{i \geq 1} \Delta_i^{-2}. \quad (22)$$

Here we used $\Delta_i < 1$ for the first inequality. On the other hand, we have the following reverse bound for the complexities

$$\frac{1}{\Delta_1} \sqrt{\sum_{i \geq 1} \Delta_i^{-2}} \geq \sqrt{\sum_{i \geq 1} \Delta_i^{-4}} \geq \frac{1}{\sqrt{N}} \sum_{i \geq 1} \Delta_i^{-2}. \quad (23)$$

Here we used $\Delta_1 = p_0 - p_1 \leq p_0 - p_i = \Delta_i$ for the first inequality and the inequality between arithmetic and quadratic mean for the second inequality. Thus, the speedup of the empirical risk minimization setting compared to the (conjectured) complexity of the reusable oracles setting is at most $\Delta_1^{-1} = (p_0 - p_1)^{-1}$ while the speedup between the (conjectured) complexity of the reusable oracles setting and the quantum channel oracle is at most $\sqrt{N}$ but both can be less, depending

on $\mathbf{p}$. To illustrate this further we consider as a prototypical example a reward vector $\mathbf{p}$ with $p_0 > p_1 = \dots = p_{N-1}$ with $p_0 - p_1 = \varepsilon$. Then the query complexities are

$$T_{\text{term}} \approx \sqrt{\frac{N}{\varepsilon^2}} = \frac{\sqrt{N}}{\varepsilon} < T_{\text{reusable}} \approx \frac{\sqrt{N}}{\varepsilon^2} < T_{\text{classical}} \approx T_{\text{one-time}} \approx \frac{N}{\varepsilon^2}. \quad (24)$$

For this setting the two difficulties can be well separated: The expected reward of each arm needs to be estimated (statistical complexity) and the correct arm needs to be searched. Classically the statistical complexity is ε^{-2} but it can be reduced to ε^{-1} in the empirical risk minimization setting (similar to quantum metrology [36]). Not having access to a superposition of basis states $|\omega\rangle$ prevents this speedup. The complexity of the search of the best arm is $\sqrt{N}$ in a quantum setting compared to the complexity of N in the classical setting or a noisy quantum setting. This gives rise to the difference of the query complexities for the one-time oracle and the classical setting or the reusable oracle. Thus, even under the favorable assumptions we made, e.g., Bernoulli distributed rewards, quantum algorithms offer no improvement in query complexity when $\mathcal{H}_P$ is not part of the computing device.

3 Proof techniques and overview

In this section we give an overview of the techniques used in the proof of our main result, Theorem 6. The proof is rather involved and technical, so we collect and review the main ingredients here, along with some results that are of independent interest. We first introduce some additional notation.

To clarify the setting and to cover general oracles we assume that the oracle acts on a Hilbert space given by $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_R$ where $\mathcal{H}_A = \langle |i\rangle, 1 \leq i \leq N \rangle$ and $\mathcal{H}_R$ is the space where the output is written, which will typically be a single qubit space. We assume that we are given oracles O_i acting by

$$O_i |i\rangle \otimes |w\rangle = |i\rangle \otimes U|w\rangle, \quad O_i |j\rangle \otimes |w\rangle = |j\rangle \otimes |w\rangle \quad \text{for } j \neq i \quad (25)$$

where U denotes a unitary map. This covers the case of the phase flip ($|i\rangle \otimes |w\rangle \rightarrow -|i\rangle \otimes |w\rangle$) and the bit flip oracle ($|i\rangle \otimes |w\rangle \rightarrow |i\rangle \otimes |w \oplus 1\rangle$). A central role is played by the quantum channels $\mathcal{F}_i^p$ acting by

$$\mathcal{F}_i^p(\rho) = (1-p)\rho + pO_i\rho O_i^\dagger. \quad (26)$$

The oracle $\mathcal{F}_i^p$ implements the pull of arm i with a $\text{Ber}(p)$ distributed arm. Note that the channels $\mathcal{F}_i^{p_i}$ and $\mathcal{F}_j^{p_j}$ commute for $i \neq j$ because O_i and O_j commute for $i \neq j$.

To relate this to the setting of Theorem 6 we consider probability vectors $\mathbf{p} \in [0, 1]^N$ and define further

$$\mathcal{E}^{\mathbf{p}} = \mathcal{F}_1^{p_1} \circ \dots \circ \mathcal{F}_N^{p_N}. \quad (27)$$

When O_i denotes the bit flip of the reward register on arm $|i\rangle$ the channel $\mathcal{E}^{\mathbf{p}}$ agrees with the definition in (11) where the vector $\mathbf{p}$ indicates the mean rewards, i.e.,

$$\mathcal{E}^{\mathbf{p}}(\rho) = \sum_{x \in \{0,1\}^N} \mathbb{P}(x) O_x \rho O_x^\dagger \quad \text{where} \quad O_x = \prod_{i:x_i=1} O_i \quad \text{and} \quad \mathbb{P}(x) = \prod_i p_i^{x_i} (1-p_i)^{1-x_i}. \quad (28)$$

Let us now introduce a class of probability vectors $\mathbf{p}$, which we will use to establish our lower bounds. Specifically, we adopt the setup previously employed in the proof of Theorem 1 (see [22, 23]), where the lower bound is demonstrated for this particular class of reward vectors. Extending these results to a setting as in Theorem 2 is a promising direction for future research. We consider a probability vector $\mathbf{p} = (p_0, \dots, p_N) \in [0, 1]^{N+1}$ with $p_0 > p_1 > p_2 \geq \dots \geq p_N$. As before, we denote $\Delta_i = p_0 - p_i$. Given $\mathbf{p}$ we then define $N+1$ probability vectors $\mathbf{p}^i \in [0, 1]^N$ given by $\mathbf{p}_j^i = p_j$ for $i \neq j$ and $\mathbf{p}_i^i = p_0$ and $\mathbf{p}^0$ given by $\mathbf{p}_i^0 = p_i$. In other words, $\mathbf{p}^0 = (p_1, \dots, p_N)$ and $\mathbf{p}^i$ is obtained from $\mathbf{p}^0$ by replacing the i -th reward by p_0 . In particular, for every $i \geq 1$ the vectors

$\mathbf{p}^0$ and $\mathbf{p}^i$ differ only in the entry i and arm i has the highest reward for reward vector $\mathbf{p}^i$. We make the additional assumption that $\Delta_1 = p_0 - p_1 = p_1 - p_2$. This ensures (see Lemma 7)

$$\frac{1}{4}H(\mathbf{p}^0) \leq H(\mathbf{p}^i) \leq 2H(\mathbf{p}^0) \quad (29)$$

for all i . We will also use the shorthand $\mathcal{E}^{\mathbf{p}^j} = \mathcal{E}_j$. Our goal is then to show that it is hard to distinguish which of the reward vectors $\mathbf{p}^i$ corresponds to a given oracle O . More precisely, we show that there is an index i such that at least $\Omega(H(\mathbf{p}^0))$ oracle calls are necessary to identify arm i when the true reward vector is $\mathbf{p}^i$.

Now that we introduced the relevant notation, we can present the three main ingredients and techniques used in the proof. First, we show general lower bounds for the query complexity to distinguish the quantum channels $\mathcal{F}_i^p$ from the identity channel. Then we carefully analyze the change in fidelity when applying $\mathcal{F}_i^p$ and $\mathcal{F}_i^q$ to two states ρ and σ for $|p - q|$ small. Finally, we use decompositions of density matrices and coupling arguments that allow us to combine the previous two ingredients. The next three subsections cover those aspects in more detail. We also refer to Appendix G where we give a quantum inspired proof of the fixed confidence best arm identification problem for classical multi-armed bandits that shares many ideas with the proof of the quantum result.

3.1 Quantum channel oracles

In this section we consider the query complexity for the decision problem whether we are given an oracle as in (26) or the trivial quantum channel. This is a simplified setting of the more general bandit problem. In fact, it corresponds to the special case that all mean rewards are 0, except for one arm with mean reward p . Most of the work related to oracle query complexity has focused on oracles that act as a unitary map. There is a large body of research work on quantum channels also with a focus on quantum channel discrimination and general lower and upper bounds were derived [37, 38, 39]. However, the application of these general bounds is mostly targeted towards rather simple channels, in particular channels that implement error mechanisms present in quantum devices. Those general results are not directly applicable here, and we will consider bounds targeted at our specific setting. Phrased as above, the setting in this section essentially agrees with [19] where they considered the quantum search problem given an oracle with a certain failure probability $(1 - p)$ (note that we exchanged p and $1 - p$ compared to their convention, which is more natural in our bandit setting). Their main result is that N/p queries are required to identify i . In particular, no speedup compared to classical algorithms is possible.

Below, we will sketch a different proof of their result because we believe it makes the main ideas slightly clearer and because the proof strategy is an important ingredient used in the proof of Theorem 6. The intuition of the proof is that the progress we make is directly related to the decoherence of the state, as measured by its purity. As the purity is lower bounded, this gives us tight control of the progress.

We could make the output space explicit by decomposing the work space $\mathcal{H}_A$ but this is not necessary. As above, we assume that we have oracle access to one of the quantum channels $\mathcal{F}_i(\rho)$ as in (26) which we seek to identify. For convenience we define $\mathcal{F}_0 = \text{Id}$. Then we have the following slightly extended version of Theorem 1 in [19].

Theorem 7. *Any algorithm that can decide whether $\mathcal{F} = \mathcal{F}_i^p$ for some $i > 0$ or $\mathcal{F} = \mathcal{F}_0$ with probability $1 - \delta$ requires at least*

$$T \geq \frac{(1 - p)(1 - 4\delta(1 - \delta))^2}{p} N \quad (30)$$

calls to the channel.

Remark 1. 1. We emphasize again that the original proof in [19] immediately generalizes to the slightly different setting considered here.

2. Note that the bound becomes vacuous for $p \rightarrow 1$. Then we recover the setting of Grover's algorithm, where the well known lower bound scales as $\sqrt{N}$. For $p = 1 - \sqrt{N}^{-1}$ the bound

(30) agrees with the Grover lower bound, so it is possible that a small error probability which depends on the number of arms does not impede the quadratic speedup that is possible in the noiseless case. Similar questions were investigated in [40].

Here we discuss the key elements of the proof, while some calculations are delegated to Appendix D.

Proof. Let us denote by Φ_U the quantum channel acting by the unitary U , i.e., $\Phi_U(\rho) = U\rho U^\dagger$. We consider an algorithm acting by $\Phi_{U_T} \circ \mathcal{F} \circ \Phi_{U_{T-1}} \circ \dots \circ \Phi_{U_1} \circ \mathcal{F} \circ \Phi_{U_0}(\rho_0)$ on some initial state $\rho_0 = |\Omega\rangle\langle\Omega|$ for some pure state Ω . We define

$$\tilde{\rho}_t^i = \Phi_{U_t}(\rho_t^i), \quad \rho_t^i = \mathcal{F}_i(\tilde{\rho}_{t-1}^i), \quad \rho_0^i = \rho_0. \quad (31)$$

Note that for $i = 0$ the state remains pure during the entire algorithm and we denote it by ψ_t and $\tilde{\psi}_t$. We now define

$$R_t^i = \text{tr}(\rho_t^i)^2, \quad (32)$$

$$F_t^i = F(\rho_t^i, \rho_t^0), \quad (33)$$

i.e., the purity of the state and the fidelity (defined by $F(\sigma, \rho) = (\text{tr} \sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})^2$) of the state with respect to the state corresponding to the trivial oracle. Let us remark on the notation, that we always write $\text{tr}(\rho)^2 = \text{tr}(\rho\rho)$ for the trace of the square of an operator while $(\text{tr} \rho)^2$ denotes the squared trace. For a brief overview of properties of the fidelity, we refer to Appendix B. We show that the changes of the two quantities are directly related, i.e., for every increase in distance (loss in fidelity) we have to pay with a loss in purity, i.e., decoherence.

We define the projections $P_i = |i\rangle\langle i| \otimes \text{Id}$. Then it can be shown (see full proof in Appendix D) that

$$F_{t-1}^i - F_t^i \leq 2p\|P_i\psi_t\| \left(\text{tr}(\tilde{\rho}_{t-1}^i - O_i\tilde{\rho}_{t-1}^iO_i^\dagger)^2 \right)^{\frac{1}{2}} \leq 2\|P_i\psi_t\| \sqrt{\frac{p}{1-p}} \sqrt{R_{t-1}^i - R_t^i} \quad (34)$$

Note that the initial values of F and R are $F_0^i = R_0^i = 1$ and $R_t^i \geq 0$. Thus, we conclude using Cauchy-Schwarz

$$\begin{aligned} \sum_i (F_0^i - F_T^i) &= \sum_{i,t} (F_{t-1}^i - F_t^i) \leq \sum_{i,t} 2\|P_i\psi_t\| \sqrt{\frac{p}{1-p}} \sqrt{R_{t-1}^i - R_t^i} \\ &\leq 2\sqrt{\frac{p}{1-p}} \left(\sum_{i,t} \|P_i\psi_t\|^2 \right)^{\frac{1}{2}} \left(\sum_{i,t} R_{t-1}^i - R_t^i \right)^{\frac{1}{2}} \\ &\leq 2\sqrt{\frac{p}{1-p}} \left(\sum_t \|\psi_t\|^2 \right)^{\frac{1}{2}} \left(\sum_i R_0^i - R_T^i \right)^{\frac{1}{2}} \leq 2\sqrt{\frac{p}{1-p}} \sqrt{T} \sqrt{N}. \end{aligned} \quad (35)$$

Finally we use our assumption that the algorithm can decide whether the oracle is trivial $\mathcal{E} = \mathcal{E}_0$ or not with probability $1 - \delta$ for some $\delta < 1/2$. From this we can conclude that the output of the algorithm for oracles $\mathcal{F}^0$ and $\mathcal{F}^i$ must be sufficiently different. Formally, success of the algorithm implies (see (54) and (61) in Appendix B) that for each i

$$1 - 2\delta \leq T(\rho_T^i, \rho_T^0) \leq \sqrt{1 - F(\rho_T^i, \rho_T^0)} \quad (36)$$

where $T(\rho, \sigma) = \frac{1}{2}\|\rho - \sigma\|_{\text{tr}}$ denotes the trace distance. This implies the bound $F_T^i \leq 4\delta(1 - \delta)$. We conclude that

$$2\sqrt{\frac{p}{1-p}} \sqrt{T} \sqrt{N} \geq N(1 - 4\delta(1 - \delta)) \Rightarrow T \geq \frac{N(1-p)(1 - 4\delta(1 - \delta))^2}{p}. \quad (37)$$

□

A natural question suggested by this result and also our main result is whether speedups can be obtained for non-unitary oracles. This question was also posed in [19]. We now show that this is not true (without making additional assumptions). The simplest example is a faulty oracle that indicates its own failure. To define this we consider oracles O_i acting by

$$O_i |i\rangle |c\rangle = -|i\rangle |c \oplus 1\rangle, \quad O_i |j\rangle |c\rangle = |j\rangle |c \oplus 1\rangle \quad \text{for } j \neq i, \quad (38)$$

i.e., the bit flip indicates the marked element i and the change in the second register $|c\rangle \rightarrow |c \oplus 1\rangle$ is used to store that the oracle worked. We consider the faulty versions of these oracles given by

$$\mathcal{F}_i(\rho) = p O_i \rho O_i^\dagger + (1-p)\rho. \quad (39)$$

Then we can obtain the same speedup as with the usual oracle except that we need to correct for the number of times the oracle is not working. This is not in contradiction to the previous result, as this oracle is not of the form defined in (25). In particular, the action of the oracle in (38) is not trivial on $|j\rangle \otimes |c\rangle$.

Theorem 8. *The channel i can be identified with probability at least $1/4$ using $\lfloor \pi/(2\theta p) \rfloor$ queries to the oracle where $\theta = 2 \arcsin(\sqrt{N}^{-1}) \approx 2\sqrt{N}^{-1}$.*

Remark 2. *Note that up to constant factors we need $\sqrt{N}/p$ queries of which typically $\sqrt{N}$ queries work, this is the same scaling as the usual Grover algorithm. As usual, this bound can also be obtained if p is unknown by iteratively increasing the number of iterations in the algorithm described below.*

The proof can be found in Appendix J. While this result is simple and not very surprising, it underlines that it will be difficult to obtain general results showing that no quantum speedup is possible.

3.2 Optimal Fidelity Bounds and Implications for non-adaptive algorithms

In this section we state optimal fidelity bounds when applying the channels $\mathcal{F}_i^p$ and $\mathcal{F}_i^q$ to two states ρ and σ . Then we show how this allows us to derive bounds for non-adaptive algorithms and suboptimal bounds for adaptive algorithms. While those results also directly follow from Theorem 6 we think it is nevertheless helpful to include those as they motivate the result and clarify the scaling.

We start with a lemma that controls the fidelity between applications of the oracle. This result provides a sharp bound that might be of independent interest.

Lemma 1. *Assume that O_i is self-adjoint and unitary and acts as in (25). For density matrices ρ, σ and $p, q \in [\eta, 1-\eta]$ the following bound holds*

$$\sqrt{F}(\mathcal{F}_i^p(\rho), \mathcal{F}_i^q(\sigma)) \geq \sqrt{F}(\rho, \sigma) - \frac{(p-q)^2}{\eta(1-\eta)} \sqrt{\text{tr}(P_i \rho) \text{tr}(P_i \sigma)} \quad (40)$$

where $P_i = |i\rangle \langle i| \otimes \text{Id}$ denotes as before the projection on state $|i\rangle$.

The proof of this lemma can be found in Appendix E. Note the quadratic scaling in $p-q$. The condition that p and q are away from 0 and 1 is necessary because otherwise the optimal bound only scales with $|p-q|$ (note that the bound in Lemma 1 is vacuous as $\eta \rightarrow 0$). These results mirror the classical results about discerning $\text{Ber}(p)$ and $\text{Ber}(q)$ variables. We state one direct consequence of the previous lemma.

Corollary 1. *Let $\mathbf{p}, \mathbf{p}' \in [\eta, 1-\eta]^N$. Then*

$$\sqrt{F}(\mathcal{E}^{\mathbf{p}}(\rho), \mathcal{E}^{\mathbf{p}'}(\sigma)) \geq \sqrt{F}(\rho, \sigma) - \sum_i \frac{(\mathbf{p}_i - \mathbf{p}'_i)^2}{2\eta(1-\eta)} \sqrt{\text{tr}(P_i \rho) \text{tr}(P_i \sigma)}. \quad (41)$$

Proof. We note that $[P_i, O_j] = 0$ for $i \neq j$ and as O_j is unitary we get

$$\text{tr}(P_i \mathcal{F}_j^p(\rho)) = \text{tr}(P_i((1-p)\rho + pO_j \rho O_j^\dagger)) = \text{tr}(P_i((1-p)\rho + p\rho O_j^\dagger O_j)) = \text{tr}(P_i \rho). \quad (42)$$

Then Lemma 1 can be applied inductively to the relation (27) to obtain the claim. $\square$

From this result we conclude that any non-adaptive algorithm requires the same amount of oracle queries as the best classical algorithm.

Corollary 2. Assume that $\mathbf{p}^j$ are as introduced at the beginning of this section with $p_i \in [\eta, 1 - \eta]$ for some $\eta > 0$. Fix a density matrix ρ . We have access to m copies of the state $\mathcal{E}(\rho)$ and it is known that $\mathcal{E}$ is as in (11) where the mean reward vector is in $\{\mathbf{p}^0, \dots, \mathbf{p}^N\}$. If the best arm of $\mathcal{E}$ can be identified with probability at least $1 - \delta$ for some $\delta < 1/2$ then

$$m \geq \frac{\eta(1-\eta)(1-2\sqrt{\delta(1-\delta)})}{16} H(\mathbf{p}) = \Omega(H(\mathbf{p})). \quad (43)$$

The short proof of this result can be found in Appendix E.3.

We can similarly derive a suboptimal bound for adaptive algorithms. Here we lose a $\sqrt{N}$ factor compared to the tight result stated in Theorem 6.

The reason that the techniques used in the proof of Theorem 7 do not provide optimal lower bounds in the more general setting is that the argument uses in an essential way that one of the density matrices is pure. Indeed, we show that the state of the other oracle decoheres with respect to this pure reference state. In the setting here, both density matrices are highly mixed, so it is more subtle to formalize their decoherence.

Corollary 3. Assume that $\mathbf{p}^j$ are as introduced at the beginning of Section G with $p_i \in [\eta, 1 - \eta]$ for some $\eta > 0$. Any quantum algorithm that identifies the best arm when it is known that the reward vector is in $\{\mathbf{p}^0, \dots, \mathbf{p}^N\}$ with probability at least $1 - \delta$ for some $\delta < \frac{1}{2}$ requires at least

$$T \geq \frac{\eta(1-\eta)(1-2\sqrt{\delta(1-\delta)})}{16\sqrt{N}} H(\mathbf{p}) = \Omega(H(\mathbf{p})/\sqrt{N}) \quad (44)$$

calls to the oracle $\mathcal{E}_i(\rho) = \mathcal{E}^{\mathbf{p}^i}(\rho)$.

Proof. The proof is close to the proof of Corollary 2. We assume we are given any algorithm $(\mathcal{E}_i \otimes \text{Id}) \circ \mathcal{E}_{U_T} \circ \dots \circ (\mathcal{E}_i \otimes \text{Id}) \circ \mathcal{E}_{U_1}$ where U_i are arbitrary unitary maps. We denote the state using the oracle i before the t -th invocation of the oracle by ρ_t^i . Using the invariance of the fidelity under unitary maps and Corollary 1 we bound

$$\sqrt{F}(\rho_T^i, \rho_T^0) \geq 1 - \sum_t \frac{4\Delta_i^2}{\eta(1-\eta)} \sqrt{\text{tr}(P_i \rho_t^i) \text{tr}(P_i \rho_t^0)}. \quad (45)$$

As the algorithm can discern the oracles $\mathcal{E}_i$ and $\mathcal{E}_0$ with probability at least $1 - \delta$ we have the bound

$$1 - 2\delta \leq T(\rho_T^i, \rho_T^0) \leq \sqrt{1 - F(\rho_T^i, \rho_T^0)}. \quad (46)$$

We conclude that

$$2\sqrt{\delta(1-\delta)} \geq \sqrt{F}(\rho_T^i, \rho_T^0) \geq 1 - \sum_t \frac{4\Delta_i^2}{\eta(1-\eta)} \sqrt{\text{tr}(P_i \rho_t^i) \text{tr}(P_i \rho_t^0)}. \quad (47)$$

Thus we get using $\sum_i \text{tr}(P_i \rho_0^i) = 1$ and $\text{tr}(P_i \rho_t^i) \leq 1$

$$\begin{aligned} \frac{\eta(1-\eta)(1-2\sqrt{\delta(1-\delta)})}{4} \sum_{i \geq 2} \Delta_i^{-2} &\leq \sum_{t,i} \sqrt{\text{tr}(P_i \rho_t^i) \text{tr}(P_i \rho_t^0)} \\ &\leq \left(\sum_{t,i} \text{tr}(P_i \rho_t^i) \right)^{\frac{1}{2}} \left(\sum_{t,i} \text{tr}(P_i \rho_t^0) \right)^{\frac{1}{2}} \leq \sqrt{NT} \sqrt{T}. \end{aligned} \quad (48)$$

Reorganizing this we obtain the claimed result. $\square$

3.3 Coupling Arguments and Density Matrix Decompositions

Roughly, we have seen so far two ingredients to prove lower bounds. First, we considered in Section 3.1 the relation between purity and fidelity and used this to show that no $\sqrt{N}$ Grover type speedup is possible. Then, in Section 3.2 we derived optimal bounds of the fidelity loss giving the right scaling in Δ_i^{-2} but which is not sufficient to exclude the Grover search speedup, see Corollary 3. Now, we show how those two lines can be combined to give optimal bounds. Let us denote by ρ_t^j the state of the algorithm after t steps when the oracle $\mathcal{E}^j$ is used. As in Section 3.1 we would like to use that ρ_t^j decoheres with respect to ρ_t^0 . However, this is more difficult to formalize as ρ_t^0 now also is a highly mixed state, and we did not succeed in finding a suitable quantity that captures this. Thus, we decompose ρ_t^j into a mixture of pure states (not its spectral decomposition, but we split each pure state in two whenever applying an oracle $\mathcal{F}_i^p$). We also define a corresponding decomposition of ρ_t^j so that the reasoning of Section 3.1 can be applied term by term. Thus, our proof relies on a coupling argument, a technique that is standard in the theory of stochastic processes but not so much in quantum information theory. For an overview of this technique, we refer to [41].

To implement this, we need a strengthened version of Lemma 1 which not only has the optimal scaling but in addition provides a decomposition $\rho = p_0\rho_0 + p_1\rho_1$ and $\sigma = q_0\sigma_0 + q_1\sigma_1$ such that the lower bound also applies to $\sqrt{p_0q_0}\sqrt{F}(\rho_0, \sigma_0) + \sqrt{p_1q_1}\sqrt{F}(\rho_1, \sigma_1)$ (which lower bounds the fidelity of the mixture). Let us state the key lemma.

Lemma 2. *Consider a density matrix ρ and a pure state ψ . Consider a unitary and self-adjoint map U and the channel $\mathcal{E}_U^p$ defined by $\mathcal{E}_U^p(\rho) = pU\rho U^\dagger + (1-p)\rho$. Let $0 < \eta < 1/2$ and $p, q \in [\eta, 1-\eta]$. Then there are density matrices $\rho_0 = \rho$ and $\rho_1 = U\rho U^\dagger$ and pure states ψ_0 and ψ_1 and a real number q' such that*

$$\mathcal{E}_U^p(\rho) = p\rho_1 + (1-p)\rho_0, \quad \mathcal{E}_U^q(|\psi\rangle\langle\psi|) = q'|\psi_1\rangle\langle\psi_1| + (1-q')|\psi_0\rangle\langle\psi_0| \quad (49)$$

and the following bound holds. Let $S = \sqrt{F}(\rho, |\psi\rangle\langle\psi|) = \sqrt{\langle\psi|\rho|\psi\rangle}$ denote the initial fidelity. Then we have

$$\begin{aligned} \sqrt{F}(\mathcal{E}_U^p(\rho), \mathcal{E}_U^q(|\psi\rangle\langle\psi|)) &\geq \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, |\psi_0\rangle\langle\psi_0|) + \sqrt{pq'}\sqrt{F}(\rho_1, |\psi_1\rangle\langle\psi_1|) \\ &\geq S - \frac{(p-q)^2|\langle\psi, (U\rho U^\dagger - \rho)\psi\rangle|}{2\eta S} - \frac{(p-q)^2|\text{Re}(\langle\psi, U\rho\psi\rangle) - (\langle\psi, \rho\psi\rangle)|^2}{8\eta^2 S^3}. \end{aligned} \quad (50)$$

Remark 3. *Let us give some explanation regarding this lemma.*

1. Note that for $p = q$ we recover the result that quantum channels can only increase the fidelity for our specific channel.
2. The problem that this lemma solves is that we need a bound on the loss in fidelity that has, firstly, the optimal quadratic rate in $(p - q)$, secondly, the bound needs to have a form that allows to exploit the specific structure of the oracles O_i which act non-trivially only on a small subspace, and, thirdly, we later want to use that the density matrices ρ decohere with respect to the state ψ . The last point will become clearer in the proof of Theorem 6 in Appendix H, but note that the expression $\langle\psi, (U\rho U^\dagger - \rho)\psi\rangle$ appeared already in the proof of Theorem 7 (see Equation (34)) which indicates that similar arguments can be applied. We remark that Lemma 1 above already satisfied the first two requirements, but the lack of the third requirement allowed us to only show the suboptimal bound in Corollary 3. It is also quite straightforward to satisfy the second and the third requirement with the suboptimal rate $|p - q|$. But this also gives only a suboptimal bound.
3. The second error term does not require all desiderata outlined above, but it is of higher order (note the extra square) which is sufficient to control it.

The proof of this lemma can be found in Appendix F, there we also state a slight generalization that is used in the proof of Theorem 6 in Appendix H.

4 Discussion

In this work we investigated quantum algorithms for multi-armed bandit problems. It was shown earlier in [17] that quantum algorithms for best arm identification with fixed confidence can have a quadratic speedup compared to their classical counterparts. This result is based on the assumption that the arms and the randomness of the rewards of the arms can be both queried in superposition. These assumptions are reasonable in the setting of empirical risk minimization, where we can evaluate loss values in superposition. However, there are many settings, e.g., motivated by quantum sensing, where it might not be possible to query the internal randomness of the bandits in superposition. Instead, every pull of the lever returns a single random reward. We then show that in this setting no speedup compared to classical algorithms is possible.

This highlights that classical randomness pose a major challenge for quantum algorithms. In our case the randomness of the rewards even prevent Grover type speedups of the search problem that one would naively expect to arise from the search part of the multi-armed bandit problem. Note that such a speedup is possible in the intermediate regime that we considered. When we can select the state of the internal randomness of the oracle (but not query it in superposition) the statistical complexity of the problem remains the same, but we can search through the arms faster, providing some speedup.

There are many open questions related to this work and we will briefly mention two. Firstly, classical randomness appears frequently in different settings. This has been studied a lot in the context of noise channels but not so much in other contexts, e.g., machine learning.

Secondly, our proofs proceed by directly controlling the fidelity between the quantum states when invoking different oracles. From a methodological side, it would be interesting to see to what degree the well-known strategies to lower bound the query complexity like the polynomial [42] or the adversarial method [43] extend to non-unitary oracles.

References

- [1] P. Shor. “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer”. *SIAM Journal on Computing* **26**, 1484–1509 (1997).
- [2] Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. “Quantum algorithm for linear systems of equations”. *Phys. Rev. Lett.* **103**, 150502 (2009).
- [3] Lov K. Grover. “A fast quantum mechanical algorithm for database search”. In Gary L. Miller, editor, *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing*, Philadelphia, Pennsylvania, USA, May 22-24, 1996. Pages 212–219. ACM (1996).
- [4] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. “Quantum machine learning”. *Nature* **549**, 195–202 (2017).
- [5] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. “Quantum support vector machine for big data classification”. *Physical Review Letters* **113** (2014).
- [6] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. “Quantum principal component analysis”. *Nature Physics* **10**, 631–633 (2014).
- [7] Iordanis Kerenidis and Anupam Prakash. “Quantum recommendation systems”. *CoRRabs/1603.08675* (2016). [arXiv:1603.08675](https://arxiv.org/abs/1603.08675).
- [8] Esma Aïmeur, Gilles Brassard, and Sébastien Gambs. “Quantum speed-up for unsupervised learning”. *Mach. Learn.* **90**, 261–287 (2013).
- [9] Iordanis Kerenidis, Jonas Landman, Alessandro Luongo, and Anupam Prakash. “q-means: A quantum algorithm for unsupervised machine learning”. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Volume 32. Curran Associates, Inc. (2019). url: <https://proceedings.neurips.cc/paper/2019/file/16026d60ff9b54410b3435b403afd226-Paper.pdf>.
- [10] Daoyi Dong, Chunlin Chen, Han-Xiong Li, and Tzyh Jong Tarn. “Quantum reinforcement learning”. *IEEE Trans. Syst. Man Cybern. Part B* **38**, 1207–1220 (2008).

- [11] Carlo Ciliberto, Mark Herbster, Alessandro Davide Ialongo, Massimiliano Pontil, Andrea Rocchetto, Simone Severini, and Leonard Wossnig. “Quantum machine learning: a classical perspective”. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **474**, 20170551 (2018).
- [12] William R. Thompson. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. *Biometrika* **25**, 285–294 (1933).
- [13] Herbert Robbins. “Some aspects of the sequential design of experiments”. *Bulletin of the American Mathematical Society* **58**, 527 – 535 (1952). url: <https://www.ams.org/journals/bull/1952-58-05/S0002-9904-1952-09620-8/S0002-9904-1952-09620-8.pdf>.
- [14] T.L Lai and Herbert Robbins. “Asymptotically efficient adaptive allocation rules”. *Advances in Applied Mathematics* **6**, 4–22 (1985).
- [15] Sébastien Bubeck and Nicolò Cesa-Bianchi. “Regret analysis of stochastic and nonstochastic multi-armed bandit problems”. *Found. Trends Mach. Learn.* **5**, 1–122 (2012).
- [16] T. Lattimore and C. Szepesvári. “Bandit algorithms”. *Cambridge University Press*. (2020).
- [17] Daochen Wang, Xuchen You, Tongyang Li, and Andrew M. Childs. “Quantum exploration algorithms for multi-armed bandits”. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 10102–10110 (2021).
- [18] Stefano Pirandola, Riccardo Laurenza, Cosmo Lupo, and Jason L. Pereira. “Fundamental limits to quantum channel discrimination”. *npj Quantum Information* **5**, 50 (2019).
- [19] Oded Regev and Liron Schiff. “Impossibility of a quantum speed-up with a faulty oracle”. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfssdóttir, and Igor Walukiewicz, editors, *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7–11, 2008, Proceedings, Part I: Track A: Algorithms, Automata, Complexity, and Games. Volume 5125 of Lecture Notes in Computer Science*, pages 773–781. Springer (2008).
- [20] Tze Leung Lai. “Adaptive Treatment Allocation and the Multi-Armed Bandit Problem”. *The Annals of Statistics* **15**, 1091 – 1114 (1987).
- [21] Peter Auer. “Using confidence bounds for exploitation-exploration trade-offs”. *J. Mach. Learn. Res.* **3**, 397–422 (2003). url: <https://jmlr.org/papers/v3/auer02a.html>.
- [22] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. “PAC bounds for multi-armed bandit and markov decision processes”. In Jyrki Kivinen and Robert H. Sloan, editors, *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8–10, 2002, Proceedings. Volume 2375 of Lecture Notes in Computer Science*, pages 255–270. Springer (2002).
- [23] Shie Mannor and John N. Tsitsiklis. “The sample complexity of exploration in the multi-armed bandit problem”. *J. Mach. Learn. Res.* **5**, 623–648 (2004). url: <http://jmlr.org/papers/volume5/mannor04b/mannor04b.pdf>.
- [24] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. “Best arm identification in multi-armed bandits”. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27–29, 2010*. Pages 41–53. Omnipress (2010). url: <https://www.learningtheory.org/colt2010/conference-website/papers/59Audibert.pdf>.
- [25] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. “Best arm identification: A unified approach to fixed budget and fixed confidence”. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States. Pages 3221–3229*. (2012). url: <https://proceedings.neurips.cc/paper/2012/hash/8b0d268963dd0cfb808aac48a549829f-Abstract.html>.

- [26] Aurélien Garivier and Emilie Kaufmann. “Optimal best arm identification with fixed confidence”. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, 29th Annual Conference on Learning Theory. Volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027. Columbia University, New York, New York, USA (2016). PMLR. url: <https://proceedings.mlr.press/v49/garivier16a.html>.
- [27] Kevin G. Jamieson, Matthew Malloy, Robert D. Nowak, and Sébastien Bubeck. “lil’ UCB : An optimal exploration algorithm for multi-armed bandits”. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*. Volume 35 of *JMLR Workshop and Conference Proceedings*, pages 423–439. JMLR.org (2014). url: <http://proceedings.mlr.press/v35/jamieson14.html>.
- [28] Lijie Chen, Jian Li, and Mingda Qiao. “Towards instance optimal bounds for best arm identification”. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*. Volume 65 of *Proceedings of Machine Learning Research*, pages 535–592. PMLR (2017). url: <http://proceedings.mlr.press/v65/chen17b.html>.
- [29] Balthazar Casalé, Giuseppe Di Molfetta, Hachem Kadri, and Liva Ralaivola. “Quantum bandits”. *Quantum Mach. Intell.* **2**, 1–7 (2020).
- [30] M. A. Nielsen and I. L. Chuang. “Quantum Computation and quantum Information”. *Cambridge University Press*. (2000).
- [31] Zongqi Wan, Zhijie Zhang, Tongyang Li, Jialin Zhang, and Xiaoming Sun. “Quantum multi-armed bandits and stochastic linear bandits enjoy logarithmic regrets”. *CoRRabs/2205.14988* (2022). [arXiv:2205.14988](https://arxiv.org/abs/2205.14988).
- [32] Patrick Rebentrost, Yassine Hamoudi, Maharshi Ray, Xin Wang, Siyi Yang, and Miklos Santha. “Quantum algorithms for hedging and the learning of ising models”. *Phys. Rev. A* **103**, 012418 (2021).
- [33] Andris Ambainis. “Variable time amplitude amplification and quantum algorithms for linear algebra problems”. In Christoph Dürr and Thomas Wilke, editors, 29th International Symposium on Theoretical Aspects of Computer Science, STACS 2012, February 29th - March 3rd, 2012, Paris, France. *Volume 14 of LIPIcs*, pages 636–647. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2012).
- [34] C. L. Degen, F. Reinhard, and P. Cappellaro. “Quantum sensing”. *Rev. Mod. Phys.* **89**, 035002 (2017).
- [35] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean. “Quantum advantage in learning from experiments”. *Science* **376**, 1182–1186 (2022).
- [36] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. “Advances in quantum metrology”. *Nature Photonics* **5**, 222–229 (2011).
- [37] A Acín. “Statistical distinguishability between unitary operations”. *Physical review letters* **87**, 177901 (2001).
- [38] Stefano Pirandola, Riccardo Laurenza, Cosmo Lupo, and Jason L. Pereira. “Fundamental limits to quantum channel discrimination”. *npj Quantum Information* **5**, 50 (2019).
- [39] Quntao Zhuang and Stefano Pirandola. “Ultimate limits for multiple quantum channel discrimination”. *Phys. Rev. Lett.* **125**, 080505 (2020).
- [40] Neil Shenvi, Kenneth R. Brown, and K. Birgitta Whaley. “Effects of a random noisy oracle on search algorithm complexity”. *Phys. Rev. A* **68**, 052313 (2003).
- [41] T. Lindvall. “Lectures on the coupling method”. *Dover Books on Mathematics*. Dover Publications. (2012).
- [42] Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald de Wolf. “Quantum lower bounds by polynomials”. *J. ACM* **48**, 778–797 (2001).

- [43] Andris Ambainis. “Quantum lower bounds by quantum arguments”. *J. Comput. Syst. Sci.* **64**, 750–767 (2002).
- [44] Yeong-Cherng Liang, Yu-Hao Yeh, Paulo E M F Mendonça, Run Yan Teh, Margaret D Reid, and Peter D Drummond. “Quantum fidelity measures for mixed states”. *Reports on Progress in Physics* **82**, 076001 (2019).

A Overview of the Appendix

In this appendix we collect all the proofs of the results in the paper. We start with a brief review of distance measure of quantum states in Section B where we list some well-known results on distance measures of quantum states for reference. Then, in Section C, we collect some simple auxiliary results that will be used in the proofs later on.

The following sections of this appendix then roughly follow the outline given in Section 3 in the main text and add the missing proofs. In particular, in Appendix D we give the complete proof of Theorem 7, in Appendix E we first prove a simpler version of Lemma 1 stated in Lemma 8, and then give the proofs of Lemma 1 and Corollary 1. In Appendix F we prove Lemma 2 and provide a small technical extension that is used in the proof of the main result. In Appendix G we give a complete quantum inspired proof of the lower bound for classical bandits. This proof clarifies some of the choices in the proof of Theorem 6 which can be found along with the more precise statement of the result in Theorem 10 in Appendix H. Finally, the Appendices I, J and K contain the proofs of Theorem 4, Theorem 8 and Theorem 5 respectively. The proof of Theorem 4 shares some ideas with, e.g., the quantum inspired proof for the classical bandits, the other two proofs use different ideas than the remaining results of this paper and are independent of the other sections.

B A Brief Review of Distance Measures for Quantum States

For the convenience of the reader we give a brief review of distance measures for quantum states. Textbooks on quantum computation, e.g., [30] discuss this thoroughly. For a review on fidelities we refer to [44]. We consider the trace distance which is defined by

$$T(\rho, \sigma) = \frac{1}{2} \|\rho - \sigma\|_{\text{tr}} \quad (51)$$

where the norm indicates the trace norm defined by $\|A\|_{\text{tr}} = \text{tr}(\sqrt{A^\dagger A})$. It has the property that for any POVM $\{E_i\}$ the outcome probabilities

$$p_i = \text{tr}(E_i \rho), \quad q_i = \text{tr}(E_i \sigma) \quad (52)$$

the total variation distance between the probability vectors p_i and q_i satisfy

$$\frac{1}{2} \sum_i |p_i - q_i| \leq T(\rho, \sigma). \quad (53)$$

The Helmsstrom measurement gives the optimal discrimination probability of two states and has success probability

$$p_{\text{success}} = \frac{1}{2} + \frac{1}{2} T(\rho, \sigma). \quad (54)$$

For many applications the fidelity is a more useful distance measure to obtain optimal bounds. It is defined by

$$\sqrt{F}(\rho, \sigma) = \text{tr} \left(\sqrt{\rho^{\frac{1}{2}} \sigma \rho^{\frac{1}{2}}} \right) = \|\rho^{\frac{1}{2}} \sigma^{\frac{1}{2}}\|_{\text{tr}}. \quad (55)$$

Some authors instead call the square of this expression the fidelity, and to clarify our convention we added the square root. As suggested by the notation we set $F = \sqrt{F}^2$. We collect some properties of the fidelity that we will use frequently.

1. For a density matrix ρ and a pure state ψ the fidelity is given by

$$\sqrt{F}(|\psi\rangle\langle\psi|, \rho) = \sqrt{\langle\psi, \rho\psi\rangle}. \quad (56)$$

2. For any density matrices ρ, σ and a quantum channel $\mathcal{E}$ the following bound holds

$$\sqrt{F}(\rho, \sigma) \leq \sqrt{F}(\mathcal{E}(\rho), \mathcal{E}(\sigma)). \quad (57)$$

3. If the quantum channel $\mathcal{E}$ acts by a unitary matrix, i.e., $\mathcal{E}(\rho) = U\rho U^\dagger$ then

$$\sqrt{F}(\rho, \sigma) = \sqrt{F}(\mathcal{E}(\rho), \mathcal{E}(\sigma)). \quad (58)$$

4. The fidelity is strongly concave

$$\sqrt{F}\left(\sum_i p_i \rho_i, \sum_i q_i \sigma_i\right) \geq \sum_i \sqrt{p_i q_i} \sqrt{F}(\rho_i, \sigma_i). \quad (59)$$

This directly implies concavity

$$\sqrt{F}\left(\sum_i p_i \rho_i, \sum_i p_i \sigma_i\right) \geq \sum_i p_i \sqrt{F}(\rho_i, \sigma_i). \quad (60)$$

5. Fidelity and trace distance are related by

$$1 - \sqrt{F}(\rho, \sigma) \leq T(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)}. \quad (61)$$

Those properties can be proved using Uhlmann's Theorem which states that

$$\sqrt{F}(\rho, \sigma) = \max_{\varphi, \psi} \langle \varphi, \psi \rangle \quad (62)$$

where the maximum is over all purifications ψ and φ of ρ and σ , respectively. We use this result to bound the fidelity change of certain quantum operations (see Lemma 1 and 2).

C Auxiliary Lemmas

Here we include simple, mostly algebraic lemmas that are used in the proof of Theorem 7 and in the proofs in Appendix E. The first lemma is a simple Cauchy-Schwarz estimate that in addition exploits invariant subspaces of an operator O . It is used in the proof of Theorem 7. Recall our convention that $\text{tr}(\rho)^2 = \text{tr} \rho \rho$.

Lemma 3. *Let O be a unitary operator and P a self-adjoint orthogonal projections such that $(\text{Id} - P)O = \text{Id} - P$, i.e., O acts trivially on the orthogonal complement of the projection P . Then, for any vector $|\varphi\rangle$ and density matrix σ the bound*

$$|\langle \varphi | (O\sigma O^\dagger - \sigma) | \varphi \rangle| \leq 2\|P\varphi\| \|\varphi\| \left(\text{tr} (O\sigma O^\dagger - \sigma)^2 \right)^{\frac{1}{2}} \quad (63)$$

holds.

Proof. We define $Q = \text{Id} - P$. Then we have $QO = Q$. By assumption, we can decompose

$$\begin{aligned} \sigma - O\sigma O^\dagger &= (P + Q)(\sigma - O\sigma O^\dagger)(P + Q) \\ &= (\sigma - O\sigma O^\dagger)P + P(\sigma - O\sigma O^\dagger)Q \end{aligned} \quad (64)$$

where we used $QO\sigma O^\dagger Q = Q\sigma Q$. Using Cauchy-Schwarz for the Hilbert-Schmidt scalar product we can bound for $M = M^\dagger$

$$|\langle \varphi_1 | M | \varphi_2 \rangle| = |\text{tr}(|\varphi_2\rangle\langle\varphi_1| M)| \leq (\text{tr}(|\varphi_2\rangle\langle\varphi_1| \langle\varphi_1|\varphi_1\rangle \langle\varphi_2|) \text{tr} M^2)^{\frac{1}{2}} = \|\varphi_1\| \cdot \|\varphi_2\| (\text{tr} M^2)^{\frac{1}{2}}. \quad (65)$$

Using (64) and (65) we can continue to estimate

$$\begin{aligned} |\langle \varphi | (O\sigma O^\dagger - \sigma) \varphi \rangle| &\leq (\|P\varphi\| \|\varphi\| + \|P\varphi\| \|Q\varphi\|) \left(\text{tr} (O\sigma O^\dagger - \sigma)^2 \right)^{\frac{1}{2}} \\ &\leq 2\|P\varphi\| \cdot \|\varphi\| \left(\text{tr} (O\sigma O^\dagger - \sigma)^2 \right)^{\frac{1}{2}}. \end{aligned} \quad (66)$$

This ends the proof. $\square$

The next lemma states two simple algebraic bounds.

Lemma 4. *For $p, q \in [c, 1 - c]$ the following bounds hold*

$$\sqrt{(1-p)(1-q)} + \sqrt{pq} \geq 1 - \frac{|p-q|^2}{4c(1-c)}, \quad (67)$$

$$\left| \sqrt{(1-p)q} - \sqrt{(1-q)p} \right| \leq \frac{|p-q|}{2\sqrt{c(1-c)}}. \quad (68)$$

Proof. We first consider the second inequality. We note that $|\sqrt{x} - \sqrt{y}| \leq |x - y|/(\sqrt{x} + \sqrt{y})$ and thus

$$\left| \sqrt{(1-p)q} - \sqrt{(1-q)p} \right| \leq \frac{|p-q|}{\sqrt{p(1-q)} + \sqrt{q(1-p)}} \leq \frac{|p-q|}{2\sqrt[4]{p(1-q)q(1-p)}} \leq \frac{|p-q|}{2\sqrt{c(1-c)}} \quad (69)$$

where we used the arithmetic geometric mean inequality in the middle step. To prove the first bound, we note that

$$\left(\sqrt{(1-p)(1-q)} + \sqrt{pq} \right)^2 + \left(\sqrt{(1-p)q} - \sqrt{(1-q)p} \right)^2 = 1 \quad (70)$$

This implies

$$\begin{aligned} \sqrt{(1-p)(1-q)} + \sqrt{pq} &= \sqrt{1 - \left(\sqrt{(1-p)q} - \sqrt{(1-q)p} \right)^2} \\ &\geq 1 - \left(\sqrt{(1-p)q} - \sqrt{(1-q)p} \right)^2 \geq 1 - \frac{|p-q|^2}{4c(1-c)}. \end{aligned} \quad (71)$$

$\square$

The following simple lemma provides a lower bound on the square root that is used in the proof of Lemma 2 below.

Lemma 5. *Let s, t be real numbers such that $s + t \geq -1$. Then the bound*

$$\sqrt{1 + s + t} \geq 1 - |s| + \frac{t}{2} - \frac{t^2}{2} \quad (72)$$

holds.

Proof. First, we note that elementary manipulations show that for all $t \in \mathbb{R}$ the bound

$$\sqrt{\max(1+t, 0)} \geq 1 + \frac{t}{2} - \frac{t^2}{2} \quad (73)$$

holds. We now consider $s > 0$. In this case, we can conclude

$$\sqrt{1 + s + t} \geq \sqrt{\max(1+t, 0)} \geq 1 + \frac{t}{2} - \frac{t^2}{2} \geq 1 + \frac{t}{2} - \frac{t^2}{2} - |s|. \quad (74)$$

Note that $x - y = (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) > (\sqrt{x} - \sqrt{y})$ if $x > 1$ and $x > y$. This implies for $s < 0$ and $t \geq 0$ that

$$\sqrt{1 + t + s} \geq \sqrt{1 + t} - |s| \geq 1 + \frac{t}{2} - \frac{t^2}{2} - |s|. \quad (75)$$

Finally, we consider the case $t, s < 0$ where we get from (73)

$$\sqrt{1+t+s} \geq 1 + \frac{t+s}{2} - \frac{(t+s)^2}{2} = 1 + \frac{t}{2} - \frac{t^2}{2} + s \left(\frac{1-t-s}{2} \right) \geq 1 + \frac{t}{2} - \frac{t^2}{2} - |s| \quad (76)$$

using $-t-s \leq 1$ in the last step. $\square$

We also state a simple fact on the relation of partial trace and operators.

Lemma 6. *Let ρ be an operator on the system $Q \otimes R$. Let O be a linear operator on Q . Then*

$$\text{tr}_R((O \otimes \text{Id})\rho) = O \text{tr}_R(\rho). \quad (77)$$

Proof. By linearity it is sufficient to consider $\rho = S \otimes T$. But then

$$\text{tr}_R((O \otimes \text{Id})\rho) = \text{tr}_R(OS \otimes T) = \text{tr}(T)OS = O \text{tr}_R(\rho). \quad (78)$$

$\square$

Finally, we state an elementary result about the reward vectors $\mathbf{p}^j$.

Lemma 7. *Let $\mathbf{p}^j \in [0, 1]^N$ be reward vectors as defined at the beginning of Section 3, i.e., there is $\mathbf{p} = (p_0, \dots, p_N) \in [0, 1]^{N+1}$ with $p_0 > p_1 > p_2 \geq \dots \geq p_N$ and then $\mathbf{p}_i^j = p_i$ for $1 \leq i \neq j$ and $\mathbf{p}_j^j = p_0$. Assume that $p_0 - p_1 = p_1 - p_2$ and let $\Delta_i = p_0 - p_i$. Then for all $1 \leq j \leq N$*

$$\frac{1}{4}H(\mathbf{p}^0) \leq H(\mathbf{p}^j) \leq 2H(\mathbf{p}^0) \quad (79)$$

Proof. We note that for $j \geq 1$

$$\begin{aligned} H(\mathbf{p}^j) &= \sum_{i \neq j} (p_0 - p_i)^{-2} \leq \sum_{i \geq 1} \Delta_i^{-2} = (p_0 - p_1)^{-2} + \sum_{i > 1} (p_0 - p_i)^{-2} \\ &\leq 2 \sum_{i > 1} (p_1 - p_i)^2 = 2H(\mathbf{p}^0) \end{aligned} \quad (80)$$

where we used $p_0 - p_1 = p_1 - p_2$ in the last step. Similarly, we obtain

$$\begin{aligned} \sum_{i \geq 1} \Delta_i^{-2} &\geq H(\mathbf{p}^j) = \sum_{i \neq j} (p_0 - p_i)^{-2} = \sum_{i \neq j} (p_1 - p_i + \Delta_1)^{-2} \\ &\geq \sum_{i > 1} (p_1 - p_i + \Delta_1)^{-2} \geq \frac{1}{4} \sum_{i > 1} (p_1 - p_i)^2 = \frac{1}{4}H(\mathbf{p}^0). \end{aligned} \quad (81)$$

Here we used in the third step that $p_1 - p_1 + \Delta_1 \leq p_1 - p_i + \Delta_1$ for any $i \geq 1$ and $p_1 - p_i + \Delta_1 \leq 2(p_1 - p_i)$ in the following inequality. This ends the proof. $\square$

D Proof of Theorem 7

In this section we provide the missing parts of the proof of Theorem 7. To have a complete proof in one place, we repeat the parts that are already contained in the main part of the paper.

Proof of Theorem 7. Denote by Φ_U the quantum channel acting by the unitary U , i.e., $\Phi_U(\rho) = U\rho U^\dagger$. We consider an algorithm acting by $\Phi_{U_T} \circ \mathcal{F} \circ \Phi_{U_{T-1}} \circ \dots \circ \Phi_{U_1} \circ \mathcal{F} \circ \Phi_{U_0}(\rho_0)$ on some initial state $\rho_0 = |\Omega\rangle\langle\Omega|$ for some pure state Ω . We define

$$\tilde{\rho}_t^i = \Phi_{U_t}(\rho_t^i), \quad \rho_t^i = \mathcal{F}_i(\tilde{\rho}_{t-1}^i), \quad \rho_0^i = \rho_0. \quad (82)$$

Note that for $i = 0$ the state remains pure during the entire algorithm, and we denote it by ψ_t and $\tilde{\psi}_t$. We now define

$$R_t^i = \text{tr}(\rho_t^i)^2, \quad (83)$$

$$F_t^i = F(\rho_t^i, \rho_t^0), \quad (84)$$

i.e., the purity of the state and the fidelity (defined by $F(\sigma, \rho) = (\text{tr} \sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})^2$) of the state with respect to the state corresponding to the trivial oracle. As explained in the main part of this paper, we will now show that any decrease in the fidelity must be paid for with a decrease in purity of ρ_t^j . Fidelity and purity are invariant under unitary maps and therefore $R_t^i = \text{tr}(\tilde{\rho}_t^i)^2$ and $F_t^i = F(\tilde{\rho}_t^i, \tilde{\rho}_t^0)$. We control, using that O_i is unitary,

$$\begin{aligned}
R_{t-1}^i - R_t^i &= \text{tr}(\tilde{\rho}_{t-1}^i)^2 - \text{tr}(\rho_t^i)^2 \\
&= \text{tr}(\tilde{\rho}_{t-1}^i)^2 - \text{tr}(pO_i\tilde{\rho}_{t-1}^iO_i^\dagger + (1-p)\tilde{\rho}_{t-1}^i)^2 \\
&= \text{tr}(\tilde{\rho}_{t-1}^i)^2 - (p^2 + (1-p)^2)\text{tr}(\tilde{\rho}_{t-1}^i)^2 - 2p(1-p)\text{tr}(O_i\tilde{\rho}_{t-1}^iO_i^\dagger\tilde{\rho}_{t-1}^i) \\
&= 2p(1-p)\left(\text{tr}(\tilde{\rho}_{t-1}^i)^2 - \text{tr}(O_i\tilde{\rho}_{t-1}^iO_i^\dagger\tilde{\rho}_{t-1}^i)\right) \\
&= p(1-p)\text{tr}(\tilde{\rho}_{t-1}^i - O_i\tilde{\rho}_{t-1}^iO_i^\dagger)^2
\end{aligned} \tag{85}$$

Similarly we estimate the change in fidelity using that ρ_0^t is a pure state and $\tilde{\psi}_{t-1} = \psi_t$ (because $\mathcal{E}_0 = \text{Id}$)

$$\begin{aligned}
F_{t-1}^i - F_t^i &= F(\tilde{\rho}_{t-1}^i, \tilde{\rho}_{t-1}^0) - F(\rho_t^i, \rho_t^0) \\
&= \langle \tilde{\psi}_{t-1}, \tilde{\rho}_{t-1}^i \tilde{\psi}_{t-1} \rangle - \langle \psi_t, \rho_t^i \psi_t \rangle \\
&= \langle \psi_t, (\tilde{\rho}_{t-1}^i - (1-p)\tilde{\rho}_{t-1}^i - pO_i\tilde{\rho}_{t-1}^iO_i^\dagger)\psi_t \rangle \\
&= p\langle \psi_t, (\tilde{\rho}_{t-1}^i - O_i\tilde{\rho}_{t-1}^iO_i^\dagger)\psi_t \rangle.
\end{aligned} \tag{86}$$

We now relate the change of R_t and F_t . Suppose that there is an orthogonal projection P and a unitary O such that $(\text{Id} - P)O = \text{Id} - P$, i.e., O acts trivially on the complement of the image of P . Then Lemma 3 in Appendix C establishes the bound

$$|\langle \varphi | (O\sigma O^\dagger - \sigma) | \varphi \rangle| \leq 2\|P\varphi\|\|\varphi\| \left(\text{tr}(O\sigma O^\dagger - \sigma)^2 \right)^{\frac{1}{2}}. \tag{87}$$

We define projections $P_i = |i\rangle\langle i| \otimes \text{Id}$. Note that, by definition of O_i we have $(\text{Id} - P_i)O_i = \text{Id} - P_i$. Applying Lemma 3, i.e., (87) (with $P = P_i$, $\sigma = \tilde{\rho}_{t-1}^i$, $\varphi = \psi_t$) we can continue to estimate (86) as follows

$$\begin{aligned}
F_{t-1}^i - F_t^i &\leq 2\|P_i\psi_t\| \left(\text{tr}(\tilde{\rho}_{t-1}^i - O_i\tilde{\rho}_{t-1}^iO_i^\dagger)^2 \right)^{\frac{1}{2}} \\
&\leq 2\|P_i\psi_t\| \sqrt{\frac{p}{1-p}} \sqrt{R_{t-1}^i - R_t^i}
\end{aligned} \tag{88}$$

Note that the initial values of F and R are $F_0^i = R_0^i = 1$ and $R_t^i \geq 0$. Thus we can conclude

$$\begin{aligned}
\sum_i (F_0^i - F_T^i) &= \sum_{i,t} (F_{t-1}^i - F_t^i) \\
&\leq \sum_{i,t} 2\|P_i\psi_t\| \sqrt{\frac{p}{1-p}} \sqrt{R_{t-1}^i - R_t^i} \\
&\leq 2\sqrt{\frac{p}{1-p}} \left(\sum_{i,t} \|P_i\psi_t\|^2 \right)^{\frac{1}{2}} \left(\sum_{i,t} R_{t-1}^i - R_t^i \right)^{\frac{1}{2}} \\
&\leq 2\sqrt{\frac{p}{1-p}} \left(\sum_t \|\psi_t\|^2 \right)^{\frac{1}{2}} \left(\sum_i R_0^i - R_T^i \right)^{\frac{1}{2}} \\
&\leq 2\sqrt{\frac{p}{1-p}} \sqrt{T} \sqrt{N}.
\end{aligned} \tag{89}$$

Finally we use our assumption that the algorithm is able to decide whether the oracle is trivial $\mathcal{E} = \mathcal{E}_0$ or not with probability $1 - \delta$ for some $\delta < 1/2$. From here we can conclude that the output of the algorithm for oracles $\mathcal{F}^0$ and $\mathcal{F}^j$ must be sufficiently different. Formally success of the algorithm implies (see (54) and (61) in Appendix B) that for each i

$$1 - 2\delta \leq T(\rho_T^i, \rho_T^0) \leq \sqrt{1 - F(\rho_T^i, \rho_T^0)} \quad (90)$$

where $T(\rho, \sigma) = \frac{1}{2}\|\rho - \sigma\|_{\text{tr}}$ denotes the trace distance. This implies the bound

$$F_T^i \leq 4\delta(1 - \delta). \quad (91)$$

We conclude that

$$2\sqrt{\frac{p}{1-p}}\sqrt{T}\sqrt{N} \geq N(1 - 4\delta(1 - \delta)) \Rightarrow T \geq \frac{N(1-p)(1 - 4\delta(1 - \delta))^2}{p}. \quad (92)$$

□

E Fidelity Loss of Oracle Calls

In this section we prove Lemma 1, i.e., the loss in fidelity when applying the oracles $\mathcal{F}_i^p$ and $\mathcal{F}_i^q$ to density matrices ρ and σ where we recall that $\mathcal{F}_i^p(\rho) = (1-p)\rho + pO_i\rho O_i^\dagger$. We also prove the slight improvement stated in Corollary 2. First, however, we consider a simpler version of this result in Lemma 8 below, which is required in the proof of Theorem 4.

E.1 Fidelity bound for invariant operators

Here we discuss a lemma that is useful to bound the loss in fidelity $\sqrt{F}(\rho, \sigma) - \sqrt{F}(O\rho O^\dagger, \sigma)$ when it is known that $O\psi = \psi$ for many states ψ (the eigenvalue 1 has large multiplicity). Note that in terms of the oracles $\mathcal{F}_i^p$ this corresponds to the case $p = 1$ and $q = 0$.

Lemma 8. *Let O be a unitary operator and P a hermitian projection such that*

$$P(O - \text{Id}) = (O - \text{Id}), \quad (93)$$

i.e., $1 - P$ projects on a subspace of the eigenspace of eigenvalue 1 of O and $[P, O] = 0$. Let ρ, σ be two density matrices. Then the bound

$$\sqrt{F}(\rho, \sigma) - \sqrt{F}(\rho, O\sigma O^\dagger) \leq 2\sqrt{\text{tr}(P\rho)\text{tr}(P\sigma)} \quad (94)$$

holds.

Proof. Call the system on which ρ, σ act Q . Let R be a copy of Q . Let φ and ψ be purifications of ρ and σ on the system QR such that $\sqrt{F}(\rho, \sigma) = \langle \varphi, \psi \rangle$. Then $(O \otimes \text{Id})\psi$ is a purification of σ and we get

$$\begin{aligned} \sqrt{F}(\rho, O\sigma O^\dagger) &\geq \langle \varphi, (O \otimes \text{Id})\psi \rangle = \langle \varphi, \psi \rangle - \langle \varphi, ((O - \text{Id}) \otimes \text{Id})\psi \rangle \\ &= \sqrt{F}(\rho, \sigma) - \langle \varphi, (P(O - \text{Id}) \otimes \text{Id})\psi \rangle \end{aligned} \quad (95)$$

We now bound using $[P, O] = 0$ and Lemma 6

$$\begin{aligned} \langle \varphi, (P(O - \text{Id}) \otimes \text{Id})\psi \rangle &\leq \langle (P \otimes \text{Id})\varphi, (P \otimes \text{Id})((O - \text{Id}) \otimes \text{Id})\psi \rangle \\ &\leq \|(P \otimes \text{Id})\varphi\| \cdot \|((O - \text{Id}) \otimes \text{Id})(P \otimes \text{Id})\psi\| \\ &\leq 2(\text{tr tr}_R((P \otimes \text{Id})|\varphi\rangle\langle\varphi| (P \otimes \text{Id})^\dagger) \text{tr tr}_R((P \otimes \text{Id})|\psi\rangle\langle\psi| (P \otimes \text{Id})^\dagger))^{\frac{1}{2}} \\ &\leq 2\sqrt{\text{tr}(P\rho)\text{tr}(P\sigma)}. \end{aligned} \quad (96)$$

□

E.2 Proof of Lemma 1

Proof. Call the system on which ρ, σ act Q and we denote $\bar{\rho} = \mathcal{F}_i^p(\rho)$ and $\bar{\sigma} = \mathcal{F}_i^q(\sigma)$. Let R be a copy of Q . Let φ and ψ be purifications of ρ and σ on the system QR such that $\sqrt{F}(\rho, \sigma) = \langle \varphi, \psi \rangle$. We use the shorthand $\bar{O}_i = O_i \otimes \text{Id}_R$ in the following. Let S be a system consisting of a single qubit. We consider the following state on the system QRS

$$\omega = \sqrt{1-p} |\varphi, 0\rangle + \sqrt{p} |\bar{O}_i \varphi, 1\rangle. \quad (97)$$

It is easy to check that ω is a purification of $\bar{\rho}$

$$\text{tr}_{QR} |\omega\rangle \langle \omega| = \text{tr}_Q ((1-p) |\varphi\rangle \langle \varphi| + p |\bar{O}_i \varphi\rangle \langle \bar{O}_i \varphi|) = (1-p)\rho + p O_i \rho O_i^\dagger = \bar{\rho}. \quad (98)$$

To obtain a purification of $\bar{\sigma}$ we define for an angle α the state

$$\zeta = \sqrt{1-q} \cos(\alpha) |\psi, 0\rangle + \sqrt{q} \sin(\alpha) |\bar{O}_i \psi, 0\rangle - \sqrt{1-q} \sin(\alpha) |\psi, 1\rangle + \sqrt{q} \cos(\alpha) |\bar{O}_i \psi, 1\rangle. \quad (99)$$

It is easy to check that $\|\zeta\| = 1$. We now check that this is a purification of $\bar{\sigma}$. Note that the cross terms $|\psi\rangle \langle \bar{O}_i \psi|$ and $|\bar{O}_i \psi\rangle \langle \psi|$ cancel, and thus

$$\begin{aligned} \text{tr}_S |\zeta\rangle \langle \zeta| &= (1-q) \cos^2(\alpha) |\psi\rangle \langle \psi| + q \sin^2(\alpha) |\bar{O}_i \psi\rangle \langle \bar{O}_i \psi| \\ &\quad + (1-q) \sin^2(\alpha) |\psi\rangle \langle \psi| + q \cos^2(\alpha) |\bar{O}_i \psi\rangle \langle \bar{O}_i \psi|. \end{aligned} \quad (100)$$

$$= (1-q) |\psi\rangle \langle \psi| + q |\bar{O}_i \psi\rangle \langle \bar{O}_i \psi|.$$

We calculate

$$\begin{aligned} \langle \omega | \zeta \rangle &= \sqrt{(1-p)(1-q)} \cos(\alpha) \langle \varphi | \psi \rangle + \sqrt{(1-p)q} \sin(\alpha) \langle \varphi | \bar{O}_i \psi \rangle \\ &\quad - \sqrt{p(1-q)} \sin(\alpha) \langle \bar{O}_i \varphi | \psi \rangle + \sqrt{pq} \cos(\alpha) \langle \bar{O}_i \varphi | \bar{O}_i \psi \rangle. \end{aligned} \quad (101)$$

Using that O_i is self-adjoint and unitary we obtain

$$\langle \eta | \zeta \rangle = \left(\sqrt{(1-p)(1-q)} + \sqrt{pq} \right) \cos(\alpha) \langle \varphi | \psi \rangle + \left(\sqrt{(1-p)q} - \sqrt{p(1-q)} \right) \sin(\alpha) \langle \varphi | \bar{O}_i \psi \rangle. \quad (102)$$

Now we set

$$\sin(\alpha) = \sqrt{(1-p)q} - \sqrt{p(1-q)} \quad (103)$$

$$\cos(\alpha) = \sqrt{(1-p)(1-q)} + \sqrt{pq} \quad (104)$$

and obtain

$$\begin{aligned} \langle \omega | \zeta \rangle &= \cos^2(\alpha) \langle \varphi | \psi \rangle + \sin(\alpha)^2 \langle \bar{O}_i \varphi | \psi \rangle \\ &= \langle \varphi | \psi \rangle + \sin(\alpha)^2 \langle \bar{O}_i \varphi - \varphi | \psi \rangle. \end{aligned} \quad (105)$$

We conclude that

$$\sqrt{F}(\bar{\rho}, \bar{\sigma}) \geq |\langle \eta | \zeta \rangle| \geq \sqrt{F}(\rho, \sigma) - |\sin^2(\alpha) \langle \bar{O}_i \varphi - \varphi | \psi \rangle|. \quad (106)$$

As above we write $\bar{P}_i = P_i \otimes \text{Id}$ and $P_{-i} = \text{Id} - P_i$ and we get

$$\bar{O}_i \varphi - \varphi = (\bar{P}_i + \bar{P}_{-i})(\bar{O}_i \varphi - \varphi) = \bar{P}_i(\bar{O}_i \varphi - \varphi) + \bar{P}_{-i}(\bar{O}_i \varphi - \varphi) = \bar{P}_i(\bar{O}_i \varphi - \varphi). \quad (107)$$

Then we control using Lemma 6

$$\begin{aligned} |\langle \bar{O}_i \varphi - \varphi | \psi \rangle|^2 &= |\langle (\bar{O}_i \varphi - \varphi) | \bar{P}_i \psi \rangle|^2 \leq \|\bar{P}_i(\bar{O}_i \varphi - \varphi)\|^2 \|\bar{P}_i \psi\|^2 \\ &= \text{tr} \bar{P}_i \left(|\bar{O}_i \varphi - \varphi\rangle \langle \bar{O}_i \varphi - \varphi| \bar{P}_i^\dagger \right) \text{tr} (|\bar{P}_i \psi\rangle \langle \bar{P}_i \psi|) \\ &= \text{tr} \text{tr}_R \left((\bar{O}_i - \text{Id}) \bar{P}_i |\varphi\rangle \langle \varphi| \bar{P}_i^\dagger (\bar{O}_i^\dagger - \text{Id}) \right) \text{tr} \text{tr}_R \left(\bar{P}_i |\psi\rangle \langle \psi| \bar{P}_i^\dagger \right) \\ &\leq 4 \text{tr} (P_i \rho) \text{tr} (P_i \sigma). \end{aligned} \quad (108)$$

Using the bound (68) from Lemma 4 implies

$$|\sin(\alpha)| = \sqrt{(1-p)q} - \sqrt{p(1-q)} \leq \frac{|p-q|}{2\sqrt{\eta(1-\eta)}}. \quad (109)$$

From (106), (108), and (109) we conclude that

$$\sqrt{F}(\bar{\rho}, \bar{\sigma}) \geq |\langle \eta | \zeta \rangle| \geq \sqrt{F}(\rho, \sigma) - (p-q)^2 \frac{\sqrt{\text{tr}(P_i \rho) \text{tr}(P_i \sigma)}}{2\eta(1-\eta)}. \quad (110)$$

□

E.3 Proof of Corollary 2

Proof. We write $\mathcal{E}_i = \mathcal{E}^i$. We obtain using Corollary 1 (which can be applied since the bit flip operation is self adjoint) for m copies

$$\begin{aligned} \sqrt{F}(\mathcal{E}_i(\rho) \otimes^m, \mathcal{E}_0(\rho) \otimes^m) &= \sqrt{F}(\mathcal{E}_i(\rho), \mathcal{E}_0(\rho))^m \\ &\geq \left(1 - \frac{(p_0 - p_i)^2}{c(1-c)} \text{tr}(P_k \rho)\right)^m \geq 1 - m \frac{4\Delta_k^2}{\eta(1-\eta)} \text{tr}(P_k \rho) \end{aligned} \quad (111)$$

where we used the Bernoulli inequality in the last step. From

$$1 - 2\delta \leq T(\mathcal{E}_i(\rho) \otimes^m, \mathcal{E}_0(\rho) \otimes^m) \leq \sqrt{1 - F(\mathcal{E}_i(\rho) \otimes^m, \mathcal{E}_0(\rho) \otimes^m)} \quad (112)$$

we conclude that

$$2\sqrt{\delta(1-\delta)} \geq \sqrt{F}(\mathcal{E}_i(\rho) \otimes^m, \mathcal{E}_0(\rho) \otimes^m) \geq 1 - m \frac{4\Delta_i^2}{\eta(1-\eta)} \text{tr}(P_i \rho). \quad (113)$$

Equivalently

$$\text{tr}(P_i \rho) \geq \frac{\eta(1-\eta)(1 - 2\sqrt{\delta(1-\delta)})}{4m\Delta_i^2} \quad (114)$$

Using $\text{tr}(\sum_i P_i \rho) = 1$ and summing over $i \geq 2$ we conclude

$$1 \geq \frac{\eta(1-\eta)(1 - 2\sqrt{\delta(1-\delta)})}{4m} \sum_{i \geq 2}^N \Delta_i^{-2}. \quad (115)$$

Using $\sum_{i \geq 2}^N \Delta_i^{-2} \geq \frac{1}{4} \sum_{i=2}^N (p_1 - p_i)^{-2}$ ends the proof. □

F State Decompositions and Fidelity Bounds

In this section we prove Lemma 2 and state a slight extension that will be used in the proof of Theorem 6. Let us restate the lemma including the definition of ψ_0 and ψ_1 for the convenience of the reader.

Lemma 9 (Lemma 2 restated). *Consider a density matrix ρ and a pure state ψ . Consider a unitary and self-adjoint map U and the channel $\mathcal{E}_U^p$ defined by $\mathcal{E}_U^p(\rho) = pU\rho U^\dagger + (1-p)\rho$. Let $0 < \eta < 1/2$ and $p, q \in [\eta, 1-\eta]$. We define*

$$\bar{\rho}_0 = (1-p)\rho, \quad \rho_0 = \rho, \quad \bar{\rho}_1 = pU\rho U^\dagger, \quad \rho_1 = U\rho U^\dagger \quad (116)$$

and

$$\bar{\psi}_0 = \sqrt{1-q} \cos(\alpha) |\psi\rangle + \sqrt{q} \sin(\alpha) |U\psi\rangle, \quad \bar{\psi}_1 = \sqrt{q} \cos(\alpha) |U\psi\rangle - \sqrt{1-q} \sin(\alpha) |\psi\rangle \quad (117)$$

$$\psi_0 = \bar{\psi}_0 / \|\bar{\psi}_0\|, \quad \psi_1 = \bar{\psi}_1 / \|\bar{\psi}_1\| \quad (118)$$

where

$$\cos(\alpha) = \sqrt{pq} + \sqrt{(1-p)(1-q)}, \quad \sin(\alpha) = \sqrt{(1-p)q} - \sqrt{(1-q)p}. \quad (119)$$

We define $q' = \|\bar{\psi}_1\|^2$. Then

$$\mathcal{E}_U^q(|\psi\rangle\langle\psi|) = q' |\psi_1\rangle\langle\psi_1| + (1-q') |\psi_0\rangle\langle\psi_0|. \quad (120)$$

Let $S = \sqrt{F}(\rho, |\psi\rangle\langle\psi|) = \sqrt{\langle\psi|\rho|\psi\rangle}$ denote the initial fidelity. Then the following bound holds

$$\begin{aligned} \sqrt{F}(\mathcal{E}_U^p(\rho), \mathcal{E}_U^q(|\psi\rangle\langle\psi|)) &\geq \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, |\psi_0\rangle\langle\psi_0|) + \sqrt{pq'}\sqrt{F}(\rho_1, |\psi_1\rangle\langle\psi_1|) \\ &\geq S - \frac{(p-q)^2 |(\psi, (U\rho U^\dagger - \rho)\psi)|}{2\eta S} - \frac{(p-q)^2 |\text{Re}(\psi, U\rho\psi) - (\psi, \rho\psi)|^2}{8\eta^2 S^3}. \end{aligned} \quad (121)$$

To clarify the origin of the expressions for $\sin(\alpha)$ and $\cos(\alpha)$ we remark that if we choose β, γ such that $\sqrt{p} = \sin(\beta)$, $\sqrt{q} = \sin(\gamma)$, then $\alpha = \gamma - \beta$. This also explains the simplifications in the formula below, in particular (125) and (128) below are just the trigonometric identities for angle sums.

Proof. First, simple algebra shows $\|\bar{\psi}_0\|^2 = 1 - q'$ and

$$\mathcal{E}_U^p(\rho) = \bar{\rho}_0 + \bar{\rho}_1 = (1-p)\rho_0 + p\rho_1 \quad (122)$$

$$\mathcal{E}_U^q(|\psi\rangle\langle\psi|) = |\bar{\psi}_0\rangle\langle\bar{\psi}_0| + |\bar{\psi}_1\rangle\langle\bar{\psi}_1| = (1-q') |\psi_0\rangle\langle\psi_0| + q' |\psi_1\rangle\langle\psi_1|, \quad (123)$$

in particular (120) holds. Strong concavity of the fidelity implies the first bound of (121). We now address the second estimate. We can express, using that $U = U^\dagger$ is self adjoint

$$\begin{aligned} \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, \psi_0) &= \sqrt{(1-p)(1-q')\langle\psi_0|\rho|\psi_0\rangle} = \sqrt{(1-p)\langle\bar{\psi}_0|\rho|\bar{\psi}_0\rangle} \\ &= \sqrt{(1-p)} \left((1-q) \cos^2(\alpha) \langle\psi|\rho|\psi\rangle + q \sin^2(\alpha) \langle\psi|U\rho U^\dagger|\psi\rangle \right. \\ &\quad \left. + 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle\psi|\rho U|\psi\rangle) \right)^{\frac{1}{2}} \\ &= \sqrt{(1-p)} \left(\left((1-q) \cos^2(\alpha) + q \sin^2(\alpha) + 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \right) S^2 \right. \\ &\quad \left. + q \sin^2(\alpha) \langle\psi|(U\rho U^\dagger - \rho)|\psi\rangle \right. \\ &\quad \left. + 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle\psi|(\rho U - \rho)|\psi\rangle) \right)^{\frac{1}{2}} \end{aligned} \quad (124)$$

Now we calculate using the definition of α

$$\begin{aligned} \sqrt{(1-p)} \left(\sqrt{1-q} \cos(\alpha) + \sqrt{q} \sin(\alpha) \right) \\ = \sqrt{(1-p)(1-q)} (\sqrt{(1-p)(1-q)} + \sqrt{pq}) + \sqrt{(1-p)q} (\sqrt{(1-p)q} - \sqrt{(1-q)p}) \\ = (1-p)(1-q) + (1-p)q = 1-p. \end{aligned} \quad (125)$$

Then we obtain

$$\begin{aligned} \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, \psi_0) \\ = (1-p)S \left(1 + \frac{q \sin^2(\alpha) \langle\psi|(U\rho U^\dagger - \rho)|\psi\rangle + 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle\psi|(\rho U - \rho)|\psi\rangle)}{(1-p)S^2} \right)^{\frac{1}{2}}. \end{aligned} \quad (126)$$

Similarly the second term can be expressed as

$$\begin{aligned} \sqrt{pq'}\sqrt{F}(\rho_1, \psi_1) &= \sqrt{p} \left(\left(q \cos^2(\alpha) + (1-q) \sin^2(\alpha) - 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \right) S^2 \right. \\ &\quad \left. + (1-q) \sin^2(\alpha) \langle\psi|(U\rho U^\dagger - \rho)|\psi\rangle - 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle\psi|(\rho U - \rho)|\psi\rangle) \right)^{\frac{1}{2}}. \end{aligned} \quad (127)$$

We can calculate

$$\begin{aligned}
& \sqrt{p} \left(\sqrt{q} \cos(\alpha) - \sqrt{1-q} \sin(\alpha) \right) \\
&= \sqrt{p} \left(\sqrt{q}(\sqrt{(1-p)(1-q)} + \sqrt{pq}) - \sqrt{1-q}(\sqrt{(1-p)q} - \sqrt{(1-q)p}) \right) \\
&= pq + (1-q)p = p.
\end{aligned} \tag{128}$$

we get

$$\begin{aligned}
& \sqrt{pq'}\sqrt{F}(\rho_1, \psi_1) = \\
&= pS \left(1 + \frac{(1-q) \sin^2(\alpha) \langle \psi | (U\rho U^\dagger - \rho) | \psi \rangle - 2\sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)}{pS^2} \right)^{\frac{1}{2}}.
\end{aligned} \tag{129}$$

We continue to estimate the square root terms. Note that by first order Taylor expansion the mixed last terms of the two expressions (126) and (129) cancel and only the first term and higher order corrections remains. To make this rigorous we use Lemma 5 in Appendix C which states that for $s + t \geq -1$ the bound

$$\sqrt{1 + s + t} \geq 1 + \frac{t}{2} - \frac{t^2}{2} - |s| \tag{130}$$

holds. We apply this to (126) and (129) where s corresponds to the term involving $\langle \psi | (U\rho U^\dagger - \rho) | \psi \rangle$ and t corresponds to the term involving $\text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)^2$. Then we get from (126)

$$\begin{aligned}
& \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, \psi_0) \\
&\geq (1-p)S \left(1 + \frac{-q \sin^2(\alpha) |\langle \psi | (U\rho U^\dagger - \rho) | \psi \rangle| + \sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)}{(1-p)S^2} \right. \\
&\quad \left. - \frac{q(1-q) \cos^2(\alpha) \sin^2(\alpha) \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)^2}{2(1-p)^2 S^4} \right).
\end{aligned} \tag{131}$$

From (129) we get similarly

$$\begin{aligned}
& \sqrt{pq'}\sqrt{F}(\rho_1, \psi_1) = \\
&\geq pS \left(1 + \frac{-(1-q) \sin^2(\alpha) \langle \psi | (U\rho U^\dagger - \rho) | \psi \rangle - \sqrt{q(1-q)} \cos(\alpha) \sin(\alpha) \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)}{pS^2} \right. \\
&\quad \left. - \frac{q(1-q) \cos^2(\alpha) \sin^2(\alpha) \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)^2}{2p^2 S^4} \right).
\end{aligned} \tag{132}$$

We notice that the linear terms in $\text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)$ cancel, and we get

$$\begin{aligned}
& \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, \psi_0) + \sqrt{pq'}\sqrt{F}(\rho_1, \psi_1) \\
&\geq S - \frac{\sin^2(\alpha) |\langle \psi | (U\rho U^\dagger - \rho) | \psi \rangle|}{S} - \left(\frac{1}{p} + \frac{1}{1-p} \right) \frac{q(1-q) \cos^2(\alpha) \sin^2(\alpha) \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)^2}{2S^3}.
\end{aligned} \tag{133}$$

Finally we can estimate using the assumption $p, q \in [\eta, 1 - \eta]$ and Lemma 4

$$\begin{aligned}
& \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, \psi_0) + \sqrt{pq'}\sqrt{F}(\rho_1, \psi_1) \\
&\geq S - \frac{|p-q|^2 \cdot |\langle \psi | (U\rho U^\dagger - \rho) | \psi \rangle|}{4\eta(1-\eta)S} - \left(\frac{1}{\eta} + \frac{1}{1-\eta} \right) \frac{|p-q|^2 \text{Re}(\langle \psi | (\rho U - \rho) | \psi \rangle)^2}{32\eta(1-\eta)S^3}.
\end{aligned} \tag{134}$$

This ends the proof. $\square$

The proof of Theorem 6 requires a slight extension of the lemma above. Unfortunately, we cannot directly derive the result, but we need to slightly modify the proof of Lemma 1.

Corollary 4. *Assume the same setting as in Lemma 1, with the following changes. We consider another self adjoint traceless operator σ and we define*

$$\bar{\rho}_0 = (1-p)(\rho + p\sigma) \quad \rho_0 = \rho + p\sigma \quad \bar{\rho}_1 = pU\rho U^\dagger + p(1-p)\sigma \quad \rho_1 = U\rho U^\dagger + (1-p)\sigma. \quad (135)$$

We assume that ρ_0 and ρ_1 are density matrices, i.e., non-negative. Then the following bound holds

$$\begin{aligned} & \sqrt{(1-p)(1-q')}\sqrt{F}(\rho_0, |\psi_0\rangle\langle\psi_0|) + \sqrt{pq'}\sqrt{F}(\rho_1, |\psi_1\rangle\langle\psi_1|) - S \geq \\ & - \frac{(p-q)^2|(\psi, (U\rho U^\dagger - \rho)\psi)|}{2\eta S} - \frac{(p-q)^2|\text{Re}(\psi, U\rho\psi) - (\psi, \rho\psi)|^2}{8\eta^2 S^3} - \frac{p|\langle\bar{\psi}_0, \sigma\bar{\psi}_0\rangle| + (1-p)|\langle\bar{\psi}_1, \sigma\bar{\psi}_1\rangle|}{S}. \end{aligned} \quad (136)$$

Proof. The proof proceeds exactly as the proof of Lemma 1 with the following minor modifications. We have to insert an additional term $p\langle\bar{\psi}_0, \sigma\bar{\psi}_0\rangle$ in (124) and a term $-(1-p)\langle\bar{\psi}_1, \sigma\bar{\psi}_1\rangle$ in (127). We carry those terms and when we estimate the square-root terms we add them to the s part in the bound (130), thus we end up with an additional term $-p|\langle\bar{\psi}_0, \sigma\bar{\psi}_0\rangle|/S$ in (131) and $-(1-p)|\langle\bar{\psi}_1, \sigma\bar{\psi}_1\rangle|/S$ in (132) and thus their sum appears in (134). This ends the proof. $\square$

G Complexity bounds for classical bandits with a quantum perspective

Before addressing the case of quantum bandits, we revisit the classical bandit problem and give a different proof for the required number of rounds in the fixed confidence setting. This section serves two purposes. It shows that the fidelity of probability distributions is a useful distance measure to analyze classical bandit problems which offers the additional advantage that it readily generalizes to quantum states. Moreover, this section is a preparation that introduces some notation for the more involved proof in the quantum setting in the next section. In fact, the proof for the quantum result shares some ideas, and it is essentially based on a combination of the proof given in this section with the optimal fidelity estimates discussed above. Recall the setting introduced at the beginning of Section 3, in particular the definition of $\mathbf{p}^j$ (for reference, we recapitulate the setting at the beginning of the next section). Then the following result implies Theorem 1.

Theorem 9. *Let $\delta < 1/2$. Assume that $\mathbf{p}^j$ are as defined in Section 3 with $p_i \in [\eta, 1-\eta]$ for some $\eta > 0$. Any classical algorithm that identifies the best arm when it is known that the reward vector is in $\{\mathbf{p}^0, \dots, \mathbf{p}^n\}$ with probability at least $1 - \delta$ requires at least*

$$T \geq cH(\mathbf{p}^1) = c \sum_{j=2}^n \Delta_j^{-2} \quad (137)$$

rounds where $c = c(\delta, \eta) > 0$.

Since this result is well known, the proof serves merely pedagogical purposes to illustrate our approach to the quantum setting. Therefore, we do not give the most concise presentation, but instead highlight the main difference to the standard proofs of this result.

Proof. Suppose we are given an algorithm A . In each step t the algorithm picks an arm a_t depending on all earlier outcomes and receives a (binary) reward $r_t \in \{0, 1\}$. We introduce the variables $x_t = (a_t, r_t) \in [N] \times \{0, 1\}$ encoding the path of the algorithm. We denote by $z_t = (x_1, \dots, x_t)$ the entire history of the exploration. Note that a_t only depends on the outcomes of the previous rounds and therefore is a deterministic function of z_{t-1} , i.e., $a_t = a_t(z_{t-1})$. When the rewards follow the distribution $\mathbf{p}^j$ this induces a distribution on x_t and z_t and we denote the corresponding random variables by Z_t^j and X_t^j and the distribution by $\mathbb{P}^j$. The main idea of the proof is to bound the fidelity of random variables Z_t^j and Z_t^0 for each t from below. On the other hand, we can upper bound the fidelity because the algorithm can identify the best arm for the reward distributions $\mathbf{p}^j$ and $\mathbf{p}^0$ and those arms are different for $j > 1$. Together, those two bounds will imply the claim.

The proof will rely on the fidelity $\sqrt{F}$ of two discrete probability distributions p_x and q_x which is defined by

$$\sqrt{F}(p, q) = \sum_x \sqrt{p_x q_x}. \quad (138)$$

We refer to Appendix B for a brief summary of distance measures, here we only need the definition and the bound by the total variation distance (defined by the first equality)

$$d_{\text{TV}}(p, q) = \frac{1}{2} \sum_x |p_x - q_x| \leq \sqrt{1 - F(p, q)}. \quad (139)$$

We now discuss the simple upper bound on the fidelity after the final round T coming from the assumption that the algorithm succeeds with high probability. Let M_j be the disjoint sets of outcomes z_T such that arm j is selected by the algorithm. By assumption $\mathbb{P}^j(M_j) > 1 - \delta$ and $\mathbb{P}^0(M_1) > 1 - \delta$ and therefore $\mathbb{P}^0(M_j) < \delta$. This implies for $j > 1$ (see Appendix B for a brief summary of distance measures)

$$1 - 2\delta < \mathbb{P}^j(M_j) - \mathbb{P}^0(M_j) \leq d_{\text{TV}}(Z_T^j, Z_T^0) \leq \sqrt{1 - F(Z_T^j, Z_T^0)}. \quad (140)$$

We conclude that

$$\sqrt{F}(Z_T^j, Z_T^0) \leq 2\sqrt{\delta(1 - \delta)}. \quad (141)$$

We now bound the fidelity from below. The first bound will not be sufficient to conclude, but it is nevertheless instructive to understand the difficulties of the quantum setting and the relation to earlier proofs. We will later refine the following estimates. We bound

$$\begin{aligned} \sqrt{F}(Z_t^j, Z_t^0) &= \sum_{z_t} \sqrt{\mathbb{P}^j(z_t) \mathbb{P}^0(z_t)} \\ &= \sum_{r_t=0}^1 \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(r_t | a_t(z_{t-1})) \mathbb{P}^j(z_{t-1}) \mathbb{P}^0(r_t | a_t(z_{t-1})) \mathbb{P}^0(z_{t-1})} \\ &= \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} \sum_{r_t=0}^1 \sqrt{\mathbb{P}^j(r_t | a_t(z_{t-1})) \mathbb{P}^0(r_t | a_t(z_{t-1}))}. \end{aligned} \quad (142)$$

We now consider two cases $a_t(z_{t-1}) = j$ and $a_t(z_{t-1}) \neq j$. In the latter case

$$\mathbb{P}^j(r_t | a_t(z_{t-1})) = \mathbb{P}^0(r_t | a_t(z_{t-1})) = p_j \quad (143)$$

and thus for $a_t(z_{t-1}) \neq j$

$$\sum_{r_t=0}^1 \sqrt{\mathbb{P}^j(r_t | a_t(z_{t-1})) \mathbb{P}^0(r_t | a_t(z_{t-1}))} = \sum_{r_t=0}^1 \mathbb{P}^0(r_t | a_t(z_{t-1})) = 1. \quad (144)$$

For $a_t(z_{t-1}) = j$ we use the simple bound (67) from Lemma 4 in Appendix C. bounding for $p, q \in [c, 1 - c]$

$$\sqrt{F}(\text{Ber}(p), \text{Ber}(q)) \geq 1 - \frac{|p - q|^2}{4c(1 - c)}. \quad (145)$$

This implies

$$\sum_{r_t=0}^1 \sqrt{\mathbb{P}^j(r_t | a_t = j) \mathbb{P}^0(r_t | a_t = j)} \geq 1 - \frac{|p_j - p_0|^2}{4\eta(1 - \eta)}. \quad (146)$$

We can now use the last two displays to continue to estimate (142)

$$\begin{aligned}\sqrt{F}(Z_t^j, Z_t^0) &\geq \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1})\mathbb{P}^0(z_{t-1})} - \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1})\mathbb{P}^0(z_{t-1})} \mathbf{1}_{a_t(z_{t-1})=j} \frac{\Delta_j^2}{4\eta(1-\eta)} \\ &= \sqrt{F}(Z_{t-1}^j, Z_{t-1}^0) - \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1})\mathbb{P}^0(z_{t-1})} \mathbf{1}_{a_t(z_{t-1})=j} \frac{\Delta_j^2}{4\eta(1-\eta)}.\end{aligned}\quad (147)$$

Using this iteratively we obtain (using $\sqrt{F}(Z_0^j, Z_0^0) = 1$) the bound

$$2\sqrt{\delta(1-\delta)} \geq \sqrt{F}(Z_0^j, Z_0^0) \geq 1 - \sum_{t=1}^T \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1})\mathbb{P}^0(z_{t-1})} \mathbf{1}_{a_t(z_{t-1})=j} \frac{\Delta_j^2}{4\eta(1-\eta)}. \quad (148)$$

Now the standard way to proceed from here is to show that with high probability $\mathbb{P}^j(z_{t-1})$ and $\mathbb{P}^0(z_{t-1})$ are similar using tail bounds for random variables (note that we already control the fidelity). Suppose that up to small errors we could replace $\mathbb{P}^j$ by $\mathbb{P}^0$. Then we could conclude from (148) that

$$\frac{\Delta_j^2}{4\eta(1-\eta)} \sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^0(z_{t-1}) \mathbf{1}_{a_t(z_{t-1})=j} \geq 1 - 2\sqrt{\delta(1-\delta)} > 0. \quad (149)$$

Dividing by Δ_j^2 and summing over j this would imply that there is a constant $c > 0$ such that

$$\sum_j \Delta_j^{-2} \leq c \sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^0(z_{t-1}) \sum_j \mathbf{1}_{a_t(z_{t-1})=j} = cT. \quad (150)$$

This approach cannot be extended to the quantum setting. The reason is that the tail bounds rely on the fact that when we use a total of $\mathcal{O}(H)$ queries then we cannot query all arms more often than Δ_j^{-2} . On the other hand, in the quantum setting we query in superposition so that we cannot simply count the number of pulls on an arm. We now show how the tail bounds can be avoided in a way that can similarly be generalized to the quantum setting. Let us denote by $n_j(z_t) = |\{s : a_s(z_t) = j\}|$ the number of times we queried the j -th arm. We introduce the decay factor

$$d^j(z_t) = \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)}\right)^{n_j(z_t)}. \quad (151)$$

Then we will be interested in bounding from below the following (the lower bound on the fidelity of $\mathbb{P}^j$ and $\mathbb{P}^0$)

$$\sum_{z_t} \sqrt{\mathbb{P}^j(z_t)\mathbb{P}^0(z_t)} d^j(z_t) \quad (152)$$

Introducing this decay factor artificially will allow us to derive stronger bounds. Note that

$$d_j(z_t) = d_j(z_{t-1}) \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)} \mathbf{1}_{a_t(z_{t-1})=j}\right). \quad (153)$$

Now we can bound using the same reasoning as in (142) and (147) and the display above

$$\begin{aligned}
\sum_{z_t} \sqrt{\mathbb{P}^j(z_t) \mathbb{P}^0(z_t)} d^j(z_t) &= \sum_{z_t} \sqrt{\mathbb{P}^j(z_t) \mathbb{P}^0(z_t)} d^j(z_{t-1}) \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)} \mathbf{1}_{a_t(z_{t-1})=j}\right) \\
&\geq \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)} \mathbf{1}_{a_t(z_{t-1})=j}\right) \\
&\quad - \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)} \mathbf{1}_{a_t(z_{t-1})=j}\right) \mathbf{1}_{a_t(z_{t-1})=j} \frac{\Delta_j^2}{4\eta(1-\eta)} \\
&\geq \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) - \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \frac{\Delta_j^2}{2\eta(1-\eta)} \mathbf{1}_{a_t(z_{t-1})=j} \\
&\geq \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) - \frac{\Delta_j^2}{2\eta(1-\eta)} \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \mathbf{1}_{a_t(z_{t-1})=j}.
\end{aligned} \tag{154}$$

In particular, we find that the decay factor is chosen such that up to a constant the same bound before holds (see (147)), except that we introduced the terms $d^j(z_t)$ in the loss terms which makes it easier to control them.

Using (141) and a telescopic series we conclude that

$$\begin{aligned}
2\sqrt{\delta(1-\delta)} &\geq \sqrt{F}(Z_T^j, Z_T^0) \geq \sum_{z_T} \sqrt{\mathbb{P}^j(z_T) \mathbb{P}^0(z_T)} d^j(z_T) \\
&\geq 1 - \frac{\Delta_j^2}{2\eta(1-\eta)} \sum_{t=1}^T \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \mathbf{1}_{a_t(z_{t-1})=j}.
\end{aligned} \tag{155}$$

Equivalently, this can be rewritten as

$$(1 - 2\sqrt{\delta(1-\delta)}) \frac{2\eta(1-\eta)}{\Delta_j^2} \leq \sum_{t=1}^T \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \mathbf{1}_{a_t(z_{t-1})=j}. \tag{156}$$

It remains to bound the right-hand side of this inequality. For $t < T$ we write $z_{T|t} = (x_1, \dots, x_t)$ where $z_T = (x_1, \dots, x_t, \dots, x_T)$, i.e., $z_{T|t}$ denotes the restriction of the history z_T to the first t steps. Then we have by definition of $\mathbb{P}$

$$\sum_{z_T} \mathbb{P}^j(z_T) f(z_{T|t}) = \sum_{z_t} \mathbb{P}^j(z_t) f(z_t). \tag{157}$$

Moreover, we note that for all z_T we can bound

$$\begin{aligned}
\sum_{t=1}^T d^j(z_{T|(t-1)})^2 \mathbf{1}[a_t(z_{T|(t-1)})=j] &= \sum_{n=0}^{n_j(z_T)-1} \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)}\right)^{2n} \\
&\leq \sum_{n=0}^{\infty} \left(1 - \frac{\Delta_j^2}{4\eta(1-\eta)}\right)^n = \frac{4\eta(1-\eta)}{\Delta_j^2}.
\end{aligned} \tag{158}$$

We can bound using the Cauchy-Schwarz estimate, (157), and

$$\begin{aligned}
& \sum_{t=1}^T \sum_{z_{t-1}} \sqrt{\mathbb{P}^j(z_{t-1}) \mathbb{P}^0(z_{t-1})} d^j(z_{t-1}) \mathbf{1}_{a_t(z_{t-1})=j} \\
& \leq \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^0(z_{t-1}) \mathbf{1}_{a_t(z_{t-1})=j} \right)^{\frac{1}{2}} \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^j(z_{t-1}) d^j(z_{t-1})^2 \mathbf{1}_{a_t(z_{t-1})=j} \right)^{\frac{1}{2}} \\
& \leq \left(\sum_{t=1}^T \mathbb{P}^0(A_t = j) \right)^{\frac{1}{2}} \left(\sum_{z_T} \mathbb{P}^j(z_T) \sum_{t=1}^T d^j(z_{T|(t-1)})^2 \mathbf{1}_{a_t(z_{T|(t-1)})=j} \right)^{\frac{1}{2}} \\
& \leq \left(\sum_{t=1}^T \mathbb{P}^0(A_t = j) \right)^{\frac{1}{2}} \left(\sum_{z_T} \mathbb{P}^j(z_T) \frac{4\eta(1-\eta)}{\Delta_j^2} \right)^{\frac{1}{2}} \\
& \leq \left(\sum_{t=1}^T \mathbb{P}^0(A_t = j) \right)^{\frac{1}{2}} \frac{2\sqrt{\eta(1-\eta)}}{\Delta_j}.
\end{aligned} \tag{159}$$

Plugging this in (156), dividing by $2\sqrt{\eta(1-\eta)}$ and summing over $j > 1$ gives

$$\begin{aligned}
(1 - 2\sqrt{\delta(1-\delta)})\sqrt{\eta(1-\eta)} \sum_{j=2}^N \Delta_j^{-2} & \leq \sum_{j=2}^N \left[\Delta_j^{-1} \left(\sum_{t=1}^T \mathbb{P}^0(A_t = j) \right)^{\frac{1}{2}} \right] \\
& \leq \left(\sum_{j=2}^N \Delta_j^{-2} \right)^{\frac{1}{2}} \left(\sum_{j=2}^n \sum_{t=1}^T \mathbb{P}^0(A_t = j) \right)^{\frac{1}{2}} \\
& = \left(\sum_{j=2}^n \Delta_j^{-2} \right)^{\frac{1}{2}} \left(\sum_{t=1}^T \sum_{j=1}^n \mathbb{P}^0(A_t = j) \right)^{\frac{1}{2}} \\
& = \sqrt{H(\mathbf{p}^1)} \sqrt{T}.
\end{aligned} \tag{160}$$

Squaring this relation ends the proof. $\square$

H Proof of Theorem 6

In this section, we finally provide a proof of our main result. Let us for reference first summarize the setting that we introduced at the beginning of Section 3. We consider vectors $\mathbf{p} = (p_0, \dots, p_N) \in [\eta, 1-\eta]^{N+1}$ with p_i decreasing and then we let $\mathbf{p}^i \in [\eta, 1-\eta]^N$ be the reward vectors with $\mathbf{p}_j^i = \mathbf{p}_j$ if $i \neq j$ and $\mathbf{p}_i^i = \mathbf{p}_0$. Moreover, $\mathbf{p}_i^0 = \mathbf{p}_i$ for $1 \leq i \leq N$ and note that $\mathbf{p}^i$ and $\mathbf{p}^0$ only differ in entry i and recall that $\Delta_i = p_0 - p_i$. We introduced the oracles $\mathcal{F}_i^p(\rho) = (1-p)\rho + pO_i\rho O_i^\dagger$ (where O_i acts on $|i\rangle$ only) and then defined

$$\mathcal{E}_i = \mathcal{F}_1^{\mathbf{p}^1} \circ \dots \circ \mathcal{F}_N^{\mathbf{p}^N}. \tag{161}$$

We then consider any algorithm start with an initial state ρ_0 and whose output is given by a POVM measurement of the state

$$\mathcal{T}_i \rho_0 = (\mathcal{E}_i \otimes \text{Id}) \circ \mathcal{E}_{U_T} \circ \dots \circ (\mathcal{E}_i \otimes \text{Id}) \circ \mathcal{E}_{U_1} \rho_0 \tag{162}$$

where U_i are unitary maps. We denote the intermediate states of the algorithm by

$$\tilde{\rho}_t^i = \mathcal{E}_{U_t} \rho_t^i \quad \rho_{t+1}^i = (\mathcal{E}_i \otimes \text{Id}) \tilde{\rho}_t^i, \tag{163}$$

i.e., ρ_t^i denotes the state after t oracle invocations and $\tilde{\rho}_t^i$ the state before the $(t+1)$ -th oracle call.

Let us now state a formal version of the main result, Theorem 6.

Theorem 10. Let $\delta < 1/2$. Assume that $\mathbf{p}^j$ are reward vectors as introduced at the beginning of this section, where $p_i \in [\eta, 1 - \eta]$ for some $\eta > 0$. Any quantum algorithm that identifies the best arm when it is known that the reward vector is in $\{\mathbf{p}^0, \dots, \mathbf{p}^n\}$ with probability at least $1 - \delta$ requires at least

$$T \geq cH(\mathbf{p}^1) = c \sum_{i=2}^N \Delta_i^{-2} \quad (164)$$

calls to the oracle $\mathcal{E}_i$ where $c = c(\delta, \eta)$ where an explicit expression under the condition $\Delta_N^2/\eta < 1/2$ is given by

$$c(\delta, \eta) = \left(\frac{\eta(1 - 2\sqrt{\delta(1 - \delta)})}{20} \right)^2. \quad (165)$$

Remark 4. We only do the proof under the condition that $\Delta_N^2/\eta \leq 1/2$ (the behaviour for small Δ_i is the main interest anyway). If this does not hold, some definitions in the proof of Proposition 1 below need to be slightly adjusted (starting with (174) and (175)) but the final result will be the same except that the constant has a poorer dependence on η .

The key ingredient in the proof is a lower bound on the fidelity between the states obtained when applying the different oracles. Let us state this as a separate proposition

Proposition 1. Let $\delta < 1/2$. Assume that $\mathbf{p}^j$ are reward vectors as introduced at the beginning of this section, where $p_i \in [\eta, 1 - \eta]$ for some $\eta > 0$. Then the following lower bound on the fidelity holds for $1 \leq i \leq N$

$$\sqrt{F}(\rho_T^i, \rho_T^0) \geq 1 - \frac{20}{\eta} \Delta_i \left(\sum_{t=1}^T \text{tr } P_i \rho_t^0 \right)^{\frac{1}{2}} \quad (166)$$

for $\Delta_i^2/\eta < 1/2$ and otherwise the bound holds for some constant $c(\eta)$ instead of $20/\eta$.

Once this proposition is proved, the proof of the main result is straightforward.

Proof of Theorem 10. As in the proof of Theorem 7 we have that success of the algorithm implies that

$$\sqrt{F}(\rho_T^i, \rho_T^0) \geq 2\sqrt{\delta(1 - \delta)} \quad (167)$$

holds for $i \geq 2$ (the first arm has the highest mean reward for $\mathbf{p}^1$ and $\mathbf{p}^0$ and thus no bound can be derived for $i = 1$). Combining this with (166) we obtain

$$\begin{aligned} \frac{\eta(1 - 2\sqrt{\delta(1 - \delta)})}{20} \sum_{i \geq 2} \Delta_i^{-2} &\leq \sum_{i \geq 2} \Delta_i^{-1} \left(\sum_{t=1}^T \text{tr } P_i \rho_t^0 \right)^{\frac{1}{2}} \\ &\leq \left(\sum_{i \geq 2} \Delta_i^{-2} \right)^{\frac{1}{2}} \left(\sum_{i \geq 2} \sum_{t=1}^T \text{tr } P_i \rho_t^0 \right)^{\frac{1}{2}} \leq \left(\sum_{i \geq 2} \Delta_i^{-2} \right)^{\frac{1}{2}} \sqrt{T}. \end{aligned} \quad (168)$$

This ends the proof. $\square$

It remains to prove Proposition 1.

Proof of Proposition 1. The proof is a bit technical and lengthy, and we therefore split it into several steps and provide intermediate results.

Overview and notation. Let us first introduce some additional notation and give a high-level overview of the proof. Note that we reviewed the general setting at the beginning of the section.

Our general strategy is to first decompose the corresponding density matrices ρ^i and ρ^0 into a sum of density matrices and then apply strong concavity to lower bound $\sqrt{F}(\rho_T^i, \rho_T^0)$. To control the loss of fidelity that we incur during a single oracle call we will rely on Corollary 4. Aggregating the loss terms this gives rise to a lower bound on the fidelity which, however, involves several error terms that are not straightforward to control. Bounding those error terms will rely on strategies that are similar to the one used in the proof of Theorem 9.

Let us now sketch how we decompose the density matrices before we give the actual definitions. To do this we introduce some notation. To denote the rewards in step t we consider $x_t \in \{0, 1\}^N$ and we collect those rewards in the vector $z_t = (x_1, \dots, x_t)$. As before we consider the measure $\mathbb{P}^i$ on sequences z_T which has the property that $\mathbb{P}^i((x_t)_j = 1) = \mathbf{p}_j^i$ and $\mathbb{P}^i((x_t)_j = 0) = 1 - \mathbf{p}_j^i$ and those variables are independent.

We generally split all states in two after each invocation of an oracle $\mathcal{F}_j^{p_j}$ depending on the realization of the randomness. For $j \neq i$ this will be the natural separation in reward 0 and 1 respectively, but for $j = i$ more complex decompositions need to be used to obtain optimal bounds. Indeed, we have seen in Lemma 2 that the optimal loss in fidelity when applying $\mathcal{E}_i$ and $\mathcal{E}_0$ is of the order Δ_i^2 and we essentially use the decomposition constructed there.

For ρ_t^i we consider a decomposition given by

$$\rho_t^i = \sum_{z_t} \mathbb{P}^i(z_t) \rho(z_t). \quad (169)$$

For ρ_t^0 we consider a decomposition into pure states depending on i , i.e., we construct a distribution $\mathbb{Q}^i$ on sequences z_T and states $\psi(z_t)$ such that

$$\rho_t^0 = \sum_{z_t} \mathbb{Q}^i(z_t) |\psi(z_t)\rangle \langle \psi(z_t)|. \quad (170)$$

We emphasize again that while ρ_t^0 does not depend on i the decomposition does depend on i because it is used to bound the distance to ρ_t^i . To simplify the notation, we drop the i dependence of the decomposition into $\rho(z_t)$ and $\psi(z_t)$. Note that the decomposition for ρ^0 involves the complexity, while the decomposition of ρ^i will be relatively straightforward. Let us now define those decompositions formally.

Decomposition of ρ_t^i . We decompose ρ_t^i roughly as

$$\rho_t(z_t) \approx |\varphi(z_t)\rangle \langle \varphi(z_t)|, \quad \varphi(z_t = (x_1, \dots, x_t)) = (O_{x_t} \otimes \text{Id}) U_t \dots (O_{x_1} \otimes \text{Id}) U_1, \quad (171)$$

i.e., we just decompose it according to the realizations of the rewards. However, we in addition need to ensure that the density matrices $\rho_t(z_t)$ decohere with respect to $\psi(z_t)$. For the definition we introduce the notation $\hat{x}_t \in \{0, 1\}^N$ for the vector x_t with the i -th entry set to 0, i.e., $(\hat{x}_t)_j = (x_t)_j$ for $j \neq i$ and $(\hat{x}_t)_i = 0$. With this notation we get the following decomposition.

Lemma 10. Assume $\Delta_i^2/\eta < 1/2$. We define (recall that ρ_0 denotes the initial state of the algorithm)

$$\rho(z_0) = \rho_0, \quad (172)$$

$$\tilde{\rho}(z_t) = \mathcal{E}_{U_{t+1}}(\rho(z_t)), \quad (173)$$

$$\tilde{\rho}_0(z_t) = \left(1 - \frac{p_0 \Delta_i^2}{\eta}\right) \tilde{\rho}(z_t) + \frac{p_0 \Delta_i^2}{\eta} O_i \tilde{\rho}(z_t) O_i^\dagger, \quad (174)$$

$$\tilde{\rho}_1(z_t) = \left(1 - \frac{(1-p_0) \Delta_i^2}{\eta}\right) O_i \tilde{\rho}(z_t) O_i^\dagger + \frac{(1-p_0) \Delta_i^2}{\eta} \tilde{\rho}(z_t), \quad (175)$$

$$\rho(z_{t+1}) = \mathcal{E}_{O_{\hat{x}_{t+1}}}(\tilde{\rho}_{(x_{t+1})_i}(z_t)). \quad (176)$$

Then (169) holds, i.e.,

$$\tilde{\rho}_t^i = \sum_{z_t} \mathbb{P}(z_t) \tilde{\rho}(z_t). \quad (177)$$

We emphasize that here the notation ρ_0 and ρ_1 corresponds (roughly) to reward 0 or 1 on arm i . Note that the reason to slightly perturb $\tilde{\rho}_0(z_t)$ and $\tilde{\rho}_1$ from their natural definitions $\tilde{\rho}(z_t)$ and $O_i \tilde{\rho}(z_t) O_i^\dagger$ is that our definition ensures up to a constant the same loss in fidelity of order Δ_i^2 but in addition we induce decoherence of $\tilde{\rho}$ so that we can argue as in the proof of Theorem 7. The factor η leads to slightly simpler expressions and tighter bounds, but is not strictly necessary. Indeed, if $\Delta_i^2/\eta > 1$ the definition needs to be adapted by removing the η which results in a weaker η -dependence.

Proof. We argue by induction. We have

$$\tilde{\rho}_t^i = \mathcal{E}_{U_{t+1}}(\rho_t^i) = \mathcal{E}_{U_{t+1}}\left(\sum_{z_t} \mathbb{P}^i(z_t) \rho(z_t)\right) = \sum_{z_t} \mathbb{P}^i(z_t) \mathcal{E}_{U_{t+1}}(\rho(z_t)) = \sum_{z_t} \mathbb{P}^i(z_t) \tilde{\rho}(z_t). \quad (178)$$

Next we note that

$$p_0 \tilde{\rho}_1(z_t) + (1 - p_0) \tilde{\rho}_0(z_t) = p_0 O_i \tilde{\rho}(z_t) O_i^\dagger + (1 - p_0) \tilde{\rho}(z_t) = \mathcal{F}_i^{p_0}(\tilde{\rho}(z_t)). \quad (179)$$

This implies together with the definition of $\mathbb{P}^i$ that

$$\begin{aligned} \sum_{x_{t+1}} \mathbb{P}^i(x_{t+1}) \rho((z_t, x_{t+1})) &= \sum_{x_{t+1}} \mathbb{P}^i(x_{t+1}) \mathcal{E}_{O_{\hat{x}_{t+1}}}(\tilde{\rho}_{(x_{t+1})_i}(z_t)) = \sum_{x_{t+1}} \mathbb{P}^i(x_{t+1}) \mathcal{E}_{O_{\hat{x}_{t+1}}} \circ \mathcal{F}_i^{p_0}(\tilde{\rho}(z_t)) \\ &= \sum_{x_{t+1}} \mathbb{P}^i(x_{t+1}) \mathcal{E}_{O_{x_{t+1}}}(\tilde{\rho}(z_t)) = \mathcal{E}_i(\tilde{\rho}(z_t)). \end{aligned} \quad (180)$$

Here we used that $\mathbb{P}^i((x_{t+1})_i = 1) = \mathbf{p}_i^i = p_0$. Using the induction hypothesis, we conclude that

$$\sum_{z_{t+1}} \mathbb{P}^i(z_{t+1}) \rho(z_{t+1}) = \sum_{z_t} \mathbb{P}^i(z_t) \sum_{x_{t+1}} \mathbb{P}(x_{t+1}) \rho((z_t, x_{t+1})) = \sum_{z_t} \mathbb{P}^i(z_t) \mathcal{E}_i(\tilde{\rho}(z_t)) = \mathcal{E}_i(\tilde{\rho}_t^i) = \rho_{t+1}^i. \quad (181)$$

□

Decomposition of ρ_t^0 . To define the decomposition of ρ_t^0 we first define several quantum states. Let

$$\tilde{\psi}(z_t) = U_t(\psi(z_t)). \quad (182)$$

Define (essentially as in Lemma 2)

$$\bar{\psi}_1(z_t) = \sqrt{1 - p_i} \cos(\alpha) \tilde{\psi}(z_t) + \sqrt{p_i} \sin(\alpha) O_i \tilde{\psi}(z_t) \quad (183)$$

$$\tilde{\psi}_1(z_t) = \bar{\psi}_1(z_t) / \|\bar{\psi}_1(z_t)\| \quad (184)$$

$$\bar{\psi}_0(z_t) = -\sqrt{1 - p_i} \sin(\alpha) \tilde{\psi}(z_t) + \sqrt{p_i} \cos(\alpha) O_i \tilde{\psi}(z_t) \quad (185)$$

$$\tilde{\psi}_0(z_t) = \bar{\psi}_0(z_t) / \|\bar{\psi}_0(z_t)\| \quad (186)$$

where $\alpha \in [0, \pi)$ is defined through $\cos(\alpha) = \sqrt{p_i p_0} + \sqrt{(1 - p_i)(1 - p_0)}$ (i.e., as in (119) with p and q replaced by p_i and p_0). Finally, let

$$\psi(z_{t+1}) = O_{\hat{x}_{t+1}} \tilde{\psi}_{(x_{t+1})_i}(z_t). \quad (187)$$

Next we consider a probability distribution $\mathbb{Q}^i$ on sequences $z_t = (x_1, \dots, x_t)$ as before. It will factorize according to

$$\mathbb{Q}^i(z_{t+1}) = \mathbb{Q}^i(z_t) \mathbb{Q}^i(x_{t+1}|z_t) = \mathbb{Q}^i(z_t) \mathbb{Q}^i((x_{t+1})_i|z_t) \prod_{j \neq i} p_j^{(x_{t+1})_j} (1 - p_j)^{1 - (x_{t+1})_j}. \quad (188)$$

In other words $\mathbb{Q}^i$ is the unique distribution on z_t such that the variables $(x_t)_j$ for $i \neq j$ are independent of everything and distributed according to $\text{Ber}(\mathbf{p}_j^0) = \text{Ber}(p_j)$ and, moreover, we require

$$\begin{aligned}\mathbb{Q}^i\left((x_{t+1})_i = 1 \mid z_t\right) &= \|\bar{\psi}_1(z_t)\|^2 \\ \mathbb{Q}^i\left((x_{t+1})_i = 0 \mid z_t\right) &= \|\bar{\psi}_0(z_t)\|^2 = 1 - \|\bar{\psi}_1(z_t)\|^2.\end{aligned}\tag{189}$$

Clearly this defines uniquely a probability distribution. With this definition we can state the following lemma.

Lemma 11. *The following identity holds*

$$\rho_t^0 = \sum_{z_t} \mathbb{Q}^i(z_t) |\psi(z_t)\rangle \langle \psi(z_t)|, \tag{190}$$

$$\tilde{\rho}_t^0 = \sum_{z_t} \mathbb{Q}^i(z_t) |\tilde{\psi}(z_t)\rangle \langle \tilde{\psi}(z_t)|. \tag{191}$$

Proof. To show this, we argue by induction. The first step is simple

$$\begin{aligned}\sum_{z_t} \mathbb{Q}^i(z_t) |\tilde{\psi}(z_t)\rangle \langle \tilde{\psi}(z_t)| &= \sum_{z_t} \mathbb{Q}^i(z_t) |U_{t+1}\psi(z_t)\rangle \langle U\psi(z_t)| \\ &= U_{t+1} \left(\sum_{z_t} \mathbb{Q}^i(z_t) |\psi(z_t)\rangle \langle \psi(z_t)| \right) U_{t+1}^\dagger = \mathcal{E}_{U_{t+1}}(\rho_t^0) = \tilde{\rho}_t^0.\end{aligned}\tag{192}$$

By (120) in the proof of Lemma 2 the following relation holds

$$\mathcal{F}_i^{p_i}(\tilde{\psi}(z_t)) = \mathbb{Q}^i\left((x_{t+1})_i = 1 \mid z_t\right) |\tilde{\psi}_1(z_t)\rangle \langle \tilde{\psi}_1(z_t)| + \mathbb{Q}^i\left((x_{t+1})_0 = 1 \mid z_t\right) |\tilde{\psi}_0(z_t)\rangle \langle \tilde{\psi}_0(z_t)|. \tag{193}$$

Using this relation we conclude that

$$\begin{aligned}\sum_{z_{t+1}} \mathbb{Q}^i(z_{t+1}) |\psi_0(z_{t+1})\rangle \langle \psi_0(z_{t+1})| &= \sum_{z_{t+1}} \mathbb{Q}^i(z_{t+1}) |O_{\hat{x}_{t+1}} \tilde{\psi}_{(x_{t+1})_i}(z_t)\rangle \langle O_{\hat{x}_{t+1}} \tilde{\psi}_{(x_{t+1})_i}(z_t)| \\ &= \mathcal{F}_1^{p_1} \circ \dots \circ \mathcal{F}_{i-1}^{p_{i-1}} \circ \mathcal{F}_{i+1}^{p_{i+1}} \circ \dots \circ \mathcal{F}_N^{p_N} \left(\sum_{z_t} \mathbb{Q}^i(z_t) \sum_{s=0}^1 \mathbb{Q}^i((x_{t+1})_i = s \mid z_t) |\tilde{\psi}_s(z_t)\rangle \langle \tilde{\psi}_s(z_t)| \right) \\ &= \mathcal{F}_1^{p_1} \circ \dots \circ \mathcal{F}_{i-1}^{p_{i-1}} \circ \mathcal{F}_{i+1}^{p_{i+1}} \circ \dots \circ \mathcal{F}_N^{p_N} \circ \mathcal{F}_i^{p_i} \left(\sum_{z_t} \mathbb{Q}^i(z_t) |\tilde{\psi}(z_t)\rangle \langle \tilde{\psi}(z_t)| \right) \\ &= \mathcal{E}_0(\tilde{\rho}_t^0) = \rho_{t+1}^0.\end{aligned}\tag{194}$$

□

Bounding the fidelity loss in a single step. We now start to estimate the fidelity of ρ_t^k and ρ_t^0 . The first step is to bound the fidelity when making a single oracle call on a single term in the decomposition. The loss only occurs when passing from $\tilde{\rho}(z_t)$ and $\tilde{\psi}(z_t)$ to $\tilde{\rho}_{0/1}(z_t)$ and $\tilde{\psi}_{0/1}(z_t)$. We can show the following bound.

Lemma 12. *Let $z_t \in \{0, 1\}^{N_t}$ and set*

$$S = \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) \tag{195}$$

and assume that $S \geq \frac{1}{2}$. Then the following bound holds with $R(z_t) = \text{tr}(\tilde{\rho}(z_t)^2)$

$$\begin{aligned}&\sqrt{\mathbb{P}^i((x_{t+1})_i = 0) \mathbb{Q}^i((x_{t+1})_i = 0 \mid z_t)} \sqrt{F}(\tilde{\rho}_0(z_t), \tilde{\psi}_0(z_t)) \\ &\quad + \sqrt{\mathbb{P}^i((x_{t+1})_i = 1) \mathbb{Q}^i((x_{t+1})_i = 1 \mid z_t)} \sqrt{F}(\tilde{\rho}_1(z_t), \tilde{\psi}_1(z_t)) \\ &\geq S - \frac{5\Delta_i}{\sqrt{2}\eta} \|P_i \tilde{\psi}(z_t)\| \sqrt{R(z_t) - \max_{x_{t+1}} R((z_t, x_{t+1}))} - \frac{4\Delta_i^2}{\eta^2} \|P_i \tilde{\psi}(z_t)\|^2.\end{aligned}\tag{196}$$

Proof. We will bound this loss using Corollary 4 above. The additional flexibility of the σ term in this corollary allows us to apply this to our setting where $\tilde{\rho}_{0/1}$ are defined as in (174) and (175). Specifically, we apply this Corollary with $\rho = \tilde{\rho}(z_t)$, $\psi = \tilde{\psi}(z_t)$, $p = p_0$, $q = p_i$, $U = O_i$ and $\sigma = \Delta_i^2(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))$. Then we note that the definition of $\tilde{\psi}_{0/1}$ agrees with the definition of $\psi_{0/1}$ in Lemma 1 and $\tilde{\rho}_{0/1}(z_t)$ agrees with $\rho_{0/1}$ and $q' = \mathbb{Q}^i((x_{t+1})_i = 1|z_t)$, $p = \mathbb{P}^i(x_i = 1)$. We conclude from Corollary 4 that

$$\begin{aligned} & \sqrt{(1-p_0)\mathbb{Q}^i((x_{t+1})_i = 0|z_t)}\sqrt{F}(\tilde{\rho}_0(z_t), \tilde{\psi}_0(z_t)) + \sqrt{p_0\mathbb{Q}^i((x_{t+1})_i = 1|z_t)}\sqrt{F}(\tilde{\rho}_1(z_t), \tilde{\psi}_1(z_t)) \\ & \geq S - \Delta_i^2 \left(\frac{|\langle \tilde{\psi}(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\tilde{\psi}(z_t) \rangle|}{\eta} + \frac{\left| \text{Re} \left(\langle \tilde{\psi}(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\tilde{\psi}(z_t) \rangle \right) \right|^2}{\eta^2} \right. \\ & \quad \left. + \frac{2(1-p)}{\eta} \left| \langle \bar{\psi}_1(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\bar{\psi}_1(z_t) \rangle \right| + \frac{2p}{\eta} \left| \langle \bar{\psi}_0(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\bar{\psi}_0(z_t) \rangle \right| \right). \end{aligned} \quad (197)$$

We control the right-hand side of this expression by exploiting the specific structure of the oracle O_i . As in the proof of Theorem 7 we use that $P_i = |i\rangle\langle i| \otimes \text{Id}$ satisfies $(1 - P_i)O_i = (1 - P_i)$ and apply Lemma 3. We get

$$|\langle \tilde{\psi}(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\tilde{\psi}(z_t) \rangle| \leq 2\|P_i\tilde{\psi}(z_t)\| \left(\text{tr} \left(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t) \right)^2 \right)^{\frac{1}{2}}. \quad (198)$$

For the second term we use that $O_i - \text{Id} = (\text{Id} - P_i + P_i)(O_i - \text{Id}) = P_i(O_i - \text{Id})$ which implies after an application of Cauchy-Schwarz

$$\left| \text{Re} \left(\langle \tilde{\psi}(z_t), (O_i\tilde{\rho}(z_t) - \tilde{\rho}(z_t))\tilde{\psi}(z_t) \rangle \right) \right|^2 \leq \|P_i\tilde{\psi}(z_t)\|^2 \cdot \|(O_i - \text{Id})\tilde{\rho}(z_t)\tilde{\psi}(z_t)\|^2 \leq 4\|P_i\tilde{\psi}(z_t)\|^2. \quad (199)$$

For the third term we use again Lemma 3, the definition (183), and $[O_i, P_i] = 0$ and bound

$$\begin{aligned} & \left| \langle \bar{\psi}_1(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\bar{\psi}_1(z_t) \rangle \right| \leq \|P_i\bar{\psi}_1(z_t)\| \cdot \|\bar{\psi}_1(z_t)\| \cdot \left(\text{tr} \left(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t) \right)^2 \right)^{\frac{1}{2}} \\ & \leq \left(\sqrt{1-p_i} \cos(\alpha) \|P_i\tilde{\psi}(z_t)\| + \sqrt{p_i} \sin(\alpha) \|P_i O_i \tilde{\psi}(z_t)\| \right) \|\bar{\psi}_1(z_t)\| \cdot \left(\text{tr} \left(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t) \right)^2 \right)^{\frac{1}{2}} \\ & \leq 2\|P_i\tilde{\psi}(z_t)\| \left(\text{tr} \left(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t) \right)^2 \right)^{\frac{1}{2}}. \end{aligned} \quad (200)$$

The same reasoning implies

$$\left| \langle \bar{\psi}_0(z_t), (O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t))\bar{\psi}_0(z_t) \rangle \right| \leq 2\|P_i\tilde{\psi}(z_t)\| \left(\text{tr} \left(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t) \right)^2 \right)^{\frac{1}{2}}. \quad (201)$$

Plugging (198), (199), (200), and (201) in (197) we obtain

$$\begin{aligned} & \sqrt{(1-p_0)\mathbb{Q}^i((x_{t+1})_i = 0|z_t)}\sqrt{F}(\tilde{\rho}_0(z_t), \tilde{\psi}_0(z_t)) + \sqrt{p_0\mathbb{Q}^i((x_{t+1})_i = 1|z_t)}\sqrt{F}(\tilde{\rho}_1(z_t), \tilde{\psi}_1(z_t)) \\ & \geq S - \frac{5\Delta_i^2}{\eta} \|P_i\tilde{\psi}(z_t)\| \left(\text{tr} \left(O_i\tilde{\rho}(z_t)O_i^\dagger - \tilde{\rho}(z_t) \right)^2 \right)^{\frac{1}{2}} - \frac{4\Delta_i^2}{\eta^2} \|P_i\tilde{\psi}(z_t)\|^2. \end{aligned} \quad (202)$$

It remains to bound the trace term, which will be very similar to the proof of Theorem 7. We define (again not indicating the i dependence)

$$R(z_t) = \text{tr}(\tilde{\rho}(z_t)^2). \quad (203)$$

Invariance of the purity under unitary operations implies that

$$R(z_{t+1}) = \text{tr}(\tilde{\rho}_{(x_{t+1})_i}(z_t)^2). \quad (204)$$

Calculations as in (85) give us for $(x_{t+1})_i = 0$

$$R(z_t) - R(z_{t+1}) = \left(1 - \frac{p_0 \Delta_i^2}{\eta}\right) \frac{p_0 \Delta_i^2}{\eta} \cdot \text{tr}(\tilde{\rho}(z_t) - O_i \tilde{\rho}(z_t) O_i^\dagger)^2. \quad (205)$$

A similar identity for $(x_t)_i = 1$ together with the assumption $\Delta_i^2/\eta \leq 1/2$ and $\min(p_0, 1 - p_0) \geq \eta$ imply

$$\text{tr}(\tilde{\rho}(z_t) - O_i \tilde{\rho}(z_t) O_i^\dagger)^2 \leq \frac{1}{2\Delta_i^2} (R(z_t) - R(z_{t+1})). \quad (206)$$

Note that the left-hand side only depends on z_t so we conclude

$$\text{tr}(\tilde{\rho}(z_t) - O_i \tilde{\rho}(z_t) O_i^\dagger)^2 \leq \frac{1}{2\Delta_i^2} (R(z_t) - \max_{x_{t+1}} R((z_t, x_{t+1}))). \quad (207)$$

Plugging this bound in (202) ends the proof of the lemma. $\square$

The previous Lemma is the key relation that allows us to control the fidelity loss between ρ_t^0 and ρ_t^i . We will use the following corollary.

Corollary 5. *Let $z_t \in \{0, 1\}^{N_t}$ and let as before*

$$S = \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) \quad (208)$$

and assume that $S \geq \frac{1}{2}$. Then the following bound holds (recall $R(z_t) = \text{tr}(\tilde{\rho}(z_t)^2)$)

$$\begin{aligned} & \sum_{x_{t+1}} \sqrt{\mathbb{P}^i(x_{t+1}) \mathbb{Q}^i(x_{t+1}|z_t)} \sqrt{F}(\tilde{\rho}((z_t, x_{t+1})), \tilde{\psi}((z_t, x_{t+1}))) \\ & \geq \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) - \frac{5\Delta_i}{\sqrt{2}\eta} \|P_i \tilde{\psi}(z_t)\| \sqrt{R(z_t) - \max_{x_{t+1}} R((z_t, x_{t+1}))} - \frac{4\Delta_i^2}{\eta^2} \|P_i \tilde{\psi}(z_t)\|^2. \end{aligned} \quad (209)$$

Proof. We first remark that invariance of the fidelity under unitary maps implies

$$\sqrt{F}(\rho((z_t, x_{t+1})), |\psi((z_t, x_{t+1}))\rangle) = \sqrt{F}(\tilde{\rho}_{(x_{t+1})_i}(z_t), |\psi_{(x_{t+1})_i}(z_t)\rangle), \quad (210)$$

$$\sqrt{F}(\rho(z_{t+1}), |\psi(z_{t+1})\rangle) = \sqrt{F}(\tilde{\rho}(z_{t+1}), |\tilde{\psi}(z_{t+1})\rangle). \quad (211)$$

We now sum the bound in Lemma 12 over all possible values of x_{t+1} to move from step $(t+1)$ back to step t . We denote $\hat{x}_{t+1}$ the vector x_{t+1} with entry i removed. Note that under $\mathbb{P}^i$ and $\mathbb{Q}^i$ this vector is independent of z_t , $(x_{t+1})_i$ and $\mathbb{P}^i(\hat{x}_{t+1}) = \mathbb{Q}^i(\hat{x}_{t+1})$. We also introduce the notation $\hat{x}_{t+1}^c$ for the vector that has entry i equal to c . Then we get using (210) and (211) for any z_t and $c \in \{0, 1\}$

$$\begin{aligned} \sum_{\hat{x}_{t+1}} \sqrt{\mathbb{P}^i(\hat{x}_{t+1}) \mathbb{Q}^i(\hat{x}_{t+1})} \sqrt{F}(\tilde{\rho}((z_t, \hat{x}_{t+1}^c)), \tilde{\psi}((z_t, \hat{x}_{t+1}^c))) &= \sum_{\hat{x}_{t+1}} \sqrt{\mathbb{P}^i(\hat{x}_{t+1}) \mathbb{P}^i(\hat{x}_{t+1})} \sqrt{F}(\tilde{\rho}_c(z_t), \tilde{\psi}_c(z_t)) \\ &= \sqrt{F}(\tilde{\rho}_c(z_t), \tilde{\psi}_c(z_t)). \end{aligned} \quad (212)$$

Summing this over $c = 0, 1$ and using (202) we get for all z_t such that $\sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) \geq 1/2$ the bound

$$\begin{aligned}
& \sum_{x_{t+1}} \sqrt{\mathbb{P}^i(x_{t+1}) \mathbb{Q}^i(x_{t+1}|z_t)} \sqrt{F}(\tilde{\rho}((z_t, x_{t+1})), \tilde{\psi}((z_t, x_{t+1}))) \\
&= \sum_{c=0}^1 \sqrt{\mathbb{P}^i((x_{t+1})_i = c) \mathbb{Q}^i((x_{t+1})_i = c|z_t)} \sum_{\hat{x}_{t+1}} \sqrt{\mathbb{P}^i(\hat{x}_{t+1}) \mathbb{Q}^i(\hat{x}_{t+1})} \sqrt{F}(\tilde{\rho}((z_t, \hat{x}_{t+1}^c)), \tilde{\psi}((z_t, \hat{x}_{t+1}^c))) \\
&= \sum_{c=0}^1 \sqrt{\mathbb{P}^i((x_{t+1})_i = c) \mathbb{Q}^i((x_{t+1})_i = c|z_t)} \sqrt{F}(\tilde{\rho}_c(z_t), \tilde{\psi}_c(z_t)) \\
&\geq \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) - \frac{5\Delta_i}{\sqrt{2}\eta} \|P_i \tilde{\psi}(z_t)\| \sqrt{R(z_t) - \max_{x_{t+1}} R((z_t, x_{t+1}))} - \frac{4\Delta_i^2}{\eta^2} \|P_i \tilde{\psi}(z_t)\|^2.
\end{aligned} \tag{213}$$

□

Deriving an error decomposition for the fidelity. Equipped with this corollary, we now move on to control the total loss in fidelity.

The general strategy is now to use joint concavity of the fidelity and then inductive application of the estimate (202) above to lower bound the fidelity. There are two technical difficulties: When directly applying the corollary above, we do not get the optimal bound because it is difficult to control the difference between $\mathbb{P}^i$ and $\mathbb{Q}^i$. This is the same problem as in the proof of Theorem 9, and we can address this with a similar idea. The second difficulty is that the change in fidelity in Lemma 2 involves the inverse of the initial fidelity, and so we derived (213) only for fidelity $S \geq 1/2$. The high-level argument that this is sufficient is that if we know that the weighted mean of the fidelities is large, then there cannot be too many small terms in the mixture, so the condition will mostly hold.

Technically, we address both difficulties by introducing additional sequences $d(z_t)$, $s(z_t)$, and $h(z_t)$ that we smuggle into the sum. We define $d(z_0) = 1$ and then recursively for $z_t = (z_{t-1}, x_t)$

$$d(z_t) = d(z_{t-1}) \left(1 - \frac{\Delta_i^2 \|P_i \tilde{\psi}(z_{t-1})\|^2}{\eta^2} \right) = d(z_{t-1}) - d(z_{t-1}) \frac{\Delta_i^2 \|P_i \tilde{\psi}(z_{t-1})\|^2}{\eta^2}. \tag{214}$$

This is the quantum analogue of the classical definition in (153) and has the same motivation. Note that the η factor was introduced for convenience, it could be dropped at the price of a slightly worse η dependence.

Next, we define

$$s(z_t) = 1 \quad \text{iff for all } t' \leq t \text{ the bound } \sqrt{F}(\tilde{\rho}(z_{t'}), \tilde{\psi}(z_{t'})) \geq 1/2 \text{ holds.} \tag{215}$$

In other words $s(z_t) = 0$ for $z_t = (x_1, \dots, x_t)$ if there is $t' \leq t$ such that $z_{t'} = (x_1, \dots, x_{t'})$ satisfies $\sqrt{F}(\tilde{\rho}(z_{t'}), \tilde{\psi}(z_{t'})) < 1/2$. Moreover, we define $h(z_t) = 1$ if $s(z_t) = 0$ but $s(z_{t'}) = 1$ for all $z_{t'}$ as above. Put differently for $z_t = (z_{t-1}, x_t)$ the relation

$$h(z_t) = s(z_{t-1}) - s(z_t) \tag{216}$$

holds, i.e., $h(z_t)$ keeps track when the fidelity $\sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t))$ falls below $1/2$ for the first time.

We can now bound (loosely speaking) the change in fidelity $\sqrt{F}(\rho_{T-1}^k, \rho_{T-1}^0) - \sqrt{F}(\rho_T^k, \rho_T^0)$ (actually, we only bound the difference on the lower bounds). This will be achieved by combining the estimates above and the introduction of various error terms. We first note that Lemma 11 and Lemma 10 together with strong concavity of the fidelity (see (59)) and the upper bounds $s \leq$ and $d \leq 1$ imply

$$\begin{aligned}
\sqrt{F}(\tilde{\rho}_T^k, \tilde{\rho}_T^0) &\geq \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) \\
&\geq \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) s(z_T).
\end{aligned} \tag{217}$$

Now we first extract the error terms that we get from the change in the sequences d and s . We again use the notation $z_{T|t} = (x_1, \dots, x_t)$ for $t \leq T$ and $z_T = (x_1, \dots, x_T)$ and then we obtain

$$\begin{aligned}
\sqrt{F}(\tilde{\rho}_T^k, \tilde{\rho}_T^0) &\geq \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) s(z_T) \\
&= \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) (s(z_{T|T-1}) - h(z_T)) \\
&= \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) s(z_{T|T-1}) - E_T^1 \\
&= -E_T^1 + \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_{T|T-1}) s(z_{T|T-1}) \\
&\quad - \frac{\Delta_i^2}{\eta^2} \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) \|P_i \tilde{\psi}(z_{T|T-1})\|^2 d(z_{T|T-1}) s(z_{T|T-1}) \\
&= -E_T^1 - E_T^2 + \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_{T|T-1}) s(z_{T|T-1}).
\end{aligned} \tag{218}$$

Where the error terms E_T^1 and E_T^2 are defined by those equations, i.e.,

$$E_T^1 = \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) h(z_T), \tag{219}$$

$$E_T^2 = \frac{\Delta_i^2}{\eta^2} \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) \|P_i \tilde{\psi}(z_{T|T-1})\|^2 d(z_{T|T-1}) s(z_{T|T-1}). \tag{220}$$

We continue to estimate the remaining term using (209) in Corollary 5. Here we use that either the factor $s(z_{T|T-1}) = 0$ vanishes and the inequality below is trivially true (both sides are zero) or the bound $\sqrt{F}(\tilde{\rho}(z_{T|T-1}), \tilde{\psi}(z_{T|T-1})) \geq 1/2$ holds so that Corollary 5 can be applied

$$\begin{aligned}
&\sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_{T|T-1}) s(z_{T|T-1}) \\
&\geq \sum_{z_{T-1}} \sqrt{\mathbb{P}^i(z_{T-1}) \mathbb{Q}^i(z_{T-1})} d(z_{T-1}) s(z_{T-1}) \left(\sqrt{F}(\tilde{\rho}(z_{T-1}), \tilde{\psi}(z_{T-1})) \right. \\
&\quad \left. - \frac{5\Delta_i}{\sqrt{2}\eta} \|P_i \tilde{\psi}(z_{T-1})\| \left(R(z_{T-1}) - \max_{x_t} R((z_{T-1}, x_t)) \right)^{\frac{1}{2}} - \frac{4\Delta_i^2}{\eta^2} \|P_i \tilde{\psi}(z_{T-1})\|^2 \right) \\
&\geq \sum_{z_{T-1}} \sqrt{\mathbb{P}^i(z_{T-1}) \mathbb{Q}^i(z_{T-1})} \sqrt{F}(\tilde{\rho}(z_{T-1}), \tilde{\psi}(z_{T-1})) d(z_{T-1}) s(z_{T-1}) - E_T^3 - E_T^4
\end{aligned} \tag{221}$$

where we again define the error terms E_T^3 and E_T^4 implicitly through these equations, i.e.,

$$E_T^3 = \frac{5\Delta_i}{\sqrt{2}\eta} \sum_{z_{T-1}} \sqrt{\mathbb{P}^i(z_{T-1}) \mathbb{Q}^i(z_{T-1})} d(z_{T-1}) s(z_{T-1}) \|P_i \tilde{\psi}(z_{T-1})\| \left(R(z_{T-1}) - \max_{x_t} R((z_{T-1}, x_t)) \right)^{\frac{1}{2}}, \tag{222}$$

$$E_T^4 = \frac{4\Delta_i^2}{\eta^2} \sum_{z_{T-1}} \sqrt{\mathbb{P}^i(z_{T-1}) \mathbb{Q}^i(z_{T-1})} d(z_{T-1}) s(z_{T-1}) \|P_i \tilde{\psi}(z_{T-1})\|^2. \tag{223}$$

Applying this inductively together with $\sqrt{F}(\tilde{\rho}_0^i, \tilde{\rho}_0^0) = 1$ we get

$$\sqrt{F}(\tilde{\rho}_T^i, \tilde{\rho}_T^0) \geq \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) s(z_T) \geq 1 - \sum_{t=1}^T E_t^1 + E_t^2 + E_t^3 + E_t^4 \tag{224}$$

Bounding the error terms E_t^2 and E_t^4 .

Lemma 13. *The error terms E_t^2 and E_t^4 satisfy the bound*

$$\sum_{t=1}^T E_t^2 + E_t^4 \leq \frac{5\Delta_i}{\eta} \left(\sum_{t=1}^T \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}}. \quad (225)$$

Proof. We start to bound E_t^2 . We find using $\sqrt{F} \leq 1$, $s(z_t) \leq 1$

$$\begin{aligned} E_t^2 &\leq \frac{\Delta_i^2}{\eta^2} \sum_{z_{t-1}} \sqrt{\mathbb{P}^i(z_{t-1}) \mathbb{Q}^i(z_{t-1})} \|P_i \tilde{\psi}(z_{t-1})\|^2 d(z_{t-1}) \sum_{x_t} \sqrt{\mathbb{P}^i(x_t | z_{t-1}) \mathbb{Q}^i(x_t | z_{t-1})} \\ &\leq \frac{\Delta_i^2}{\eta^2} \sum_{z_{t-1}} \sqrt{\mathbb{P}^i(z_{t-1}) \mathbb{Q}^i(z_{t-1})} \|P_i \tilde{\psi}(z_{t-1})\|^2 d(z_{t-1}) \end{aligned} \quad (226)$$

Combining this with the definition of E_t^4 and using again $s(z_{t-1}) \leq 1$ we obtain using Cauchy Schwarz

$$\begin{aligned} \sum_{t=1}^T E_t^2 + E_t^4 &\leq \frac{5\Delta_i^2}{\eta^2} \sum_{t=1}^T \sum_{z_{t-1}} \sqrt{\mathbb{P}^i(z_{t-1}) \mathbb{Q}^i(z_{t-1})} \|P_i \tilde{\psi}(z_{t-1})\|^2 d(z_{t-1}) \\ &\leq \frac{5\Delta_i}{\eta} \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{Q}^i(z_{t-1}) \|P_i \tilde{\psi}(z_{t-1})\|^2 \right)^{\frac{1}{2}} \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^i(z_{t-1}) \frac{\Delta_i^2}{\eta^2} \|P_i \tilde{\psi}(z_{t-1})\|^2 d(z_{t-1}) \right)^{\frac{1}{2}}. \end{aligned} \quad (227)$$

For the second factor we apply the definition (214) combined with $\sum \mathbb{P}^i(x_t) = 1$ to rewrite the expression, and we find

$$\begin{aligned} \sum_{t=1}^T E_t^2 + E_t^4 &\leq \frac{5\Delta_i}{\eta} \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{Q}^i(z_{t-1}) \text{tr}(P_i |\tilde{\psi}(z_{t-1})\rangle \langle \tilde{\psi}(z_{t-1})| P_i) \right)^{\frac{1}{2}} \\ &\quad \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^i(z_{t-1}) \left(d(z_{t-1}) - \sum_{x_t} \mathbb{P}^i(x_t) d((z_{t-1}, x_t)) \right) \right)^{\frac{1}{2}} \\ &\leq \frac{5\Delta_i}{\eta} \left(\sum_{t=1}^T \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}} \left(\sum_{t=1}^T \sum_{z_{t-1}} \mathbb{P}^i(z_{t-1}) d(z_{t-1}) - \sum_{t=1}^T \sum_{z_t} \mathbb{P}^i(z_t) d(z_t) \right)^{\frac{1}{2}} \\ &\leq \frac{5\Delta_i}{\eta} \left(\sum_{t=1}^T \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}} \left(d(z_0) - \sum_{z_T} \mathbb{P}^i(z_T) d(z_T) \right)^{\frac{1}{2}} \\ &\leq \frac{5\Delta_i}{\eta} \left(\sum_{t=1}^T \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}} \end{aligned} \quad (228)$$

where we used Lemma 11 for the first factor and (191) and the telescopic sum together with $d(z_0) = 1$ and $0 \leq d(z_T) \leq 1$ in the last steps for the second factor. $\square$

Bounding the error term E_t^3 .

Lemma 14. *The error term E_t^3 is bounded by*

$$\sum_{t=1}^T E_t^3 \leq \frac{5\Delta_i}{\sqrt{2}\eta} \left(\sum_{t=0}^{T-1} \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}}. \quad (229)$$

We bound, using again Cauchy Schwarz and $s(z_t)d(z_t) \leq 1$,

$$\begin{aligned}
\sum_{t=1}^T E_t^3 &= \frac{5\Delta_i}{\sqrt{2}\eta} \sum_{t=0}^{T-1} \sum_{z_t} \sqrt{\mathbb{P}^i(z_t)\mathbb{Q}^i(z_t)} d(z_t) s(z_t) \|P_i \tilde{\psi}(z_t)\| \left(R(z_t) - \max_{x_{t+1}} R((z_t, x_{t+1})) \right)^{\frac{1}{2}} \\
&\leq \frac{5\Delta_i}{\sqrt{2}\eta} \left(\sum_{t=0}^{T-1} \sum_{z_t} \mathbb{Q}^i(z_t) \|P_i \tilde{\psi}(z_t)\|^2 \right)^{\frac{1}{2}} \left(\sum_{t=0}^{T-1} \sum_{z_t} \mathbb{P}^i(z_t) \left(R(z_t) - \max_{x_{t+1}} R((z_t, x_{t+1})) \right) \right)^{\frac{1}{2}} \\
&\leq \frac{5\Delta_i}{\sqrt{2}\eta} \left(\sum_{t=0}^{T-1} \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}} \left(\sum_{t=0}^{T-1} \sum_{z_t} \mathbb{P}^i(z_t) R(z_t) - \sum_{t=0}^{T-1} \sum_{z_t, x_{t+1}} \mathbb{P}^i(z_t) \mathbb{P}^i(x_{t+1}) R((z_t, x_{t+1})) \right)^{\frac{1}{2}} \\
&\leq \frac{5\Delta_i}{\sqrt{2}\eta} \left(\sum_{t=0}^{T-1} \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}} \left(\sum_{t=0}^{T-1} \sum_{z_t} \mathbb{P}^i(z_t) R(z_t) - \sum_{t=1}^T \sum_{z_t} \mathbb{P}^i(z_t) R(z_t) \right)^{\frac{1}{2}} \\
&\leq \frac{5\Delta_i}{\sqrt{2}\eta} \left(\sum_{t=0}^{T-1} \text{tr}(P_i \tilde{\rho}_t^0) \right)^{\frac{1}{2}}.
\end{aligned} \tag{230}$$

Here we applied Lemma 11 similar to the previous lemma, and we once more used the telescopic sum and the fact that $0 \leq R(z_t) \leq 1$.

Bounding the error term E_t^1 . It remains to bound the last remaining error term E_t^1 where we can prove the following bound.

Lemma 15. *The error term E_t^1 can be bounded as follows*

$$\sum_{t=1}^T E_t^1 \leq \sum_{t=1}^T E_t^2 + E_t^3 + E_t^4. \tag{231}$$

Proof. The key ingredient to bound E_t^1 is the observation that by definition of $s(z_t)$ and $h(z_t)$ we have $\sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) < 1/2$ if $h(z_t) = 1$ which implies

$$\sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) < 1 - \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)). \tag{232}$$

From here we conclude using the definition of E_t^1 (and $h(z_t) = 0$ if $h(z_t) \neq 1$)

$$\begin{aligned}
\sum_{t=1}^T E_t^1 &= \sum_{t=1}^T \sum_{z_t} \sqrt{\mathbb{P}^i(z_t)\mathbb{Q}^i(z_t)} \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) d(z_t) h(z_t) \\
&\leq \sum_{t=1}^T \sum_{z_t} \sqrt{\mathbb{P}^i(z_t)\mathbb{Q}^i(z_t)} \left(1 - \sqrt{F}(\tilde{\rho}(z_t), \tilde{\psi}(z_t)) \right) d(z_t) h(z_t) \\
&\leq \sum_{t=1}^T \sum_{z_t} \sqrt{\mathbb{P}^i(z_t)\mathbb{Q}^i(z_t)} h(z_t) - \sum_{t=1}^T E_t^1.
\end{aligned} \tag{233}$$

We now find using (224) and the last display that

$$\begin{aligned}
\sum_{z_T} \sqrt{\mathbb{P}^i(z_T)\mathbb{Q}^i(z_T)} s(z_T) &\geq \sum_{z_T} \sqrt{\mathbb{P}^i(z_T)\mathbb{Q}^i(z_T)} \sqrt{F}(\tilde{\rho}(z_T), \tilde{\psi}(z_T)) d(z_T) s(z_T) \\
&\geq 1 - \sum_{t=1}^T (E_t^1 + E_t^2 + E_t^3 + E_t^4) \\
&\geq 1 - \sum_{t=1}^T (E_t^2 + E_t^3 + E_t^4) + \sum_{t=1}^T E_t^1 - \sum_{t=1}^T \sum_{z_t} \sqrt{\mathbb{P}^i(z_t)\mathbb{Q}^i(z_t)} h(z_t).
\end{aligned} \tag{234}$$

Now it is easy to see that

$$\sum_{t=1}^T \sum_{z_t} \mathbb{P}^i(z_t) h(z_t) + \sum_{z_T} \mathbb{P}^i(z_T) s(z_T) = 1 \quad (235)$$

because this is the probability that the fidelity drops below $1/2$ under the measure $\mathbb{P}^i$ plus the probability that it is at least $1/2$. Clearly, a similar relation holds for $\mathbb{Q}^i$. Now we apply $\sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} \leq (\mathbb{P}^i(z_T) + \mathbb{Q}^i(z_T))/2$ and find

$$\begin{aligned} & \sum_{z_T} \sqrt{\mathbb{P}^i(z_T) \mathbb{Q}^i(z_T)} s(z_T) + \sum_{t=1}^T \sum_{z_t} \sqrt{\mathbb{P}^i(z_t) \mathbb{Q}^i(z_t)} h(z_t) \\ & \leq \frac{1}{2} \left(\sum_{z_T} \mathbb{P}^i(z_T) s(z_T) + \sum_{t=1}^T \sum_{z_t} \mathbb{P}^i(z_t) h(z_t) + \sum_{z_T} \mathbb{Q}^i(z_T) s(z_T) + \sum_{t=1}^T \sum_{z_t} \mathbb{Q}^i(z_t) h(z_t) \right) = 1. \end{aligned} \quad (236)$$

Using this relation in (234) we find

$$1 \geq 1 - \sum_{t=1}^T (E_t^2 + E_t^3 + E_t^4) + \sum_{t=1}^T E_t^1 \quad (237)$$

from which

$$\sum_{t=1}^T E_t^1 \leq \sum_{t=1}^T (E_t^2 + E_t^3 + E_t^4) \quad (238)$$

follows. $\square$

Conclusion. To finish the proof of Proposition 1 we only have to plug all the derived bounds in (224). Indeed, applying Lemma 13, Lemma 14, and Lemma 15 we get

$$\begin{aligned} \sqrt{F}(\rho_T^i, \rho_T^0) & \geq 1 - \sum_{t=1}^T (E_t^1 + E_t^2 + E_t^3 + E_t^4) \geq 1 - 2 \sum_{t=1}^T (E_t^2 + E_t^3 + E_t^4) \\ & \geq 1 - 20\eta^{-1} \Delta_i \left(\sum_{t=1}^T \text{tr } P_i \tilde{\rho}_t^0 \right)^{\frac{1}{2}}. \end{aligned} \quad (239)$$

This ends the proof. $\square$

I Proof of Theorem 4

Here we prove the lower bound in Theorem 4. Let us first give a precise statement of the result. We consider the same probability vectors $\mathbf{p}_i$ as introduced at the beginning of Section 3. First, we note that the result does not hold for fixed oracles O_i with reward vector $\mathbf{p}_i$ because the algorithm could exploit the specific structure of the oracles. There could be, e.g., a state ω_0 such that $O(\mathbf{p}_i) |j\rangle |\omega_0\rangle |0\rangle = |j\rangle |\omega_0\rangle |\delta_{ij}\rangle$. Then the problem reduces to the unstructured search problem when ω_0 is known. Thus, the result only holds when we assume that O_i is a random oracle with the fixed reward vector $\mathbf{p}_i$, emulating the situation where we have no additional information about the oracles. We consider for a reward vector $r \in \{0, 1\}^{|\mathcal{H}_A| \cdot |\mathcal{H}_P|}$ the oracle O^r acting as in (7), i.e.,

$$O^r |i\rangle |\omega\rangle |c\rangle = |i\rangle |\omega\rangle |c + r_i(\omega)\rangle. \quad (240)$$

We consider a random reward distribution $r(i)$ where $r(i)$ is the uniform distribution over all reward vectors with mean reward vector $\mathbf{p}_i$, i.e., $|\mathcal{H}_P|^{-1} \sum_{\omega} r(i)_j(\omega) = (\mathbf{p}_i)_j$. Then a more precise version of Theorem 4 reads as follows.

Theorem 11. Let $\delta < 1/2$. Assume that $\mathbf{p}^j$ are as before with $p_i \in [\eta, 1 - \eta]$ for some $\eta > 0$. Any algorithm that identifies the best arm with probability at least $1 - \delta$ given an oracle $O^{r(i)}$ where $r(i)$ is distributed as above requires at least

$$T \geq \frac{1}{2} \eta (1 - 2\sqrt{\delta(1 - \delta)}) \left(\sum_{i=2}^n \Delta_i^{-2} \right)^{\frac{1}{2}} \geq c(\delta, \eta) \sqrt{H(\mathbf{p}^1)} \quad (241)$$

calls to the oracle. This is still true even when it is known that the vector of mean rewards is $\{\mathbf{p}^0, \dots, \mathbf{p}^n\}$.

Proof. The proof is close to the proof of Theorem 7. We assume we are given any algorithm acting by $(\mathcal{E}_O \otimes \text{Id}) \circ \mathcal{E}_{U_T} \circ \dots \circ (\mathcal{E}_O \otimes \text{Id}) \circ \mathcal{E}_{U_1}$ where U_t are arbitrary unitary maps where O denotes the given oracle. Assume that the initial state is a fixed density matrix ρ . We denote the state using the oracle $O^{r(i)}$ before the $t + 1$ -th invocation of the oracle by $\rho_t^{r(i)}$, i.e., $\rho_0^{r(i)} = \mathcal{E}_{U_1}(\rho)$. We introduce the notation $\mathbb{E}_i$ when we average over the reward distribution $r(i)$ and we write

$$\rho_t^i = \mathbb{E}_i \rho_t^{r(i)}. \quad (242)$$

By assumption we can identify i given ρ_T^i with probability at least $1 - \delta$. This implies, as in (141), (91) for $i > 1$

$$2\sqrt{\delta(1 - \delta)} \geq \sqrt{F}(\rho_T^i, \rho_T^0). \quad (243)$$

One main ingredient of the proof is to define a suitable coupling of the random variables $r(0)$ and $r(i)$. Note that its reward vectors are $\mathbf{p}_0$ and $\mathbf{p}_i$ such that $(\mathbf{p}_i)_j = p_j = (\mathbf{p}_0)_j$ for $j \neq i$ and $(\mathbf{p}_i)_i = p_0$, $(\mathbf{p}_0)_i = p_i$. We now consider a coupling where $r(0)_j = r(i)_j$ for $j \neq i$ and $r(i)_i \geq r(0)_i$ and the distribution of $r(i)_i \in \{0, 1\}^{|\mathcal{H}_P|}$ is uniform over all rewards under this constraint for a fixed $r(0)$. Note that this entails that the distribution of $r(0)$ for a fixed $r(i)$ is also uniform over all rewards with the right mean reward satisfying $r(0) \leq r(i)$. It is straightforward to see that such a coupling exists (a possible explicit construction is to draw i.i.d. random numbers $u_i(\omega)$ and set $r(0)_i(\omega) = 1$ iff $u_i(\omega)$ is in the p_i -th quantile of the numbers $u_i(\cdot)$ and similarly for $r(i)$). We denote this coupling by $p_i(r(i), r(0))$. Observe that for this coupling satisfies, for all ω ,

$$\mathbb{P}(r(i)_j(\omega) = 1 | r(0)_j(\omega) = 0) = \mathbb{P}(r(i)_j = 1 | r(0)_j = 0) = \begin{cases} 0 & \text{for } j \neq i \\ \frac{p_0 - p_i}{1 - p_i} = \frac{\Delta_i}{1 - p_i} & \text{for } j = i. \end{cases} \quad (244)$$

Similarly, we get for all ω

$$\mathbb{P}(r(0)_j(\omega) = 0 | r(i)_j(\omega) = 1) = \mathbb{P}(r(0)_j = 0 | r(i)_j = 1) = \begin{cases} 0 & \text{for } j \neq i \\ \frac{p_0 - p_i}{p_0} = \frac{\Delta_i}{p_0} & \text{for } j = i. \end{cases} \quad (245)$$

We can bound using the concavity of the fidelity

$$\sqrt{F}(\rho_t^i, \rho_t^0) \geq \sum_{r(i), r(0)} p_i(r(i), r(0)) \sqrt{F}(\rho_t^{r(i)}, \rho_t^{r(0)}). \quad (246)$$

Now we lower bound the fidelity terms on the right-hand side based on our construction of the coupling. By construction we have $r(i) \geq r(0)$ which implies that $r(i) - r(0) \in \{0, 1\}^{|\mathcal{H}_A| \cdot |\mathcal{H}_P|}$. Note that

$$\begin{aligned} O^{r(i)} O^{r(0)} |j, \omega, c\rangle &= |j, \omega, c + (r(i))_j(\omega) + (r(0))_j(\omega)\rangle = |j, \omega, c + (r(i))_j(\omega) - (r(0))_j(\omega)\rangle \\ &= O^{r(i) - r(0)} |j, \omega, c\rangle. \end{aligned} \quad (247)$$

Let us introduce the self adjoint projection $P^{r(i), r(0)}$ given by

$$P^{r(i), r(0)} |j, \omega, c\rangle = \mathbf{1}_{(r(i))_j(\omega) \neq (r(0))_j(\omega)} |j, \omega, c\rangle. \quad (248)$$

Note that then

$$(\text{Id} - P^{r(i),r(0)})O^{r(i)-r(0)}|j, \omega, c\rangle = (\text{Id} - P^{r(i),r(0)})|j, \omega, c\rangle. \quad (249)$$

We get, using the invariance of the fidelity under unitary transformations, $(O^{r(i)})^2 = \text{Id}$, and (247)

$$\begin{aligned} \sqrt{F}(\rho_{t+1}^{r(i)}, \rho_{t+1}^{r(0)}) &= \sqrt{F}(\mathcal{E}_{O^{r(i)}}(\rho_t^{r(i)}), \mathcal{E}_{O^{r(0)}}(\rho_t^{r(0)})) = \sqrt{F}(\rho_t^{r(i)}, \mathcal{E}_{O^{r(i)}} \circ \mathcal{E}_{O^{r(0)}}(\rho_t^{r(0)})) \\ &= \sqrt{F}(\rho_t^{r(i)}, \mathcal{E}_{O^{r(i)-r(0)}}(\rho_t^{r(0)})). \end{aligned} \quad (250)$$

Next we apply Lemma 8 to $O^{r(i)-r(0)}$ and $P^{r(i),r(0)}$. The assumptions of the lemma are satisfied because (249) and all $P^{r(i),r(0)}$ and O^r commute. Lemma 8 together with the last display imply

$$\sqrt{F}(\rho_{t+1}^{r(i)}, \rho_{t+1}^{r(0)}) \geq \sqrt{F}(\rho_t^{r(i)}, \rho_t^{r(0)}) - 2\sqrt{\text{tr}(P^{r(i),r(0)}\rho_t^{r(i)})\text{tr}(P^{r(i),r(0)}\rho_t^{r(0)})} \quad (251)$$

We obtain

$$\sqrt{F}(\rho_T^i, \rho_T^0) \geq 1 - 2 \sum_{r(i), r(0)} \sum_t p_i(r(i), r(0)) \sqrt{\text{tr}(P^{r(i),r(0)}\rho_t^{r(i)})\text{tr}(P^{r(i),r(0)}\rho_t^{r(0)})}. \quad (252)$$

Using (243) followed by Cauchy-Schwarz we get

$$\begin{aligned} \frac{1}{2} - \sqrt{\delta(1-\delta)} &\leq \sum_{r(i), r(0)} \sum_t p_i(r(i), r(0)) \sqrt{\text{tr}(P^{r(i),r(0)}\rho_t^{r(i)})\text{tr}(P^{r(i),r(0)}\rho_t^{r(0)})} \\ &\leq \left(\sum_{t, r(i), r(0)} p_i(r(i), r(0)) \text{tr}(P^{r(i),r(0)}\rho_t^{r(i)}) \right)^{\frac{1}{2}} \left(\sum_{t, r(i), r(0)} p_i(r(i), r(0)) \text{tr}(P^{r(i),r(0)}\rho_t^{r(0)}) \right)^{\frac{1}{2}}. \end{aligned} \quad (253)$$

We observe that by construction of the coupling and (244) we have

$$\begin{aligned} \sum_{r(i)} p_i(r(i), r(0)) P^{r(i),r(0)} |j, \omega, c\rangle &= \sum_{r(i)} p_i(r(i), r(0)) \mathbf{1}_{(r(i))_j(\omega) \neq (r(0))_j(\omega)} |j, \omega, c\rangle \\ &= p(r(0)) \mathbb{P}((r(i))_j(\omega) = 1 | r(0)) |j, \omega, c\rangle \\ &= p(r(0)) \mathbf{1}_{i=j} \mathbf{1}_{r(0)_i(\omega)=0} \frac{\Delta_i}{1-p_i} |j, \omega, c\rangle. \end{aligned} \quad (254)$$

And similarly, using (245)

$$\sum_{r(0)} p_i(r(i), r(0)) P^{r(i),r(0)} |j, \omega, c\rangle = p(r(i)) \mathbf{1}_{i=j} \mathbf{1}_{r(i)_i(\omega)=1} \frac{\Delta_i}{p_0} |j, \omega, c\rangle. \quad (255)$$

As before, we define P_i to be the projection on arm i , i.e., $P_i |j, \omega, c\rangle = \delta_{ij} |j, \omega, c\rangle$. We conclude that

$$\sum_{r(i)} p_i(r(i), r(0)) \text{tr}(P^{r(i),r(0)}\rho_t^{r(0)}) \leq p(r(0)) \frac{\Delta_i}{\eta} \text{tr}(P_i \rho_t^{r(0)}), \quad (256)$$

$$\sum_{r(0)} p_i(r(i), r(0)) \text{tr}(P^{r(i),r(0)}\rho_t^{r(i)}) \leq p(r(i)) \frac{\Delta_i}{\eta} \text{tr}(P_i \rho_t^{r(i)}). \quad (257)$$

Combining this with (253) and (242) we obtain

$$\begin{aligned} \frac{1}{2} - \sqrt{\delta(1-\delta)} &\leq \left(\frac{\Delta_i}{\eta} \sum_{t, r(i)} p(r(i)) \text{tr}(P_i \rho_t^{r(i)}) \right)^{\frac{1}{2}} \left(\frac{\Delta_i}{\eta} \sum_{t, r(0)} p(r(0)) \text{tr}(P_i \rho_t^{r(0)}) \right)^{\frac{1}{2}} \\ &= \frac{\Delta_i}{\eta} \left(\sum_t \text{tr}(P_i \rho_t^i) \right)^{\frac{1}{2}} \left(\sum_t \text{tr}(P_i \rho_t^0) \right)^{\frac{1}{2}} \leq \frac{\Delta_i \sqrt{T}}{\eta} \left(\sum_t \text{tr}(P_i \rho_t^0) \right)^{\frac{1}{2}}. \end{aligned} \quad (258)$$

We square this relation divide by Δ_i^2 and sum over $i > 1$ and get

$$\left(\frac{1}{2} - \sqrt{\delta(1-\delta)}\right)^2 \sum_{i>1} \Delta_i^{-2} \leq \frac{T}{\eta^2} \sum_i \sum_t \text{tr}(P_i \rho_t^0) = \frac{T}{\eta^2} \sum_t \text{tr}(\rho_t^0) = \frac{T^2}{\eta^2}. \quad (259)$$

This implies

$$T \geq \frac{1}{2} \eta (1 - 2\sqrt{\delta(1-\delta)}) \sqrt{\sum_{i>1} \Delta_i^{-2}}. \quad (260)$$

□

J Proof of Theorem 8

Proof. We assume that we work on a Hilbert space $\mathcal{H} = \mathcal{H}_I \otimes \mathcal{H}_S \otimes \mathcal{H}_A$ where $\mathcal{H}_I$ is the input space, $\mathcal{H}_S$ is a single qubit state space and $\mathcal{H}_A$ consists of T ancilla qubits where $T = \mathcal{O}(\sqrt{N}/p)$ will be defined below in (269). We assume that the initial state of $\mathcal{H}_I$ is $|s\rangle = \sqrt{N}^{-1} \sum |i\rangle$ and all remaining qubits are in state $|0\rangle$. The algorithm consists of applying in turn the three operators $\mathcal{F}$, a controlled Grover-diffusion

$$U_\omega = \text{Id} \otimes |0\rangle\langle 0| \otimes \text{Id} + (2|s\rangle\langle s| - \text{Id}) \otimes |1\rangle\langle 1| \otimes \text{Id}, \quad (261)$$

and a swap operation S_t that swaps the state qubit with the t -th ancilla qubit. We can rewrite this concisely as

$$(\mathcal{E}_{S_T} \circ \mathcal{E}_{U_\omega} \circ \mathcal{F}) \circ \dots \circ (\mathcal{E}_{S_1} \circ \mathcal{E}_{U_\omega} \circ \mathcal{F})(\rho). \quad (262)$$

We denote by $\tilde{O}_i$ the standard oracle on $\mathcal{H}_I$ and denote by $G = (2|s\rangle\langle s| - \text{Id})\tilde{O}_i$ the map from Grover's algorithm acting on $\mathcal{H}_I$. We claim that the final state of the algorithm is

$$\rho_T = \sum_{x \in \{0,1\}^T} p^{|x|_1} (1-p)^{T-|x|_1} \left| G^{|x|_1} s, 0, x, 0^{A-T} \right\rangle \left\langle G^{|x|_1} s, 0, x, 0^{A-T} \right|. \quad (263)$$

The proof of this is by induction, indeed, we note

$$\begin{aligned} & (\mathcal{E}_{S_T} \circ \mathcal{E}_{U_\omega} \circ \mathcal{E}) \left(\left| G^{|x|_1} s, 0, x, 0^{A-T} \right\rangle \left\langle G^{|x|_1} s, 0, x, 0^{A-T} \right| \right) \\ &= (\mathcal{E}_{S_T} \circ \mathcal{E}_{U_\omega}) \left(p \left| \tilde{O} G^{|x|_1} s, 1, x, 0^{A-T} \right\rangle \left\langle \tilde{O} G^{|x|_1} s, 1, x, 0^{A-T} \right| \right. \\ &\quad \left. + (1-p) \left| G^{|x|_1} s, 0, x, 0^{A-T} \right\rangle \left\langle G^{|x|_1} s, 0, x, 0^{A-T} \right| \right) \\ &= \mathcal{E}_{S_T} \left(p \left| G G^{|x|_1} s, 1, x, 0^{A-T} \right\rangle \left\langle G G^{|x|_1} s, 1, x, 0^{A-T} \right| \right. \\ &\quad \left. + (1-p) \left| G^{|x|_1} s, 0, x, 0^{A-T} \right\rangle \left\langle G^{|x|_1} s, 0, x, 0^{A-T} \right| \right) \\ &= p \left| G G^{|x|_1} s, 0, x, 1, 0^{A-T-1} \right\rangle \left\langle G G^{|x|_1} s, 0, x, 1, 0^{A-T-1} \right| \\ &\quad + (1-p) \left| G^{|x|_1} s, 0, x, 0^{A-T} \right\rangle \left\langle G^{|x|_1} s, 0, x, 0^{A-T} \right| \end{aligned} \quad (264)$$

which implies (263). The reduced density matrix on $\mathcal{H}_I$ is

$$\rho_T^I = \sum_{k=0}^T p^k (1-p)^{T-k} |G^k s\rangle \langle G^k s|. \quad (265)$$

The classical analysis of the Grover algorithm proves

$$G^k |s\rangle = \cos\left(\frac{2k+1}{2}\theta\right) |s'\rangle + \sin\left(\frac{2k+1}{2}\theta\right) |i\rangle \quad (266)$$

where $s' = \sqrt{N-1}^{-\frac{1}{2}} \sum_{j \neq i} |i\rangle$ and

$$\theta = 2 \arccos\left(\sqrt{(N-1)/N}\right) = 2 \arcsin\left(\sqrt{N}^{-1}\right). \quad (267)$$

We remark that $\theta/2 \approx \sqrt{N}^{-1}$ as $N \rightarrow \infty$. Note that for

$$k \in I = (\pi/(4\theta) - 1/2, 3\pi/(4\theta) - 1/2) \quad (268)$$

we have $\sin\left(\frac{2k+1}{2}\theta\right)^2 \geq 1/2$. Let

$$T = \lfloor \pi/(2\theta p) \rfloor. \quad (269)$$

It remains to be shown that with probability at least $1/2$ a $\text{Bin}(T, p)$ distributed variable is contained in the interval I .

Using that the variance of $X \sim \text{Bin}(T, p)$ is $Tp(1-p)$ we can bound

$$\mathbb{P}(|X - pT| > pT/8) \leq \frac{64\mathbb{E}(|X - pT|^2)}{p^2 T^2} \leq \frac{64}{pT} \leq \frac{256\theta}{\pi} \leq \frac{1}{2} \quad (270)$$

for N sufficiently large. Moreover, $|X - pT| < pT/8$ implies

$$X \in \left(\frac{3pT}{8}, \frac{5pT}{8}\right) \subset \left(\frac{3\pi}{8\theta} - 1, \frac{5\pi}{8\theta}\right) \subset \left(\frac{\pi}{4\theta} - 1/2, \frac{3\pi}{4\theta} - 1/2\right) \quad (271)$$

for N sufficiently large. This ends the proof. $\square$

K Analysis of reusable oracles

Here we sketch a proof of Theorem 5. The proof essentially relies on the algorithm given in [17]. The only building block that needs to be changed is the gapped amplitude estimation (Corollary 2 in [17]). Let us explain this algorithm along with the replacement based on oracles as in (10). Gapped amplitude estimation assumes we have access to an oracle O_p and its adjoint acting via

$$O_p |0\rangle = \sqrt{p} |1\rangle + \sqrt{1-p} |0\rangle = |\text{coin}_p\rangle. \quad (272)$$

Then the following result holds.

Lemma 16 (Corollary 2 in [17]). *For $\varepsilon > 0$, $l \in [0, 1]$ and $\delta > 0$ there is a unitary procedure with $\mathcal{O}(\varepsilon^{-1} \ln(\delta))$ queries to O_p that prepares the state*

$$|\text{coin}_p\rangle (\alpha_0 |0\rangle |\psi_1\rangle + \alpha_1 |1\rangle |\psi_2\rangle) \quad (273)$$

with $\alpha_0, \alpha_1 \in [0, 1]$ and $\alpha_1 \leq \delta$ if $p \leq l - 2\varepsilon$ and $\alpha_0 \leq \delta$ if $p \geq l - \varepsilon$.

We replace this by classical estimation of the mean. Using tail bounds of random variables, we obtain the following simple variant of Hoeffding's inequality.

Lemma 17. *Let S_k be the sum of k independent random variables with distribution $\text{Ber}(p)$. For $\varepsilon > 0$ the bounds*

$$\mathbb{P}(S_k > k(p + \varepsilon)) \leq e^{-2\varepsilon^2 k}, \quad \mathbb{P}(S_k < k(p - \varepsilon)) \leq e^{-2\varepsilon^2 k} \quad (274)$$

hold.

Proof. This is just Hoeffding's inequality. $\square$

We then have the following corollary.

Corollary 6. *Given access to oracles as in (10) we can construct for any $\varepsilon, \delta > 0$ and $l \in [0, 1]$ an algorithm $\mathcal{A}$ that maps with probability at least $1 - \delta$ for all $1 \leq i \leq n$*

$$\mathcal{A} |i\rangle |0\rangle = |i\rangle |c\rangle \quad (275)$$

where $c = 1$ if $p_i < l - 2\varepsilon$ and $c = 0$ if $p_i > l - \varepsilon$ and $\mathcal{A}$ requires ε^{-2} oracle calls.

Proof. Set $k = \lceil 2\varepsilon^{-2} \ln(n/\delta) \rceil$. Then we consider the sequence

$$|i\rangle |0\rangle |0\rangle \rightarrow |i\rangle \left| \sum_{j=0}^k X_i^t \right\rangle |0\rangle \rightarrow |i\rangle \left| \sum_{j=0}^k X_i^t \right\rangle \left| \mathbf{1}_{\sum_{j=0}^k X_i^t < k(l-3\varepsilon/2)} \right\rangle \rightarrow |i\rangle |0\rangle \left| \mathbf{1}_{\sum_{j=0}^k X_i^t < k(l-3\varepsilon/2)} \right\rangle. \quad (276)$$

Where we in the first step sum the rewards using the oracles O_{X^t} , then set the flag register conditional on the sum and then uncompute the work register. This requires $2k$ oracle calls. Using Lemma 17 we conclude that for $p_i > l - \varepsilon$

$$\mathbb{P} \left(\sum_{j=0}^k X_i^t < l - 3\varepsilon/2 \right) \leq e^{-2k(\varepsilon/2)^2} \leq e^{-\ln(n/\delta)} = \frac{\delta}{n} \quad (277)$$

and a similar statement holds for $p_i < l - 2\varepsilon$. The union bound implies the statement. $\square$

Now we consider Theorem 5. Giving a full proof would require a large amount of notation that is not worth the effort. Thus, we give a very brief sketch of the argument and leave the details to the reader. Their proof relies on variable time algorithms [33] which we also do not introduce here. For readers not familiar with them, the following proof shall merely serve as a heuristic.

Sketch of the proof of Theorem 5. We apply the algorithm constructed in [17] but replace the gapped amplitude estimation by the algorithm constructed in Corollary 6. The main strategy used in their proof is to construct a variable time algorithm based on the gapped amplitude estimate that flags all arms whose reward is smaller than a given threshold. This allows to construct algorithms to count the number of flagged arms and rotate on the subspace of flagged arms using variable time amplitude amplification [33]. Those sub-routines can be used to first estimate p_1 and p_2 and then identify the best arm.

Their variable time algorithm based on the gapped amplitude estimate has query complexity on arm i is at most $\Delta_i^{-1} \log(1/a)$ with a polynomial in $\Delta_1 n$ and the l^2 averaged run-time thus amounts to

$$t_{\text{av}}^2 \leq C \frac{1}{n} \sum_{i=2}^n \Delta_i^{-2} \ln^2(1/a). \quad (278)$$

When relying on the algorithm from Corollary 6 we obtain the query complexity $\Delta_i^{-2} \log(n/\delta')$ for arm i where δ' denotes the bound on the failure probability for the algorithm and the l^2 averaged run-time then amounts to

$$t_{\text{av}}^2 \leq C \frac{1}{n} \sum_{i=2}^n \Delta_i^{-4} \ln^2(n\delta'). \quad (279)$$

Then the variable time amplitude amplification gives an algorithm with success probability more than $1/2$ and query complexity $t_{\text{av}}/\sqrt{p_{\text{succ}}} \ln(t_{\text{max}})$ where p_{succ} denotes the success probability and t_{max} the maximal complexity of the initial variable algorithm (there is another term $t_{\text{max}} \ln(t_{\text{max}})$ which is smaller in our case).

Since the initial success probability is n^{-1} we obtain the query complexity bound

$$T \leq t_{\text{av}} \sqrt{n} \ln(t_m) \leq C \left(\sum_{i=2}^n \Delta_i^{-4} \right)^{\frac{1}{2}} \ln(n/\delta') \ln(t_{\text{max}}). \quad (280)$$

We need to apply a total of $\mathcal{O}(\ln(\Delta_1^{-1}))$ such amplified variable time algorithms (essentially we perform binary search to find $l_L < l_R$ such that $p_2 < l_L - \Delta_2/4$, $l_R + \Delta_2/4 < p_1$, and $l_L + \Delta_2/4 < l_R$). To ensure that all constructed oracles as in Corollary 6 succeed with probability more than $1 - \delta$ it is thus sufficient to pick $\delta' = c\delta/\ln(\Delta_1^{-1})$. This together with $t_{\max} = \tilde{\mathcal{O}}(\Delta_1^{-2})$ ends the proof sketch. $\square$