Post-Quantum Succinct Arguments: Breaking the Quantum Rewinding Barrier

Alessandro Chiesa
alexch@berkeley.edu
UC Berkeley

Fermi Ma
fermima@alum.mit.edu
Princeton and NTT Research

Nicholas Spooner
nspooner@bu.edu
Boston University

Mark Zhandry
mzhandry@gmail.com
Princeton and NTT Research

June 3, 2021

Abstract

We prove that Kilian's four-message succinct argument system is post-quantum secure in the standard model when instantiated with any probabilistically checkable proof and any collapsing hash function (which in turn exist based on the post-quantum hardness of Learning with Errors). This yields the first post-quantum succinct argument system from any falsifiable assumption.

At the heart of our proof is a new quantum rewinding procedure that enables a reduction to repeatedly query a quantum adversary for accepting transcripts as many times as desired. Prior techniques were limited to a *constant* number of accepting transcripts.

Keywords: succinct arguments; post-quantum cryptography; quantum rewinding

Contents

1 Introduction

Quantum computers pose a growing threat to cryptography. Fully realized, quantum computers would enable an attacker to break the computational assumptions underlying many of today's public-key cryptosystems [Sho94]. Fortunately, a number of plausibly *quantum-secure* computational assumptions have emerged (e.g., lattice assumptions [Reg05]) providing a foundation for secure cryptography in a post-quantum era. But post-quantum cryptography requires more than quantum-safe assumptions: it also needs *security reductions* compatible with quantum attackers. While some classical security reductions directly translate to the quantum setting, many other security reductions do not translate because they are not compatible with quantum attackers.

Kilian's protocol [Kil92] is a fundamental result in cryptography for which no security reduction compatible with quantum attackers is known. Kilian's protocol is the canonical construction of a *succinct argument*: it uses a collision-resistant hash function to transform any probabilistically checkable proof (PCP) into an interactive protocol that achieves an exponential improvement in communication complexity over just sending the PCP. This comes at the cost of *computational* soundness, i.e., fooling the verification procedure of the protocol is intractable, not impossible. The security reduction against a classical attacker is via a *rewinding argument*: the attacker's state is saved midway through the protocol execution, and the attacker is run from this state many times to obtain many (succinct) protocol executions, from which the (long) PCP string can be extracted.

Alarmingly, Kilian's security reduction completely falls apart if the attacker has a quantum computer! The reduction has access to only a single copy of the attacker's state, due to the *no-cloning theorem*. Moreover, since quantum measurements are *destructive*, any attempt to measure the attacker's response may irreversibly damage the attacker's state, potentially rendering it useless.

Translating rewinding-based security reductions to the quantum setting has proved difficult (see e.g., [ARU14]). While there has been some progress on developing quantum techniques tailored to specific use cases [Wat06, Unr12, Unr16b], these techniques are not broadly applicable. Importantly, existing quantum rewinding techniques are limited to recording a *constant* number of attacker responses. This is particularly problematic for Kilian's protocol and beyond: all known techniques for reducing security of a succinct argument to an underlying (falsifiable) assumption require the reduction to record a super-constant (and typically polynomial) number of attacker responses.[1]

One way to avoid rewinding security reductions for succinct arguments is to rely on strong cryptographic assumptions. Kilian's protocol can be proved secure via a straightline (non-rewinding) extractor when ported to the random oracle model, and its security in the quantum random oracle model [BDF+11] follows from prior work [CMS19]. Beyond Kilian's protocol, there are constructions of succinct arguments that are proved secure directly from underlying post-quantum "knowledge" assumptions [BISW17, BISW18, GMNO18], but these assumptions are not falsifiable.[2]

In sum, the following question remains open:

Do post-quantum succinct arguments exist under standard assumptions?

1.1 Our results

We answer the question affirmatively by proving that Kilian's protocol is post-quantum secure, provided the underlying hash function is *collapsing* [Unr16b].

[1] Even if a classical security proof relies on an explicitly post-quantum assumption (e.g., [BBC+18, BLNS20]) this does not translate to *provable* post-quantum security as the rewinding security reduction is not quantum-compatible.

[2] See [Nao03, GW11] for further discussion on falsifiable assumptions.

Theorem 1.1 (Kilian's protocol is post-quantum secure)**.** *Kilian's protocol is a post-quantum succinct argument when instantiated with any PCP and any collapsing hash function. Moreover, if the underlying PCP is a proof of knowledge, Kilian's protocol is a post-quantum succinct argument of knowledge.*

Since collapsing hash functions are implied by post-quantum lossy functions [Unr16a], which exist assuming the quantum hardness of Learning with Errors (QLWE), we obtain post-quantum succinct arguments for all of NP from the same assumption. This is the first construction of post-quantum succinct arguments from *any* falsifiable assumption.

Corollary 1.2 (Post-quantum succinct arguments from QLWE)**.** *Assuming quantum hardness of LWE (QLWE), there exist post-quantum succinct arguments (of knowledge) for all of* NP*.*

The core of our proof is a new quantum extraction procedure that enables a reduction to record the prover's responses for an *arbitrary* number of random challenges. This significantly improves over prior work, which was limited to recording the prover's responses for a *constant* number of random challenges [Unr12, Unr16b, DFMS19].

Our extraction procedure applies not only to Kilian's protocol, but any *collapsing protocol* [Unr16b, LZ19, DFMS19]. A collapsing protocol refers to any public-coin interactive argument with the guarantee, roughly, that any (unitary) prover that only gives accepting responses cannot detect if its last response is measured. We show Kilian's protocol has this guarantee if it is instantiated with a collapsing hash function.

Theorem 1.3 (Quantum rewinding, informal)**.** *Given black-box access to any quantum adversary for a collapsing protocol, there is an efficient procedure to repeatedly query the adversary on random challenges and record an arbitrary number of accepting transcripts.*

Beyond our primary application to Kilian's protocol, our quantum rewinding procedure also implies that any k-special sound collapsing protocol is a post-quantum argument of knowledge, for any polynomially-bounded k.

Optimal knowledge error. Our rewinding technique achieves asymptotically optimal knowledge error. As an immediate application, our technique improves a previous result due to [Unr12, Unr16b], who showed that if a quantum attacker in a 2-special sound collapsing sigma protocol has success probability ε, then there is an extractor that can output a witness with probability $\varepsilon \cdot (\varepsilon^2 - 1/C)$, where C is the size of the challenge space. In particular, there was previously no guarantee for $1/C \leq \varepsilon \leq 1/\sqrt{C}$. Our techniques yield an extractor running in time $\mathrm{poly}(\lambda, 1/\varepsilon)$ that (given ε as input) outputs a witness with probability $\Omega(\varepsilon)$ provided that $\varepsilon \geq (1+\delta)/C$ for any constant $\delta > 0$.

2 Technical overview

2.1 Kilian's protocol

Kilian's protocol compiles any *probabilistically checkable proof* (PCP) into an interactive protocol using a Merkle tree built from a collision-resistant hash function. Recall that a PCP is a type of NP proof π that can be verified by reading only a few random positions [BFLS91, FGL+91, AS98, ALM+98]. The collision-resistant hash function enables the argument prover to send a *succinct* Merkle tree commitment to the PCP π that it can later open on any subset of positions Q with a short opening proof.

The protocol. Let $(\mathbf{P}_{\mathsf{PCP}}, \mathbf{V}_{\mathsf{PCP}})$ be a PCP proof system for an NP relation $\mathfrak{R}$, and let $\{H_\lambda\}_\lambda$ be a family of collision-resistant hash functions. The argument prover P and argument verifier V both receive as input the security parameter λ and an instance x, while the prover additionally receives a corresponding witness w (such that $(x, w) \in \mathfrak{R}$). They interact as follows.

1. V samples a collision-resistant hash function $h_{\mathsf{CRHF}} \leftarrow H_\lambda$ and sends it to P.
2. P computes a PCP string $\pi \leftarrow \mathbf{P}_{\mathsf{PCP}}(x, w)$, uses h_{CRHF} to generate a Merkle tree commitment $\mathsf{cm} \leftarrow \mathsf{Merkle.Commit}(h_{\mathsf{CRHF}}, \pi)$ to π, and sends cm to V.
3. V samples random coins $r \leftarrow R$ for the PCP verifier $\mathbf{V}_{\mathsf{PCP}}$ and sends them to P.
4. P computes the PCP indices Q that $\mathbf{V}_{\mathsf{PCP}}(x; r)$ would query, generates a Merkle opening proof pf for $\pi[Q]$, and sends the response $z \coloneqq (\pi[Q], \mathsf{pf})$ to V.[3]

Once the interaction is complete, V accepts if: (1) pf is a valid Merkle opening of cm to $\pi[Q]$ on indices Q; and (2) $\pi[Q]$ is accepted by the PCP verifier $\mathbf{V}_{\mathsf{PCP}}(x; r)$. Kilian's protocol is *publicly verifiable*: one can compute whether V accepts given only the instance x and the four-message transcript $(h_{\mathsf{CRHF}}, \mathsf{cm}, r, z)$.

The classical security reduction. Kilian's protocol ensures that an efficient extractor, given a malicious classical prover $\tilde{P}$ that convinces V with success probability 2ε, can output with overwhelming probability a PCP π such that $\Pr[\mathbf{V}^{\pi}_{\mathsf{PCP}}(x)] \geq \varepsilon/2$; the particular constants here are chosen to simplify the presentation in the following steps.

The extractor works by running $\tilde{P}$ through the first round of the protocol, obtaining a transcript prefix $\tau = (h_{\mathsf{CRHF}}, \mathsf{cm})$ and $\tilde{P}$'s intermediate state state_τ. Call state_τ "ε-good" if

$$\Pr\left[V(\tau, r, z) = 1 \;\middle|\; \begin{array}{r} r \leftarrow R \\ z \leftarrow \tilde{P}(\mathsf{state}_\tau, r) \end{array} \right] \geq \varepsilon \ .$$

By Markov's inequality, state_τ is ε-good with probability at least ε. If state_τ is ε-good, the extractor constructs a PCP proof π as follows.

Start with $\pi \coloneqq 0^\ell$ where ℓ is the PCP proof length. Repeat the loop:
1. Choose $r \leftarrow R$ uniformly at random.
2. Run $z \leftarrow \tilde{P}(\mathsf{state}_\tau, r)$.
3. If $V(\tau, r, z) = 1$, parse z as $(\pi'[Q], \mathsf{pf})$. Update π to match π' at the positions in Q.

[3] The Merkle opening for a PCP index q consists of the hash values of every vertex adjacent to the path from q to the root; the Merkle opening proof pf for a set of PCP indices Q consists of the Merkle openings for each $q \in Q$.

If the PCP has alphabet Σ and proof length ℓ, one can show that if the extractor records $k = 6\ell \cdot \log(2|\Sigma|)$ challenge-response pairs $(r_1, z_1), \ldots, (r_k, z_k)$ for distinct challenges r_i, then with probability $1 - \mathrm{negl}(\lambda)$ the PCP string π satisfies $\Pr[\mathbf{V}^{\pi}_{\mathsf{PCP}}(x)] \geq \varepsilon/2$.

This guarantee implies the *classical* security of Kilian's protocol. For instance, if the PCP system has negligible soundness error then the interactive argument has negligible soundness error.

2.2 Our approach to post-quantum security of Kilian's protocol

In this work, we prove that if the collision-resistant hash function h_{CRHF} is a *collapsing hash function* [Unr16b], then Kilian's protocol, without any additional modifications, is secure against malicious quantum provers. At a very high level, our security proof takes the following steps:

1. **Kilian's protocol is collapsing.** We prove that Kilian's protocol is a *collapsing protocol* in the sense of [LZ19, DFMS19] when the underlying hash function is collapsing; we elaborate on collapsing protocols in Section 2.3.

2. **Collapsing protocols admit quantum rewinding.** We devise a general-purpose quantum extraction procedure for collapsing protocols that enables efficiently recording any desired number of malicious prover responses. This step is our main technical contribution.

Organization. We discuss the importance of the collapsing notion in Section 2.3, but will otherwise defer the details of Step 1 to the body of the paper, since proving that Kilian's protocol is collapsing is a straightforward application of techniques from [Unr16b].

Step 2 is the primary focus of this technical overview. We summarize prior work on rewinding for collapsing protocols in Section 2.3 and explain in Section 2.4 why existing techniques are insufficient for Kilian. We then describe our extraction procedure in Sections 2.5 to 2.7.

2.3 Prior quantum techniques

We discuss prior techniques for recording responses of a malicious quantum prover in a classical interactive (public-coin) protocol. While prior works did not explicitly focus on Kilian's protocol, the abstract setting is the same. A reduction runs a malicious prover $\tilde{P}$ up to the final round of the protocol, obtaining a fixed transcript prefix τ and corresponding prover state state_τ. Assuming that $\tilde{P}(\mathsf{state}_\tau, \cdot)$ answers a random challenge $r \leftarrow R$ with *success probability* ε, the goal is to obtain some number k of accepting transcripts $(\tau, r_1, z_1), \ldots, (\tau, r_k, z_k)$ with the same prefix τ.

In the classical setting, this is an elementary task. By repeatedly sampling random challenges $r \leftarrow R$ and running $z \leftarrow \tilde{P}(\mathsf{state}_\tau, r)$, we can record any desired number of independent and identically distributed transcripts where an ε-fraction of them are accepting. Put another way:

> Given $\tilde{P}$ and state_τ, one can record k *accepting* transcripts for any desired k with probability 1 in expected time k/ε.

In the quantum setting, it is unlikely that such a statement holds: if state_τ is a quantum state $|\psi\rangle$, it is not possible in general to run $\tilde{P}(\mathsf{state}_\tau, \cdot)$ multiple times independently. This is because any measurement applied by $\tilde{P}$ may irreversibly alter the state. Indeed, Ambainis, Rosmanis, and Unruh [ARU14] show that this statement can be false relative to a (quantum) oracle, even if (P, V) is classically secure.

Collapsing protocols. Nevertheless, there *is* a class of protocols for which the statement holds in a limited sense. A public-coin interactive argument is a *collapsing protocol* [Unr16b, DFMS19, LZ19] if, given any last-round challenge r, an efficient prover which produces a superposition $|\phi\rangle$ of *accepting responses* cannot distinguish between $|\phi\rangle$ and the state that results after measuring the response in the computational basis.[4] We remark that [DFMS19, LZ19] defined collapsing protocols in the context of three-round sigma protocols, but the notion easily extends to public-coin interactive arguments.

For any collapsing protocol (P, V), Unruh's lemma [Unr12, DFMS19] gives a weaker version of the above statement. Suppose a malicious $\tilde{P}$ with state $|\psi\rangle$ has initial success probability ε, i.e., $\tilde{P}(|\psi\rangle, r)$ outputs an accepting response z on a random $r \leftarrow R$ with probability ε. Then Unruh's lemma gives the following guarantee:

> Given $\tilde{P}$ and $|\psi\rangle$, one can record k *accepting* transcripts for any desired k with probability $O(\varepsilon^{2k-1})$.

This $O(\varepsilon^{2k-1})$ probability, which does not appear in the classical statement, is over the randomness of the challenges and any quantum measurements the malicious prover performs. Notice that for constant k, this probability is still large enough to obtain meaningful guarantees. However, security of Kilian's protocol needs, at a minimum, $k = \Omega(\ell/|Q|)$ where ℓ is the PCP length and $|Q|$ is the number of queries of the PCP verifier. Thus, Unruh's lemma is insufficient since the guarantee only holds with probability $\varepsilon^{\Omega(\ell/|Q|)}$, which is negligible for any PCP with useful parameters.

2.4 A closer look at Unruh's lemma

Unruh's lemma is a quantum information-theoretic statement about any collection of binary-outcome projective measurements $\{\mathsf{A}_r\}_{r\in R}$. We write binary-outcome projective measurements as $\mathsf{A}_r = (\Pi_r, \mathbf{I} - \Pi_r)$ where Π_r is associated with outcome 1, and $\mathbf{I} - \Pi_r$ with outcome 0.

Let $\mathsf{MixM}(\{\mathsf{A}_r\}_r)$ be the corresponding *mixture* of the projective measurements $\{\mathsf{A}_r\}_r$, i.e., the procedure that chooses $r \leftarrow R$ uniformly at random, applies measurement A_r, and outputs the outcome $b \in \{0, 1\}$. Unruh's lemma [Unr12, DFMS19] concerns the measurement outcomes obtained from sequential applications of $\mathsf{MixM}(\{\mathsf{A}_r\}_r)$.

> **Unruh's lemma**: For any state $|\psi\rangle$ and any collection of binary-outcome projective measurements $\{\mathsf{A}_r\}_{r\in R}$, if applying $\mathsf{MixM}(\{\mathsf{A}_r\}_r)$ to $|\psi\rangle$ returns 1 with probability ε, then starting from $|\psi\rangle$ and applying $\mathsf{MixM}(\{\mathsf{A}_r\}_r)$ for k times in succession returns 1 all k times with probability ε^{2k-1}.

To use this lemma in the context of an interactive protocol, for each r in the challenge space R one defines $\mathsf{A}_r = (\Pi_r, \mathbf{I} - \Pi_r)$ as follows. Let U_r be the unitary describing the (purified) operation of $\tilde{P}$ in the last round on verifier message r; let $\Pi_{V,r} \coloneqq \sum_{z, V(\tau,r,z)=1} |z\rangle\langle z|$ be the projection onto responses z that the verifier $V(\tau, r, \cdot)$ accepts; and finally set $\Pi_r \coloneqq U_r^\dagger \Pi_{V,r} U_r$.

Intuitively, A_r measures *whether* $\tilde{P}$ causes V to accept on challenge r. Therefore, the probability ε in Unruh's lemma (the probability $\mathsf{MixM}(\{\mathsf{A}_r\}_r)$ applied to $|\psi\rangle$ returns 1) is the probability that

[4] More precisely, $|\phi\rangle = \sum_{y,z} \alpha_{y,z} |y, z\rangle$ where each z in the superposition satisfies $V(\tau, r, z) = 1$ for some fixed partial transcript τ, and y is the state on other registers. *Measuring the response* means measuring the register containing z.

$\tilde{P}(|\psi\rangle, \cdot)$ successfully answers a random challenge $r \leftarrow R$ in the interactive protocol. We sometimes refer to ε as the *success probability* of $|\psi\rangle$.

Thus Unruh's lemma shows that it is possible to "observe" k accepting executions with probability ε^{2k-1}, in the following sense: whenever MixM returns 1, one can apply U_r for the r sampled by MixM, and measure the adversary's response register to obtain z such that (τ, r, z) is an accepting transcript. Importantly, because Unruh's lemma only concerns *binary-outcome* projective measurements, we require an additional *collapsing* property from the underlying protocol to (undetectably) *record* any accepting responses. Thus, applied to a *collapsing protocol*, Unruh's lemma implies an extractor can record k accepting transcripts with probability $\varepsilon^{2k-1} - \mathrm{negl}(\lambda)$, since this additional measurement of the response register is (computationally) undetectable when MixM returns 1.

Consecutive measurements can destroy a state. The ε^{2k-1} probability comes in part from the fact that Unruh's lemma only captures the probability that k *consecutive* trials succeed.[5] This is a strong requirement: even in the classical setting, k consecutive trials succeed with probability ε^k. Classically this can be resolved by performing $N = k/\epsilon$ trials to obtain roughly k successful trials. One might hope that this would also work in the quantum setting: perhaps repeatedly applying $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$ some $\mathrm{poly}(k, 1/\varepsilon)$ times suffices to obtain k successful trials overall.

Unfortunately, this does not work. Adapting a counterexample of Zhandry [Zha20, Section 5], suppose the initial state $|\psi\rangle$ is $|0\rangle$, and for any desired success probability ε, define each $\mathsf{A}_r = (\Pi_r, \mathbf{I} - \Pi_r)$ so that Π_r is the rank-one projection onto $\sqrt{\varepsilon}\,|0\rangle + \sqrt{1-\varepsilon}\,|r\rangle$. Clearly, MixM applied to $|\psi\rangle$ returns 1 with probability ε, but one can verify that if repeated applications of MixM use *distinct* challenges r, then the expected number of 1 outcomes is at most $1/(2-2\varepsilon)$ regardless of the number of trials; for small ε this is close to $1/2$. This counterexample is a barrier if there are a super-polynomial number of challenges, as each trial will use a distinct r with overwhelming probability. Note that, in this example, the bound $1/(2-2\varepsilon)$ arises because the (expected) success probability of the state after j trials is exponentially small in j. In other words, the repeated applications of MixM "damage" the state.

2.5 State recovery

Given the above discussion, a natural approach is to try to recover the original state after the application of $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$. In particular, it would suffice to build a procedure that would allow recovering a state $|\psi\rangle$ after it has been perturbed by some binary projective measurement B. In our setting, $|\psi\rangle$ corresponds to the malicious prover's intermediate state, and B is the measurement M_r applied by $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$. Applying M_r to $|\psi\rangle$ disturbs the state, leaving some post-measurement state $|\phi\rangle$, and our aim is to somehow return the state back to $|\psi\rangle$. If we could do this in general (for any efficient binary projective measurement B) this would enable "perfect" quantum rewinding.

Unfortunately, this is impossible in general, but to build intuition for our eventual approach, we will show how to achieve this assuming we have access to a hypothetical additional power. In particular, suppose we can perform the binary projective measurement

$$\mathsf{Equals}_{|\psi\rangle} = (\,|\psi\rangle\langle\psi|\,, \mathbf{I} - |\psi\rangle\langle\psi|)$$

onto the one-dimensional subspace spanned by the initial state $|\psi\rangle$. If $\mathsf{Equals}_{|\psi\rangle}$ returns the outcome 1, then the post-measurement state is $|\psi\rangle$. In the remainder of this section, we use $\mathsf{Equals}_{|\psi\rangle}$ to develop a procedure that recovers the state $|\psi\rangle$ with probability close to 1.

[5] Technically, ε^{2k-1} only applies for random uncorrelated challenges, which may not be distinct. Unruh also gives a bound that applies for distinct random challenges.

The qubit case. First we consider the case where $|\psi\rangle$ is a single qubit: $|\psi\rangle$ lies in the *two-dimensional* space $\mathbb{C}^2$. If $\mathsf{B} = (\Pi, \mathbf{I} - \Pi)$ is nontrivial, then $\Pi = |\phi\rangle\langle\phi|$ and $\mathbf{I} - \Pi = |\phi^\perp\rangle\langle\phi^\perp|$ for some pair of orthogonal states $|\phi\rangle, |\phi^\perp\rangle \in \mathbb{C}^2$. This is shown in Fig. 1.

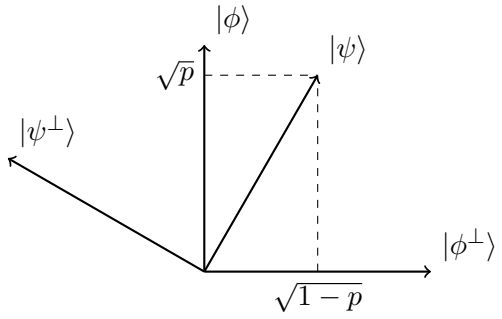

Figure 1: The quantum states $|\psi\rangle$ and $|\psi^\perp\rangle$ correspond to outcomes 1 and 0 of $\mathsf{Equals}_{|\psi\rangle} = (|\psi\rangle\langle\psi|, \mathbf{I} - |\psi\rangle\langle\psi|)$, respectively. The quantum states $|\phi\rangle$ and $|\phi^\perp\rangle$ correspond to outcomes 1 and 0 of $\mathsf{B} = (\Pi, \mathbf{I} - \Pi)$, respectively.

From Fig. 1 we see that $|\langle\phi|\psi\rangle|^2 = \langle\psi| |\phi\rangle\langle\phi| |\psi\rangle = \|\Pi|\psi\rangle\|^2 = p$. By making a suitable choice of phase, we can write

$$|\phi\rangle = \sqrt{p}\,|\psi\rangle + \sqrt{1-p}\,|\psi^\perp\rangle \quad ,$$
$$|\psi\rangle = \sqrt{p}\,|\phi\rangle + \sqrt{1-p}\,|\phi^\perp\rangle \quad .$$

Suppose that we have applied B to the state $|\psi\rangle$ and obtained the outcome 1. (The case of outcome 0 is symmetric.) The post-measurement state is then $|\phi\rangle$. A natural idea to recover the original state $|\psi\rangle$ is to apply $\mathsf{Equals}_{|\psi\rangle}$ to $|\phi\rangle$:

- With probability p, we obtain the outcome 1 and the state is $|\psi\rangle$.
- With probability $1 - p$ we obtain the outcome 0 and the state is $|\psi^\perp\rangle$ (which only holds because the space is two-dimensional).

In the first case we are done. But even in the second case we are not "stuck": if we apply B *again*, then with probability $1 - p$ we return to the state $|\phi\rangle$, and with probability p we move to the state $|\phi^\perp\rangle$. This leads to a "state recovery" procedure, which follows a technique first used by Marriott and Watrous for QMA amplification [MW05].[6] After potentially disturbing the state $|\psi\rangle$ by applying B, we can recover $|\psi\rangle$ by simply alternating the measurements

$$\mathsf{Equals}_{|\psi\rangle}, \mathsf{B}, \mathsf{Equals}_{|\psi\rangle}, \mathsf{B}, \ldots$$

until $\mathsf{Equals}_{|\psi\rangle}$ returns 1, at which point the state must be $|\psi\rangle$. In fact, the state of the system and the measurement outcomes throughout the procedure are remarkably easy to characterize. For instance, the effect of each $\mathsf{Equals}_{|\psi\rangle}$ measurement can be deduced from Fig. 1:

- Applying $\mathsf{Equals}_{|\psi\rangle}$ to $|\phi\rangle$ returns 1 with probability p resulting in $|\psi\rangle$, and returns 0 with probability $1 - p$ resulting in $|\psi^\perp\rangle$.

[6] The goal of [MW05] was not to reconstruct a particular quantum state, but to estimate the probability p.

- Applying $\mathsf{Equals}_{|\psi\rangle}$ to $|\phi^{\perp}\rangle$ returns 0 with probability p resulting in $|\psi^{\perp}\rangle$, and returns 1 with probability $1-p$ resulting in $|\psi\rangle$.

The effect of B on $|\psi\rangle$ and $|\psi^{\perp}\rangle$ is analogous. Letting b_i denote the outcome of the i-th measurement, starting from $|\psi\rangle$ and applying $\mathsf{B}, \mathsf{Equals}_{|\psi\rangle}, \ldots$ in alternating fashion (now counting the initial B as part of the sequence), the outcome sequence $b_1, b_2, \ldots$ follows a classical distribution $\mathsf{MWDist}(p)$ (for "Marriott–Watrous"):

1. Initialize $b_0 = 1$ (the initial state $|\psi\rangle$ corresponds to the 1 outcome of $\mathsf{Equals}_{|\psi\rangle}$).
2. For each $i \in \mathbb{N}$, set $b_i := b_{i-1}$ with probability p, and $b_i := 1 - b_{i-1}$ otherwise.

With this characterization, we can analyze the procedure's running time. The procedure fails to terminate at the first application of $\mathsf{Equals}_{|\psi\rangle}$, corresponding to $b_2 = 0$, with probability $2p(1-p)$. If this occurs, the next application of $\mathsf{Equals}_{|\psi\rangle}$ returns 0 with probability $1-2p(1-p)$. Continuing with this argument, the probability the procedure fails to terminate after $2T$ total measurements is

$$2p(1-p)(1-2p(1-p))^{T-1} < 1/T \ ,$$

where the inequality holds for *any* probability p.

Extending to more qubits. The analysis above relies on the fact that, in two dimensions, the system throughout the alternating measurement procedure is easily seen to lie in one of the four states $\{|\psi\rangle, |\psi^{\perp}\rangle, |\phi\rangle, |\phi^{\perp}\rangle\}$. In higher dimensions, the behavior of the system is potentially more complex.[7] We can nevertheless prove that the procedure terminates after $2T$ measurements with probability at most $1/T$.

To analyze the multi-qubit case, we use Jordan's lemma, a tool in quantum information theory that extends two-dimensional analyses of a pair of projectors to higher dimensions. Specifically, *any* two projectors $\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}$ induce a decomposition of the ambient Hilbert space into two-dimensional subspaces $\mathcal{S}_j$ such both Π_{A} and Π_{B} act as rank-one projectors within each subspace.[8]

More precisely, for each "Jordan subspace" $\mathcal{S}_j$, there exist orthogonal vectors $|v^{\mathsf{A}}_{j,1}\rangle, |v^{\mathsf{A}}_{j,0}\rangle$ that span $\mathcal{S}_j$, such that $\Pi_{\mathsf{A}}|v^{\mathsf{A}}_{j,1}\rangle = |v^{\mathsf{A}}_{j,1}\rangle$ and $\Pi_{\mathsf{A}}|v^{\mathsf{A}}_{j,0}\rangle = 0$; similarly, there exist orthogonal vectors $|v^{\mathsf{B}}_{j,1}\rangle, |v^{\mathsf{B}}_{j,0}\rangle$ that span $\mathcal{S}_j$ such that $\Pi_{\mathsf{B}}|v^{\mathsf{B}}_{j,1}\rangle = |v^{\mathsf{B}}_{j,1}\rangle$ and $\Pi_{\mathsf{B}}|v^{\mathsf{B}}_{j,0}\rangle = 0$. Defining the *eigenvalue* of $\mathcal{S}_j$ as $p_j := |\langle v_j|w_j\rangle|^2$, within each subspace $\mathcal{S}_j$ we recover a two-dimensional picture, as in Fig. 2. We refer to p_j as the "eigenvalue" of $\mathcal{S}_j$ because $|v^{\mathsf{A}}_{j,1}\rangle$ is an eigenvector of the Hermitian matrix $\Pi_{\mathsf{A}}\Pi_{\mathsf{B}}\Pi_{\mathsf{A}}$ with eigenvalue p_j (and $|v^{\mathsf{B}}_{j,1}\rangle$ is an eigenvector of $\Pi_{\mathsf{B}}\Pi_{\mathsf{A}}\Pi_{\mathsf{B}}$ with eigenvalue p_j).

[7] In the current setting, since $\mathsf{Equals}_{|\psi\rangle}$ projects onto a rank-one subspace, it turns out that even in higher dimensions the behaviour of this particular system will be two-dimensional, moving between states $|\psi\rangle, (\Pi - p\mathbf{I})|\psi\rangle, \Pi|\psi\rangle, (\mathbf{I}-\Pi)|\psi\rangle$ (appropriately normalized). Our more general treatment will be useful later on when we replace $\mathsf{Equals}_{|\psi\rangle}$ with a projection onto a higher-dimensional subspace.

[8] There are also one-dimensional subspaces, which we ignore here for the purpose of exposition; in any case, these can be treated as "degenerate" two-dimensional subspaces.

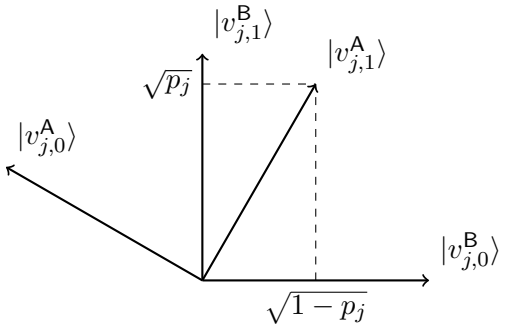

Figure 2: The states $|v^{\mathsf{A}}_{j,1}\rangle$ and $|v^{\mathsf{A}}_{j,0}\rangle$ correspond to 1 and 0 outcomes of $(\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$, respectively; $|v^{\mathsf{B}}_{j,1}\rangle$ and $|v^{\mathsf{B}}_{j,0}\rangle$ correspond to 1 and 0 outcomes of $(\Pi_{\mathsf{B}}, \mathbf{I} - \Pi_{\mathsf{B}})$, respectively.

By Jordan's lemma, a quantum state $|\phi\rangle$ satisfying $\Pi_{\mathsf{B}} |\phi\rangle = |\phi\rangle$ can be written as

$$|\phi\rangle = \textstyle\sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle \quad ,$$

where α_j is the amplitude of the state on the Jordan subspace $\mathcal{S}_j$. Starting from $|\phi\rangle$, if we alternate the binary projective measurements $(\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$ and $(\Pi_{\mathsf{B}}, \mathbf{I} - \Pi_{\mathsf{B}})$, then the distribution of the resulting measurement outcomes follows $\mathsf{MWDist}(p_j)$ with probability $|\alpha_j|^2$.

To see why this distribution arises, consider the projective measurement $\mathsf{M}_{\mathrm{Jor}} = (\Pi^{\mathrm{Jor}}_j)_j$ that projects onto the Jordan subspaces $\{\mathcal{S}_j\}_j$ and returns j as the outcome, i.e., each Π^{Jor}_j is a projection onto the $\mathcal{S}_j$ subspace. Since $\mathsf{M}_{\mathrm{Jor}}$ acts as the identity within every Jordan subspace $\mathcal{S}_j$, a consequence of Jordan's lemma is that $\mathsf{M}_{\mathrm{Jor}}$ commutes with both $(\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$ and $(\Pi_{\mathsf{B}}, \mathbf{I} - \Pi_{\mathsf{B}})$. Inserting the measurement $\mathsf{M}_{\mathrm{Jor}}$ at any point in the sequence of alternating measurements cannot change the earlier measurement outcomes, and the distribution above arises from commuting $\mathsf{M}_{\mathrm{Jor}}$ to the beginning of the procedure.

With Jordan's lemma in hand, our analysis of the "state recovery" procedure in the two-dimensional setting extends to higher dimensions by associating $(\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$ with $\mathsf{Equals}_{|\psi\rangle}$ and $(\Pi_{\mathsf{B}}, \mathbf{I} - \Pi_{\mathsf{B}})$ with B. Since the procedure's running time is determined solely by the measurement outcomes, we recover the original state $|\psi\rangle$ after $2T$ alternating measurements except with probability

$$\textstyle\sum_j |\alpha_j|^2 \cdot 2p_j(1-p_j)(1-2p_j(1-p_j))^{T-1} \leq \frac{1}{T} \sum_j |\alpha_j|^2 = 1/T \quad .$$

Summarizing, we obtain the following general lemma for binary projective measurements A, B:

> **Setup:** Fix measurements $\mathsf{A} = (\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$ and $\mathsf{B} = (\Pi_{\mathsf{B}}, \mathbf{I} - \Pi_{\mathsf{B}})$ and a state $|\psi\rangle$ in the span of Π_{A}. Apply B to $|\psi\rangle$ and let $|\phi\rangle$ be the post-measurement state.
>
> **Alternate:** Starting from $|\phi\rangle$, apply $\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B}, \ldots$ until A returns 1. The procedure requires $O(1)$ measurements in expectation.

In particular, if A is our hypothetical $\mathsf{Equals}_{|\psi\rangle}$ measurement, then after the procedure terminates, we recover the state $|\psi\rangle$.

2.6 State repair

Perhaps unsurprisingly, we cannot efficiently implement the measurement $\mathsf{Equals}_{|\psi\rangle}$, and in general we cannot *recover* the original state $|\psi\rangle$ [9]. However, our goal is to efficiently extract successful attacker responses, which "only" requires that the probability A_r for a random $r \leftarrow R$ returns 1 (the "success probability") does not significantly decay with repeated applications. One of our key observations is that we can satisfy this requirement *without* having to recover the original state.

Observation: Restoring the state's *success probability* suffices for extraction.

We refer to the process of restoring the success probability as *state repair*. Jumping ahead, the repaired state in our state repair procedure may be far in trace distance from the original state $|\psi\rangle$.

Below we explain how to modify the "state recovery" procedure from the previous subsection into a "state repair" procedure. Informally, we replace $\mathsf{Equals}_{|\psi\rangle}$ with a measurement $\mathsf{Test}_\varepsilon$ having a relaxed guarantee on post-measurement states: when $\mathsf{Test}_\varepsilon$ returns 1, the post-measurement state has the same *success probability* as $|\psi\rangle$.

Defining $\mathsf{Test}_\varepsilon$. To define a projective measurement $\mathsf{Test}_\varepsilon$ suitable for performing "state repair", it suffices to identify a linear space for which every $|\psi\rangle$ in the space has success probability at least ε. We achieve this by identifying a particular operator E with an extremely useful property: any eigenstate of E with eigenvalue p corresponds to a state $|\psi\rangle$ with success probability p. We then define $\mathsf{Test}_\varepsilon$ to be the projection onto the direct sum of eigenspaces of E with eigenvalue $p \geq \varepsilon$.

Our choice of E must somehow capture the probability that a random A_r for $r \leftarrow R$ returns 1 when applied to a state $|\psi\rangle$. Thus, a natural place to start is to consider the *purification* of $\mathsf{MixM}(\{\mathsf{A}_r\}_{r\in R})$, i.e., the procedure that applies M_r for random $r \leftarrow R$. For this, in addition to the original Hilbert space $\mathcal{H}$, we need an ancilla register $\mathcal{R}$. We initialize this register to a uniform superposition $|+_R\rangle$ over the indices $r \in R$. We then define a binary projective measurement CProj (for "controlled projection") that applies $\{\mathsf{M}_r = (\Pi_r, \mathbf{I} - \Pi_r)\}_r$ controlled on $\mathcal{R}$:

$$\mathsf{CProj} \coloneqq (\Pi_{\mathsf{CProj}}, \mathbf{I} - \Pi_{\mathsf{CProj}}) \text{ where } \Pi^{\mathsf{CProj}} \coloneqq \textstyle\sum_{r\in R} |r\rangle\langle r|^{\mathcal{R}} \otimes \Pi_r \ .$$

Letting $\mathsf{MixM}(\{\mathsf{M}_r\}_r; |\psi\rangle)$ denote the application of $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$ to $|\psi\rangle$, observe that applying CProj to $|+_R\rangle^{\mathcal{R}} \otimes |\psi\rangle$ and tracing out $\mathcal{R}$ is equivalent to $\mathsf{MixM}(\{\mathsf{M}_r\}_r; |\psi\rangle)$.

We remark that the measurement CProj represents a "superposition query" to the adversary $\tilde{P}(|\psi\rangle, \cdot)$. This is a qualitative departure from the techniques of [Unr12, DFMS19], which only make classical queries to the adversary. Superposition queries have been used in [VZ21] in the context of proofs of *quantum* knowledge. We find it interesting that superposition queries also arise in an essential way when extracting only classical knowledge.

We are now ready to define the operator E:

$$E \coloneqq |+_R\rangle\langle+_R|^{\mathcal{R}} \cdot \Pi_{\mathsf{CProj}} \cdot |+_R\rangle\langle+_R|^{\mathcal{R}} \text{ where } |+_R\rangle\langle+_R|^{\mathcal{R}} \text{ denotes } |+_R\rangle\langle+_R|^{\mathcal{R}} \otimes \mathbf{I}^{\mathcal{H}} \ .$$

As desired, any eigenstate of E with positive eigenvalue p is of the form $|+_R\rangle|\chi\rangle$ where $|\chi\rangle \in \mathcal{H}$ has success probability p:

$$\Pr\left[\mathsf{MixM}(\{\mathsf{M}_r\}; |\chi\rangle) = 1\right] = \|\Pi_{\mathsf{CProj}} |+_R\rangle |\chi\rangle\|^2 = (\langle+_R| \otimes \langle\chi|)E(|+_R\rangle \otimes |\chi\rangle) = p \ .$$

[9] One may notice that, for the setting of interactive arguments, $|\psi\rangle$ was generated by an efficient procedure. Nevertheless, there is no efficient procedure to re-generate the particular $|\psi\rangle$ that corresponds to the partial transcript seen so far. This is because $|\psi\rangle$ is the collapsed state leftover after measuring the prover's commitment message, and this may yield different outcomes every time.

We stress that this implication only goes in one direction, as it is *not true* that every state $|\psi\rangle$ with success probability p corresponds to an eigenstate $|+_R\rangle|\psi\rangle$ of E with eigenvalue p. The precise relationship is summarized in the following observation:

> **Key fact:** For every state $|\psi\rangle$ with success probability p, $|+_R\rangle|\psi\rangle$ can be written as a *linear combination of eigenstates of* E
>
> $$|+_R\rangle|\psi\rangle = \sum_j \alpha_j |+_R\rangle|\chi_j\rangle$$
>
> where each $|+_R\rangle|\chi_j\rangle$ has eigenvalue/success probability p_j, and $p = \sum_j |\alpha_j|^2 p_j$.

We now define Π_ε as the projector onto the span of eigenstates of E with eigenvalue at least ε. Let the corresponding binary-outcome measurement be $\mathsf{Test}_\varepsilon := (\Pi_\varepsilon, \mathbf{I} - \Pi_\varepsilon)$. Importantly, $\mathsf{Test}_\varepsilon$ satisfies the following properties.

- **Property 1: applied to any 2ε-successful state, $\mathsf{Test}_\varepsilon$ returns 1 with probability ε.** By the "key fact" above, any state $|+_R\rangle|\psi\rangle$ where $|\psi\rangle$ has success probability 2ε is a linear combination of eigenstates $\sum_j \alpha_j |+_R\rangle|\chi_j\rangle$ where $2\varepsilon = \sum_j |\alpha_j|^2 p_j$. By Markov's inequality, there must be at least probability mass ε on eigenstates with eigenvalue/success probability at least ε.
- **Property 2: when $\mathsf{Test}_\varepsilon$ returns 1, the post-measurement state is ε-successful.** This follows from the definition of Π_ε, since any state in the image of Π_ε is a linear combination of eigenstates $|+_R\rangle|\chi_j\rangle$ where every $|\chi_j\rangle$ has success probability at least ε.

A state repair procedure. We now present a state prepare procedure using $\mathsf{Test}_\varepsilon$. We stress that the following procedure is not yet sufficient to implement an efficient extraction procedure, since we have not specified how to implement $\mathsf{Test}_\varepsilon$.

Start with state $|+_R\rangle|\psi\rangle \in (\mathcal{R}, \mathcal{H})$ where $|\psi\rangle$ has success probability 2ε.

1. **Initialization.** Apply the measurement $\mathsf{Test}_\varepsilon$ and abort if the outcome is 0.
2. **Measure-and-repair.** Repeat the following loop as many times as desired.
   (a) (Measure step) Sample a random $r \leftarrow R$ and apply A_r to $\mathcal{H}$ to obtain an outcome b. Call this step "successful" if $b = 1$.
   (b) (Repair step) Repair the state by applying $\mathsf{Test}_\varepsilon, \mathsf{A}_r, \mathsf{Test}_\varepsilon, \mathsf{A}_r, \ldots$ until $\mathsf{Test}_\varepsilon$ outputs 1.

Since the state $|\psi\rangle$ at the beginning of the procedure has success probability at least 2ε, the initialization step aborts with probability at most $1 - \varepsilon$.

We now analyze the execution of this procedure conditioned on the event that the initialization step *does not abort*. We argue that the procedure can repeatedly iterate the measure-and-repair loop. By construction, the state after any (non-aborting) Initialization step or Repair step is in the span of Π_ε. Thus, the state at the beginning of the Measure step is always in the span of Π_ε. Since any state in the span of Π_ε is of the form $|+_R\rangle|\chi\rangle$ where $|\chi\rangle$ has success probability ε, the Measure step is equivalent to an application of $\mathsf{MixM}(\{\mathsf{A}_r\}_r)$ that succeeds with at least ε probability.

Recap. We summarize what our state repair procedure implies for extraction. Suppose we are given a malicious prover $\tilde{P}(|\psi\rangle, \cdot)$ for a collapsing interactive protocol who successfully answers a random challenge $r \leftarrow R$ with success probability ε. Moreover, assume that we can implement $\mathsf{Test}_\varepsilon$. Then for any desired $c \in \mathbb{N}$, if the initialization step does not abort, then we can repeat the measure-and-repair iteration c times and achieve the following:

- in each iteration we ask $\tilde{P}$ a random challenge $r \leftarrow R$, and record an accepting transcript (τ, r, z) with probability at least ε; and
- in expectation, the total number of measurements performed is $O(c)$.

While this is promising, we are far from done, because we do not know of a way to efficiently implement $\mathsf{Test}_\varepsilon$. Hence, in Section 2.7, we show how to replace $\mathsf{Test}_\varepsilon$ with an efficient measurement $\mathsf{ApproxTest}_\varepsilon$ that *approximates* the behavior of $\mathsf{Test}_\varepsilon$. While the idea behind $\mathsf{ApproxTest}_\varepsilon$ is natural, proving that $\mathsf{ApproxTest}_\varepsilon$ suffices for extraction is the most technically challenging part of this work.

2.7 Approximate state repair

Approximating $\mathsf{Test}_\varepsilon$. While we do not know how implement $\mathsf{Test}_\varepsilon$, we have *already* developed a way to *approximate* $\mathsf{Test}_\varepsilon$: the alternating measurements technique we used for state repair doubles as a way to *estimate* the success probability! Note that estimating success probability (not repairing the state) was the motivation for alternating measurements in [MW05, Zha20].

Let $|+_R\rangle|\chi_j\rangle$ be an eigenstate of $E = |+_R\rangle\langle+_R|^{\mathcal{R}} \Pi_{\mathsf{CProj}} |+_R\rangle\langle+_R|^{\mathcal{R}}$ with eigenvalue p_j; recall from Section 2.6 that $|\chi_j\rangle$ has success probability p_j.

An important observation is that the eigenspectrum of E corresponds to the decomposition of $(\mathcal{R}, \mathcal{H})$ induced by Jordan's lemma for Π_{CProj} and $|+_R\rangle\langle+_R|^{\mathcal{R}}$: any state in the span of $|+_R\rangle\langle+_R|^{\mathcal{R}}$ that is in the Jordan subspace $\mathcal{S}_j$ must be an eigenstate $|+_R\rangle|\chi_j\rangle$ of E with eigenvalue p_j.

Then, by the analysis in Section 2.5, if we start from $|+_R\rangle|\chi_j\rangle$ and apply the binary projective measurements $\mathsf{CProj} = (\Pi_{\mathsf{CProj}}, \mathbf{I} - \Pi_{\mathsf{CProj}})$ and $\mathsf{M}_{|+_R\rangle} = (|+_R\rangle\langle+_R|, \mathbf{I} - |+_R\rangle\langle+_R|)$ in an alternating fashion:

$$\mathsf{CProj}, \mathsf{M}_{|+_R\rangle}, \mathsf{CProj}, \mathsf{M}_{|+_R\rangle}, \ldots,$$

then the corresponding measurement outcomes $b_1, b_2, b_3, \ldots$ are distributed so that $\mathbf{1}_{b_i = b_{i+1}}$ (the indicator for the event $b_i = b_{i+1}$, where we define $b_0 \coloneqq 1$) is an independent Bernoulli random variable with expectation p_j for all $i \geq 0$.

Following [MW05, Zha20], this yields a simple, *non-projective* procedure $\mathsf{ApproxTest}_{\varepsilon,t}$:

Initial state: $|+_R\rangle|\psi\rangle$ for state $|\psi\rangle$ with success probability at least 2ε.

1. Apply $2t$ measurements $\mathsf{CProj}, \mathsf{M}_{|+_R\rangle}, \ldots, \mathsf{CProj}, \mathsf{M}_{|+_R\rangle}$. Denote the binary outcome of the i-th measurement by b_i and additionally set $b_0 \coloneqq 1$.
2. Compute $p \coloneqq \frac{1}{2t} \cdot |\{i \in \{1, \ldots, 2t\} : b_{i-1} = b_i\}|$ and output 1 if $p \geq \varepsilon$.

To analyze the distribution of outcomes from applying $\mathsf{ApproxTest}_{\varepsilon,t}$ to an arbitrary state of the form $|+_R\rangle|\psi\rangle$, we employ the method from Section 2.5 of projecting onto the Jordan subspaces $\{\mathcal{S}_j\}_j$ for the projectors Π_{CProj} and $|+_R\rangle\langle+_R|$. Since any state $|+_R\rangle|\psi\rangle$ can be written as a linear combination $\sum_j \alpha_j |+_R\rangle|\chi_j\rangle$ of eigenstates of E, the result of applying $\mathsf{ApproxTest}_{\varepsilon,t}$ to $|+_R\rangle|\psi\rangle$ can be described as follows, where $\mathsf{Test}_\varepsilon$ is included for comparison:

- $\mathsf{Test}_\varepsilon$: Sample j with probability $|\alpha_j|^2$, and then return 1 if $p_j \geq \varepsilon$ and 0 otherwise.
- $\mathsf{ApproxTest}_{\varepsilon,t}$: Sample j with probability $|\alpha_j|^2$; flip $2t$ independent Bernoulli random variables with parameter p_j; let p be the fraction of flips that return 1; output 1 if $p \geq \varepsilon$ and 0 otherwise.

Thus, we have from Section 2.6 a working extraction procedure based on $\mathsf{Test}_\varepsilon$, and now a way to efficiently approximate $\mathsf{Test}_\varepsilon$ to any desired precision using $\mathsf{ApproxTest}_{\varepsilon,t}$. However, turning this

intuition into a working extraction procedure requires overcoming a number of technical challenges, stemming from the fact that $\mathsf{ApproxTest}_{\varepsilon,t}$ as defined above is *not* a projective measurement.

Challenge: $\mathsf{ApproxTest}_{\varepsilon,t}$ is not projective. In Section 2.5 we claimed that if a state $|\psi\rangle$ initially in the span of some projector Π_{A} is disturbed by an binary-outcome measurement B, then by performing alternating measurements, we can return our state to the span of Π_{A} in $2T$ measurements except with probability $1/T$. It is not clear that such a statement holds if $\mathsf{A} = (\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$ is replaced by a *non-projective* measurement.

Concretely, we need to analyze the behavior of the alternating measurement procedure

$$\mathsf{ApproxTest}_{\varepsilon,t}, \mathsf{A}_r, \mathsf{ApproxTest}_{\varepsilon,t}, \mathsf{A}_r, \dots$$

where $\mathsf{ApproxTest}_{\varepsilon,t}$ *itself is an alternating measurements procedure*, i.e., $\mathsf{ApproxTest}_{\varepsilon,t}$ runs

$$\mathsf{CProj}, \mathsf{M}_{|+_R\rangle}, \mathsf{CProj}, \mathsf{M}_{|+_R\rangle}, \dots \ .$$

The core technical challenge is to prove that the guarantees of alternating measurements used in Section 2.6 extend to "nested" alternating measurements.

Can we appeal to trace distance? One might hope to show that for large t, the post-measurement states of $\mathsf{ApproxTest}_{\varepsilon,t}$ and $\mathsf{Test}_{\varepsilon}$ are close. If $\mathsf{ApproxTest}_{\varepsilon,t}|\psi\rangle$ were sufficiently close in trace distance to $\mathsf{Test}_{\varepsilon}|\psi\rangle$ for all $|\psi\rangle$, then we could show that any property of the procedure $\mathsf{Test}_{\varepsilon}, \mathsf{A}_r, \mathsf{Test}_{\varepsilon}, \mathsf{A}_r, \dots$ still applies if we swap out $\mathsf{Test}_{\varepsilon}$ for $\mathsf{ApproxTest}_{\varepsilon,t}$, up to a small loss.

Unfortunately, a simple example illustrates why such a claim about the trace distance is false. Suppose we have an eigenstate $|+_R\rangle|\chi_j\rangle$ of the operator E with eigenvalue $p_j = \varepsilon$. Then since $\mathsf{Test}_{\varepsilon}$ projects onto eigenspaces of E with eigenvalue $\geq \varepsilon$, applying $\mathsf{Test}_{\varepsilon}$ to this state returns 1 with probability 1. However, applying $\mathsf{ApproxTest}_{\varepsilon,t}$ returns 1 with essentially $1/2$ probability, since it performs ε-weighted coin flips and only accepts if the fraction of 1's is at least ε.

Expanding the Hilbert space. Since a trace distance argument is unlikely to work, the next idea is to simply force $\mathsf{ApproxTest}_{\varepsilon,t}$ to be projective by expanding the Hilbert space. The hope is that by making the measurement projective, we regain our ability to apply Jordan's lemma. Specifically, we introduce $2t$-qubit ancilla registers $\mathcal{L}$ to store the $2t$ outcomes of CProj and $\mathsf{M}_{|+_R\rangle}$, which we perform *coherently*, meaning that instead of actually performing the measurements, we apply corresponding unitaries to CNOT the measurement results onto the ancilla registers $\mathcal{L}$. To ensure the measurement is projective, we must also uncompute all the (coherent applications of) CProj and $\mathsf{M}_{|+_R\rangle}$ once we obtain the probability estimate p.

Technical challenge: $\mathsf{ApproxTest}_{\varepsilon,t}$ is only meaningful if $\mathcal{L}$ is $|0^{2t}\rangle$. Unfortunately, expanding the Hilbert space introduces a new problem. If $\mathsf{ApproxTest}_{\varepsilon,t}$ computes its estimate of p using a $2t$-qubit ancilla register $\mathcal{L}$, then we have to ensure the register $\mathcal{L}$ is set to $|0^{2t}\rangle$, or else the estimate of p, computed based on the contents of the $\mathcal{L}$ register, may be meaningless. A natural idea would be to ensure that, before any application of $\mathsf{ApproxTest}_{\varepsilon,t}$, we trace out the potentially non-zero registers $\mathcal{L}$ and manually reset them to $|0^{2t}\rangle$. However, doing this is equivalent to performing the original non-projective version of $\mathsf{ApproxTest}_{\varepsilon,t}$, and we would be back where we started.

Resolution: project $\mathcal{L}$ onto $|0^{2t}\rangle$. Instead we modify the measurement A_r (which originally acts as identity on the $\mathcal{L}$ registers) to additionally project $\mathcal{L}$ onto $|0^{2t}\rangle$. This modified measurement $\mathsf{A}_{r,b}$ returns 1 if and only if A_r returns b *and* the binary projective measurement of $\mathcal{L}$ onto $|0^{2t}\rangle$ returns 1; in particular, $\mathsf{A}_{r,b}$ is still a binary projective measurement. Proving that the state is repaired after the *projective* version of $\mathsf{ApproxTest}_{\varepsilon,t}$ returns 1 requires a very careful analysis of

the properties of the Jordan decomposition induced by (projective) $\mathsf{ApproxTest}_{\varepsilon,t}$ and $\mathsf{A}_{r,b}$. The analysis of this procedure is the most technical component of the paper; see Section 4.3 for details.

2.8 Quantum strategies for repeated games

Our quantum rewinding techniques can be cast in the language of *single-player games*, i.e., a referee asks a player a random question $r \leftarrow R$, the player responds with some z, and wins if $f(r, z) = 1$ for some predicate f. Mapped onto this setting, the quantum rewinding task is to transform any efficient quantum strategy for winning the game once into an efficient strategy that can win in many rounds in an n-fold sequential repetition of this game, where in each repetition the referee *only measures whether the player has won.* Importantly, we are only given one copy of the quantum state used by the one-time strategy.

In the context of rewinding, we set $f(r, z) \coloneqq V(\tau, r, z)$ to be the verifier predicate with partial transcript τ. The strategy of the prover in the last round of the protocol is then an efficient strategy for the one-time game. To obtain multiple accepting transcripts, a rewinding extractor plays the sequential repetition of the game. Note that by measuring $\mathcal{Z}$ in the computational basis if the player has won, the extractor obtains an accepting response z; collapsing ensures that this additional measurement is not detectable by an efficient strategy.

This gives a conceptually simple characterization of the quantum rewinding task, which may be of independent interest. In the body of the paper, we develop general techniques that apply to any single-player game (see Section 4).

2.9 Discussion: is collapsing necessary for Kilian's protocol?

Since collision-resistant hash functions (CRHFs) suffice in the classical setting, a natural question is whether Kilian's protocol (in its original formulation using Merkle trees) is post-quantum secure when instantiated with any post-quantum CRHF. We do not know the answer, but believe that the existing evidence points to collision resistance being *insufficient* for Kilian's protocol.

Ambainis et al. [ARU14] give a counter-example showing that, in general, collision resistance alone is likely not enough for rewinding in interactive protocols. The counter-example works by giving a construction of an *equivocal* hash function.[10] This is a hash function that is collision resistant, but where it is possible to break the security of the hash function as a commitment scheme. For example, it is possible to send a hash image y, and then upon receiving an arbitrary prefix z, "open" that image to a pre-image x of y with prefix z. Such equivocal hash functions do not exist classically, due to a rewinding argument, but Ambainis et al. [ARU14] show how to construct them relative to a quantum oracle. Amos et al. [AGKZ20] later give a construction relative to a classical oracle.

While Ambainis et al. use equivocal hash functions to give unsound interactive proofs, the results do not immediately apply to the case of Kilian's protocol. This is because Merkle trees do not necessarily preserve equivocality of the component hash function. In particular, equivocating Merkle trees would seem to require equivocating the underlying hash function on either the left half or the right half of the input. On the other hand, only a very short prefix can be equivocated by the existing works.[11]

[10] The terminology "equivocal" is due to [AGKZ20].

[11] Ambanis et al. allow for a richer class of equivocations than just prefixes, but they must still be short relative to the input length.

Nevertheless, we observe that a slight variant of Merkle trees *does* preserve the equivocality of the underlying hash function. Namely, if each node is obtained by hashing the children together with an arbitrarily long auxiliary string. By setting the length of the auxiliary strings sufficiently long, one can equivocate on a prefix long enough to arbitrarily choose the child nodes. This allows for full equivocality of Merkle trees, while still preserving collision resistance. More generally, it yields a *vector commitment* that is collision resistant, but equivocal and therefore insufficient for the post-quantum security of Kilian's protocol.

We leave as an interesting open question whether Kilian's protocol instantiated with vanilla Merkle trees using a post-quantum CRHF is sufficient for post-quantum security. We note, however, that if Kilian's protocol instantiated with a CRHF is *not* post-quantum secure, then it means the CRHF is not collapsing. As shown by Zhandry [Zha19], such a CRHF would yield strong cryptographic objects, namely "quantum lightning", which have no known instantiations under well-studied assumptions.[12]

[12]More precisely, Zhandry [Zha19] shows that non-collapsing CRHFs imply *infinitely-often secure* quantum lightning, a slightly weaker notion.

3 Preliminaries

The security parameter is denoted by λ. A function $f\colon \mathbb{N} \to [0,1]$ is *negligible*, denoted $f(\lambda) = \mathrm{negl}(\lambda)$, if it decreases faster than the inverse of any polynomial. A probability is *overwhelming* if is at least $1 - \mathrm{negl}(\lambda)$ for a negligible function $\mathrm{negl}(\lambda)$. For any positive integer n, let $[n] \coloneqq \{1, 2, \ldots, n\}$. For a set R, we write $r \leftarrow R$ to denote a uniformly random sample r drawn from R.

3.1 Concentration inequalities

We denote by $\mathsf{Bin}(n,p)$ the binomial distribution with n trials and success probability p (sum of n independent Bernoullis with parameter p). We use the following Chernoff bounds.

Proposition 3.1 (additive Chernoff bound)**.** *For $\delta, \epsilon > 0$, define $n_{\epsilon,\delta} \coloneqq \frac{\log(1/2\delta)}{2\epsilon^2}$. If $n \geq n_{\epsilon,\delta}$ then*

$$\Pr_{X \leftarrow \mathsf{Bin}(n,p)}\left[p - \epsilon \leq \frac{X}{n} \leq p + \epsilon\right] \geq 1 - \delta \ .$$

Proposition 3.2 (multiplicative Chernoff bound)**.** *Let $x_1, \ldots, x_N \in \{0,1\}$ and define $\mu \coloneqq \frac{K}{N}\sum_{i=1}^{N} x_i$. Let $X_1, \ldots, X_K$ be independent uniformly random samples from $x_1, \ldots, x_N$. Then*

$$\Pr\left[\sum_{i=1}^{K} X_K \geq (1+\delta)\mu\right] \leq e^{-\delta^2\mu/3} \ .$$

3.2 Quantum preliminaries and notation

Quantum information. A (pure) *quantum state* is a vector $|\psi\rangle$ in a complex Hilbert space $\mathcal{H}$ with $\||\psi\rangle\| = 1$; in this work, $\mathcal{H}$ is always finite-dimensional. We denote by $\mathbf{S}(\mathcal{H})$ the space of Hermitian operators on $\mathcal{H}$. A *density matrix* is a Hermitian operator $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H})$ with $\mathrm{Tr}(\boldsymbol{\rho}) = 1$. A density matrix represents a probabilistic mixture of pure states (a mixed state); the density matrix corresponding to the pure state $|\psi\rangle$ is $|\psi\rangle\langle\psi|$. Typically we divide a Hilbert space into *registers*, e.g. $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$. We sometimes write, e.g., $\boldsymbol{\rho}^{\mathcal{H}_1}$ to specify that $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H}_1)$.

A unitary operation is represented by a complex matrix U such that $UU^\dagger = \mathbf{I}$. The operation U transforms the pure state $|\psi\rangle$ to the pure state $U|\psi\rangle$, and the density matrix $\boldsymbol{\rho}$ to the density matrix $U\boldsymbol{\rho}U^\dagger$.

A *projector* Π is a Hermitian operator ($\Pi^\dagger = \Pi$) such that $\Pi^2 = \Pi$. A *projective measurement* is a collection of projectors $\mathsf{P} = (\Pi_i)_{i \in S}$ such that $\sum_{i \in S} \Pi_i = \mathbf{I}$. This implies that $\Pi_i \Pi_j = 0$ for distinct i and j in S. The application of a projective measurement to a pure state $|\psi\rangle$ yields outcome $i \in S$ with probability $p_i = \|\Pi_i|\psi\rangle\|^2$; in this case the post-measurement state is $|\psi_i\rangle = \Pi_i|\psi\rangle / \sqrt{p_i}$. We will sometimes refer to the post-measurement state $\Pi_i|\psi\rangle/\sqrt{p_i}$ as the result of applying $\mathsf{P} = (\Pi_i)_{i\in S}$ to $|\psi\rangle$ and *post-selecting* (i.e., conditioning) on outcome i. A state $|\psi\rangle$ is an *eigenstate* of P if it is an eigenstate of every Π_i.

A two-outcome projective measurement is called a *binary projective measurement*, and is written as $\mathsf{P} = (\Pi, \mathbf{I} - \Pi)$, where Π is associated with the outcome 1, and $\mathbf{I} - \Pi$ with the outcome 0.

General (non-unitary) evolution of a quantum state can be represented via a *completely-positive trace-preserving (CPTP)* map $T\colon \mathbf{S}(\mathcal{H}) \to \mathbf{S}(\mathcal{H}')$. We omit the precise definition of these maps in this work; we will only use the facts that they are trace-preserving (for every $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H})$ it holds that $\mathrm{Tr}(T(\boldsymbol{\rho})) = \mathrm{Tr}(\boldsymbol{\rho})$) and linear.

For every CPTP map $T\colon \mathbf{S}(\mathcal{H}) \to \mathbf{S}(\mathcal{H})$ there exists a *unitary dilation* U that operates on an expanded Hilbert space $\mathcal{H} \otimes \mathcal{K}$, so that $T(\boldsymbol{\rho}) = \mathrm{Tr}_{\mathcal{K}}(U(\rho \otimes |0\rangle\langle 0|^{\mathcal{K}})U^\dagger)$. This is not necessarily unique; however, if T is described as a circuit then there is a dilation U_T represented by a circuit of size $O(|T|)$.

For Hilbert spaces $\mathcal{A}, \mathcal{B}$ the *partial trace* over $\mathcal{B}$ is the unique CPTP map $\mathrm{Tr}_{\mathcal{B}}\colon \mathbf{S}(\mathcal{A}\otimes\mathcal{B}) \to \mathbf{S}(\mathcal{A})$ such that $\mathrm{Tr}_{\mathcal{B}}(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B) = \mathrm{Tr}(\boldsymbol{\rho}_B)\boldsymbol{\rho}_A$ for every $\boldsymbol{\rho}_A \in \mathbf{S}(\mathcal{A})$ and $\boldsymbol{\rho}_B \in \mathbf{S}(\mathcal{B})$.

A *general measurement* is a CPTP map $\mathsf{M}\colon \mathbf{S}(\mathcal{H}) \to \mathbf{S}(\mathcal{H} \otimes \mathcal{O})$, where $\mathcal{O}$ is an ancilla register holding a classical outcome. Specifically, given measurement operators $\{M_i\}_{i=1}^N$ such that $\sum_{i=1}^N M_i M_i^\dagger = \mathbf{I}$ and a basis $\{|i\rangle\}_{i=1}^N$ for $\mathcal{O}$, $\mathsf{M}(\boldsymbol{\rho}) \coloneqq \sum_{i=1}^N (M_i \boldsymbol{\rho} M_i^\dagger \otimes |i\rangle\langle i|^{\mathcal{O}})$. We will sometimes implicitly discard the outcome register. A projective measurement is simply a general measurement where the M_i are projectors. A measurement induces a probability distribution over its outcomes given by $\Pr[i] = \mathrm{Tr}\Big(|i\rangle\langle i|^{\mathcal{O}}\, \mathsf{M}(\boldsymbol{\rho})\Big)$; we denote sampling from this distribution by $i \leftarrow \mathsf{M}(\boldsymbol{\rho})$.

The *trace distance* between states $\boldsymbol{\rho}, \boldsymbol{\sigma}$, denoted $d(\boldsymbol{\rho}, \boldsymbol{\sigma})$, is defined as $\frac{1}{2}\mathrm{Tr}\Big(\sqrt{(\boldsymbol{\rho}-\boldsymbol{\sigma})^2}\Big)$. The trace distance is contractive under CPTP maps, i.e. for any CPTP map T, $d(T(\boldsymbol{\rho}), T(\boldsymbol{\sigma})) \le d(\boldsymbol{\rho}, \boldsymbol{\sigma})$. It follows that for any measurement M, the statistical distance between the distributions $\mathsf{M}(\boldsymbol{\rho})$ and $\mathsf{M}(\boldsymbol{\sigma})$ is bounded by $d(\boldsymbol{\rho}, \boldsymbol{\sigma})$. We have the following *gentle measurement lemma*, which bounds how much a state is disturbed by applying a measurement whose outcome is almost certain.

Lemma 3.3 (Gentle Measurement [Win99])**.** *Let $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H})$ and $\mathsf{P} = (\Pi, \mathbf{I} - \Pi)$ be a binary projective measurement on $\mathcal{H}$ such that $\mathrm{Tr}(\Pi\boldsymbol{\rho}) \ge 1 - \delta$. Let $\boldsymbol{\rho}'$ be the state after applying P to $\boldsymbol{\rho}$ and post-selecting on obtaining outcome 1. Then*

$$d(\boldsymbol{\rho}, \boldsymbol{\rho}') < 2\sqrt{\delta}.$$

Quantum algorithms. In this work, a *quantum adversary* is a family of quantum circuits $\{\mathsf{Adv}_\lambda\}_{\lambda\in\mathbb{N}}$ represented classically using some standard universal gate set. A quantum adversary is *polynomial-size* if there exists a polynomial p and $\lambda_0 \in \mathbb{N}$ such that for all $\lambda > \lambda_0$ it holds that $|\mathsf{Adv}_\lambda| \le p(\lambda)$ (i.e., quantum adversaries have classical non-uniform advice).

In this work we refer to the *expected running time* of quantum algorithms. This means that there is a classical control algorithm that applies quantum circuits of a fixed size and decides whether to terminate based on the classical outputs of those circuits. The expected running time is then the expected number of unit operations, classical or quantum, applied during this execution.

Black-box access. A circuit C with black-box access to a unitary U, denoted C^U, is a standard quantum circuit with special gates that act as U and $U^\dagger$. We also use C^T to denote black-box access to a map T, which we interpret as C^{U_T} for a unitary dilation U_T of T; all of our results will be independent of the choice of dilation. This allows, for example, the "partial application" of a projective measurement, and the implementation of a general measurement via a projective measurement on a larger space.

3.3 Jordan's lemma

We state Jordan's lemma and, for completeness, provide a proof that roughly follows [Reg06].

Lemma 3.4 ([Jor75])**.** *For any two Hermitian projectors Π_{A} and Π_{B} on a Hilbert space $\mathcal{H}$, there exists an orthogonal decomposition of $\mathcal{H} = \bigoplus_j \mathcal{S}_j$ into one-dimensional and two-dimensional subspaces $\{\mathcal{S}_j\}_j$ (the* Jordan subspaces*), where each $\mathcal{S}_j$ is invariant under both Π_{A} and Π_{B}. Moreover:*

- *in each one-dimensional space,* Π_{A} *and* Π_{B} *act as identity or rank-zero projectors; and*
- *in each two-dimensional subspace* $\mathcal{S}_j$*,* Π_{A} *and* Π_{B} *are rank-one projectors. In particular, there exist distinct orthogonal bases* $\{|v^{\mathsf{A}}_{j,1}\rangle, |v^{\mathsf{A}}_{j,0}\rangle\}$ *and* $\{|v^{\mathsf{B}}_{j,1}\rangle, |v^{\mathsf{B}}_{j,0}\rangle\}$ *for* $\mathcal{S}_j$ *such that* Π_{A} *projects onto* $|v^{\mathsf{A}}_{j,1}\rangle$ *and* Π_{B} *projects onto* $|v^{\mathsf{B}}_{j,1}\rangle$*.*

Proof. Since Π_{A} and Π_{B} are both Hermitian, their sum $\Pi_{\mathsf{A}} + \Pi_{\mathsf{B}}$ is also Hermitian. By the spectral theorem for Hermitian matrices, it follows that the eigenvectors of $\Pi_{\mathsf{A}} + \Pi_{\mathsf{B}}$ span $\mathcal{H}$. Let $|\psi\rangle$ be an eigenvector with eigenvalue p (i.e., $\Pi_{\mathsf{A}}|\psi\rangle + \Pi_{\mathsf{B}}|\psi\rangle = p|\psi\rangle$). There are two cases to consider.

If $\Pi_{\mathsf{A}}|\psi\rangle$ lies in $\mathrm{span}(|\psi\rangle)$, then $\Pi_{\mathsf{B}}|\psi\rangle$ must also be in $\mathrm{span}(|\psi\rangle)$, so $\mathrm{span}(|\psi\rangle)$ is a one-dimensional subspace invariant under both Π_{A} and Π_{B}. Since Π_{A} and Π_{B} are projectors, their eigenvalues are 0 or 1, so in $\mathrm{span}(|\psi\rangle)$ they act as identity or rank-zero projectors.

If $\Pi_{\mathsf{A}}|\psi\rangle$ does not lie in $\mathrm{span}(|\psi\rangle)$, then $\mathrm{span}(|\psi\rangle, \Pi_{\mathsf{A}}|\psi\rangle)$ is a two-dimensional subspace. This subspace is invariant under Π_{A}, which acts as a projector onto $|v^{\mathsf{A}}_{j,1}\rangle := \Pi_{\mathsf{A}}|\psi\rangle$. Moreover, this subspace can be written as $\mathrm{span}(|\psi\rangle, \Pi_{\mathsf{B}}|\psi\rangle)$, and by an identical argument, Π_{B} projects this subspace onto $|v^{\mathsf{B}}_{j,1}\rangle := \Pi_{\mathsf{B}}|\psi\rangle$.

By setting $|v^{\mathsf{A}}_{j,0}\rangle := \mathcal{S}_j \cap \ker(\Pi_{\mathsf{A}})$ (i.e. the state in $\mathcal{S}_j$ orthogonal to $|v^{\mathsf{A}}_{j,1}\rangle$)) and $|v^{\mathsf{B}}_{j,0}\rangle := \mathcal{S}_j \cap \ker(\Pi_{\mathsf{B}})$, we obtain two different orthogonal bases $\{|v^{\mathsf{A}}_{j,1}\rangle, |v^{\mathsf{A}}_{j,0}\rangle\}$ and $\{|v^{\mathsf{B}}_{j,1}\rangle, |v^{\mathsf{B}}_{j,0}\rangle\}$ for $\mathcal{S}_j$ where Π_{A} projects onto $|v^{\mathsf{A}}_{j,1}\rangle$ and Π_{B} projects onto $|v^{\mathsf{B}}_{j,1}\rangle$. □

3.4 Interactive arguments

For interactive classical algorithm V and interactive (potentially) quantum circuit A, we denote by $\langle A(|\psi\rangle), V\rangle$ the random variable corresponding to the output of V when interacting with $A(|\psi\rangle)$; note that since V is classical, the communication in this interaction is also classical. For a general formal treatment of interactive quantum circuits, see [VW16].

Definition 3.5. A (post-quantum) *interactive argument* for a relation $\mathfrak{R}$ with soundness s is a pair of interactive classical polynomial-time algorithms $\mathsf{ARG} = (P, V)$ such that the following holds.

- **Completeness.** For every $\lambda \in \mathbb{N}$ and $(x, w) \in \mathfrak{R}$, $\Pr\Big[\langle P(1^\lambda, x, w), V(1^\lambda, x)\rangle = 1\Big] = 1$.
- **Soundness.** For every $\lambda \in \mathbb{N}$, $x \notin \mathcal{L}(\mathfrak{R})$, and polynomial-size interactive quantum circuit $\tilde{P}$,

$$\Pr\Big[\langle \tilde{P}, V(1^\lambda, x)\rangle = 1\Big] \leq s(\lambda) \ .$$

We say that ARG is **succinct** if the total amount of communication between P and V is at most $c(\lambda, \log|x|)$ for some fixed polynomial c.

In this work a *round* is a back-and-forth interaction consisting of a verifier message followed by a prover message.

We also consider interactive arguments that satisfy the stronger property of *knowledge soundness*. Below we write $E^{\tilde{P}}$ for an extractor with "black-box" access to $\tilde{P}$; we will give this a precise meaning shortly. Our definition loosely follows that of [Unr12].

Definition 3.6. $\mathsf{ARG} = (P, V)$ has *knowledge soundness* with knowledge error κ if there exists an expected polynomial time quantum extractor E such that for every polynomial-size interactive quantum circuit $\tilde{P}$, quantum state $|\psi\rangle$, $\lambda \in \mathbb{N}$, instance x, and parameter $\varepsilon(\lambda) \leq \Pr\Big[\langle \tilde{P}(x, |\psi\rangle), V(1^\lambda, x)\rangle = 1\Big]$ the following holds:

$$\Pr\left[(x, w) \in \mathfrak{R} \mid w \leftarrow E^{\tilde{P}(x;|\psi\rangle)}(1^\lambda, x, 1^{1/\varepsilon})\right] = \Omega(\varepsilon(\lambda) - \kappa) \ .$$

We describe the differences between our definition and the definition of quantum proofs of knowledge given in [Unr12].

- Our definition asks that the extractor succeed with probability *linear* in $(\varepsilon(\lambda) - \kappa)$, whereas Unruh's definition only requires the extractor's success probability be $(\varepsilon(\lambda) - \kappa)^d / p(\lambda)$ for a constant $d \in \mathbb{N}$ and polynomial p.
- Our definition is incomparable to Unruh's definition when $|\psi\rangle$ is a general quantum state, since we require that the extractor be given as input a lower bound ε on the success probability of the adversary. This arises due to a technical requirement in our security proof.
- When $|\psi\rangle$ is a computational basis state (or any other efficiently-constructible state), our definition is stronger than Unruh's definition since in this case the extractor can compute for itself a lower bound on the success probability of the adversary by simply running the adversary many times (independently, from the *beginning* of the protocol).

To define black-box access to $\tilde{P}$, we will need to consider in more detail how an interactive quantum circuit is specified.

Definition 3.7 (Interactive quantum circuits)**.** A m-round interactive quantum circuit A is a sequence of unitary quantum circuits $(U^{(1)}, \ldots, U^{(m)})$ where $U^{(i)}$ operates on registers $(\mathcal{I}, \mathcal{R}_i, \mathcal{Z}_i)$.

The *size* of an interactive quantum circuit is the sum of the sizes of the circuits implementing $U^{(1)}, \ldots, U^{(m)}$.

Let $\tilde{P} \coloneqq (U^{(1)}, \ldots, U^{(m)})$; then $E^{\tilde{P}}$ is a quantum circuit with special gates corresponding to $U^{(i)}(r)$ and $(U^{(i)}(r))^\dagger$ for $i \in [m]$.

The requirement that the $U^{(i)}$ be unitary is without loss of generality, in the sense that any quantum circuit not of this form can be "purified" into a circuit of this form which is only a constant factor larger with the same observable behavior. Using this formulation, we can sample the random variable $\langle \tilde{P}, V \rangle$ equivalently as:

1. Initialize the register $\mathcal{I}$ to $|\psi\rangle$, and $\tau \coloneqq ()$.
2. For $i = 1, \ldots, m$,
   (a) Sample $r_i \leftarrow R_i$. Initialize the $\mathcal{R}_i$ register to $|r_i\rangle$.
   (b) Apply unitary $U^{(i)}$ to $(\mathcal{I}, \mathcal{R}_i, \mathcal{Z}_i)$.
   (c) Measure $\mathcal{Z}_i$ in the computational basis to obtain response z_i. Append (r_i, z_i) to τ.
3. Return the output of $V(\tau)$.

In particular, the interaction is *public coin*. Note again that we restrict the operation of A in each round to be unitary except for the measurement of $\mathcal{Z}_i$ in the computational basis.

3.5 Collapsing hash functions

Let $\mathcal{H} = \{H_\lambda\}_{\lambda \in \mathbb{N}}$ be such that each H_λ is a distribution over functions $h \colon \{0,1\}^{n(\lambda)} \to \{0,1\}^{\ell(\lambda)}$.

Definition 3.8. $\mathcal{H}$ is *post-quantum collision resistant* if for every polynomial-size quantum adversary Adv,

$$\Pr\left[\begin{array}{c|c} x \neq x' \wedge & h \leftarrow H_\lambda \\ h(x) = h(x') & (x, x') \leftarrow \mathsf{Adv}(h) \end{array}\right] = \mathrm{negl}(\lambda) \ .$$

Definition 3.9. $\mathcal{H}$ is *collapsing* [Unr16b] if for every security parameter λ and polynomial-size quantum adversary Adv,

$$\Big|\Pr[\mathsf{HCollapseExp}(0,\lambda,\mathsf{Adv})=1]-\Pr[\mathsf{HCollapseExp}(1,\lambda,\mathsf{Adv})=1]\Big|\leq \mathrm{negl}(\lambda)\ .$$

For $b\in\{0,1\}$ the experiment $\mathsf{HCollapseExp}(b,\lambda,\mathsf{Adv})$ is defined as follows:

1. The challenger samples $h\leftarrow H_\lambda$ and sends h to Adv.
2. Adv replies with a (classical) binary string $y\in\{0,1\}^{\ell(\lambda)}$ and a $n(\lambda)$-qubit quantum state on registers $\mathcal{X}$. (The requirement that y is classical can be enforced by having the challenger immediately measure these registers upon receiving them.)
3. The challenger computes h in superposition on the $n(\lambda)$-qubit quantum state, and measures the bit indicating whether the output of h equals y. If h does not equal y, the challenger aborts and outputs $\perp$.
4. If $b=0$, the challenger does nothing. If $b=1$, the challenger measures the $n(\lambda)$-qubit state in the standard basis.
5. The challenger returns contents of the registers $\mathcal{X}$ to Adv.
6. Adv outputs a bit b', which is the output of the experiment.

Claim 3.10 ([Unr16b])**.** *If $\mathcal{H}$ is collapsing then $\mathcal{H}$ is collision resistant.*

Proof. A proof can be found in [Unr16b, Lemma 25], but for convenience we include a proof here.

Let Adv be an adversary that breaks collision resistance of $\mathcal{H}$ with probability at least $\varepsilon(\lambda)$. We construct an adversary Adv' that breaks collapsing of $\mathcal{H}$ with probability at least $\varepsilon(\lambda)/2$.

The adversary Adv' works as follows. First, given as input $h\leftarrow H_\lambda$, Adv' computes $(x,x')\leftarrow\mathsf{Adv}(h)$. If (x,x') is not a valid collision (they are equal or they map to different outputs under h) then Adv' sends to the challenger an arbitrary classical bitstring y and an arbitrary quantum state on register $\mathcal{X}$, and then outputs 0 at the conclusion of the experiment. If (x,x') is a valid collision (they are distinct and they map to the same ouput under h), then Adv' sends $y:=h(x)$ and the quantum state $|\psi\rangle:=\frac{1}{\sqrt{2}}(|x\rangle+|x'\rangle)$ on register $\mathcal{X}$; when the challenger returns the contents of $\mathcal{X}$, Adv' applies the binary projective measurement $\mathsf{P}=(|\psi\rangle\langle\psi|,\mathbf{I}-|\psi\rangle\langle\psi|)$, and outputs the measurement outcome b.

In $\mathsf{HCollapseExp}(0,\lambda,\mathsf{Adv}')$, the adversary Adv' outputs 1 with probability at least $\varepsilon(\lambda)$, since as long as Adv outputs a valid collision (x,x'), the measurement P is applied to $\frac{1}{\sqrt{2}}(|x\rangle+|x'\rangle)$ and must return 1. In $\mathsf{HCollapseExp}(1,\lambda,\mathsf{Adv}')$, the adversary Adv' outputs 1 with probability at most $\varepsilon(\lambda)/2$, since as long as Adv outputs a valid collision (x,x'), the measurement P is applied to either $|x\rangle$ or $|x'\rangle$, and thus returns 1 with probability at most $1/2$. The overall difference in the two probabilities is $\varepsilon(\lambda)/2$. □

3.6 Collapsing protocols

Definition 3.11 ([Unr16b, LZ19, DFMS19])**.** We say that a protocol is collapsing if for every polynomial-size interactive quantum adversary $\tilde{P}$ and polynomial-size quantum distinguisher Adv,

$$\Big|\Pr\Big[\mathsf{CollapseExp}(0,\tilde{P},\mathsf{Adv})=1\Big]-\Pr\Big[\mathsf{CollapseExp}(1,\tilde{P},\mathsf{Adv})=1\Big]\Big|\leq\mathrm{negl}(\lambda)\ .$$

For $b\in\{0,1\}$, the experiment $\mathsf{CollapseExp}(b,\tilde{P},\mathsf{Adv})$ is defined as follows:

1. The challenger simulates $\langle \tilde{P}, V \rangle$, stopping just before the measurement of $\mathcal{Z}_m$. Let $\tau' = (r_1, z_1, \ldots, r_{m-1}, z_{m-1}, r_m)$ be the transcript up to this point (i.e., excluding the final prover message).
2. The challenger applies a unitary U that computes the bit $V(\tau', \mathcal{Z}_m)$ into a fresh ancilla, measures the ancilla, and applies $U^\dagger$. If the measurement outcome is 0, the experiment aborts.
3. If $b = 0$, the challenger does nothing. If $b = 1$, the challenger measures the $\mathcal{Z}_m$ register in the computational basis and discards the result.
4. The challenger sends all registers to Adv. Adv outputs a bit b', which is the output of the experiment.

4 Efficient quantum strategies for repeated games

We consider a classical single-player game $\mathcal{G}$ played with quantum strategies. This section makes use of the notion of quantum interaction and interactive quantum algorithms; for details on how to model this formally, see [VW16].

Definition 4.1. A game $\mathcal{G} = (R, Z, f)$ consists of a question set R, answer set Z, and win predicate $f\colon R \times Z \to \{0,1\}$. An (efficient) *quantum* strategy for $\mathcal{G}$ is an interactive quantum algorithm S with initial state $\boldsymbol{\rho}$.

The *value* of a strategy $(S, \boldsymbol{\rho})$, denoted $\omega_{\mathcal{G}}(S, \boldsymbol{\rho})$, is the probability that a player using strategy $(S, \boldsymbol{\rho})$ in the following game causes the referee to output 1: the referee sends the player a question $r \leftarrow R$, and the player answers with (classical) $z \in Z$; the referee outputs $f(r, z)$.

We now define a *quantum experiment* in which the player's answer can be an arbitrary quantum state on $\mathcal{Z}$, and the referee determines whether the player wins by computing $f(r, \mathcal{Z})$ in superposition and measuring the output; it then uncomputes f and returns $\mathcal{Z}$ to the player. The key difference between the classical and quantum experiments is that the only measurement performed in the quantum experiment is on the output of f, whereas a quantum player in a classical interaction must measure to send a classical z. While this does not affect the value of a game when played once, it is crucial when the game is repeated *sequentially*.

In more detail, our quantum experiment consists of the following quantum interaction:

1. The referee samples a question $r \leftarrow R$ and sends it to the player.
2. The player responds with a quantum state on register $\mathcal{Z}$.
3. The referee computes $f(r, \mathcal{Z})$ in superposition, measures the result to obtain an outcome $b \in \{0,1\}$, and uncomputes f. The referee then returns $\mathcal{Z}$ to the player.

It is easily verified that the probability a player following strategy $(S, \boldsymbol{\rho})$ wins in the above experiment is $\omega_{\mathcal{G}}(S, \boldsymbol{\rho})$, as in the classical experiment. Without loss of generality, we can assume that the strategy S is implemented by a unitary U_S.

We now consider the n-fold *sequential repetition* of the quantum experiment. Formally, the interaction consists of n sequential rounds, where in the ith round:

1. The referee samples a question $r_i \leftarrow R$ and sends it to the player.
2. The player responds with a quantum state on $\mathcal{Z}$.
3. The referee computes $f(r_i, \mathcal{Z})$ in superposition, measures the result to obtain an outcome $b_i \in \{0,1\}$, and uncomputes f. The referee then returns $\mathcal{Z}$ to the player, along with b_i.

Definition 4.2 (Value of a strategy in a repeated game)**.** The *value* of a strategy $(S, \boldsymbol{\rho})$ in the above experiment is denoted $\omega^n_{\mathcal{G}}(S, \boldsymbol{\rho})$, and is equal to $\mathbb{E}[\sum_{i=1}^n b_i]$, the *expected number of wins* across all trials. Note that $\omega^1_{\mathcal{G}}(S, \boldsymbol{\rho}) = \omega_{\mathcal{G}}(S, \boldsymbol{\rho})$.

When $\boldsymbol{\rho}$ is a classical state, the n-fold repetition S^n of any strategy S trivially achieves $\omega^n_{\mathcal{G}}(S^n, \boldsymbol{\rho}) = n \cdot \omega_{\mathcal{G}}(S, \boldsymbol{\rho})$. For quantum $\boldsymbol{\rho}$, this may not be true, since the state is in general disturbed by the referee's measurement. In this section we show that, given any *quantum* strategy $(S, \boldsymbol{\rho})$ for the one-round experiment, there is an efficient quantum algorithm S' that makes black-box use of U_S (and $U_S^\dagger$) such that $\omega^n_{\mathcal{G}}(S', \boldsymbol{\rho}) \approx n \cdot \omega_{\mathcal{G}}(S, \boldsymbol{\rho})$.

Theorem 4.3. *For any single-player quantum game $\mathcal{G} = (R, Z, f)$ with classically efficient predicate f, $n \in \mathbb{N}, \eta_0 \in [0,1]$, there is a quantum oracle algorithm $A_{\mathcal{G},n,\eta_0}$ such that for all $(S, \boldsymbol{\rho})$,*

$$\omega_{\mathcal{G}}^n(A^S_{\mathcal{G},n,\eta_0}, \boldsymbol{\rho}) \geq n \cdot (\omega_{\mathcal{G}}(S, \boldsymbol{\rho}) - \eta_0)$$

and A runs in expected time $\tilde{O}(|f| \cdot n/\eta_0)$ and makes an expected $\tilde{O}(n/\eta_0)$ queries to $U_S, U_S^\dagger$.

We prove the theorem using two key subroutines, ValEst and $\mathsf{ValRepair}$, which do the following:

- ValEst^S applied to $\boldsymbol{\rho}$ is an *approximate measurement* of $\omega_{\mathcal{G}}(S, \boldsymbol{\rho})$. That is, it produces an outcome p where $\mathbb{E}[p] = \omega_{\mathcal{G}}(S, \boldsymbol{\rho})$, and conditioned on obtaining outcome p the post-measurement state $\boldsymbol{\rho}'$ satisfies $\omega_{\mathcal{G}}(S, \boldsymbol{\rho}') \approx p$.
- $\mathsf{ValRepair}_p^S$ is a procedure that *repairs* a state that has been perturbed by the referee's measurement. In more detail, if $\boldsymbol{\rho}$ is the state of the system after applying ValEst^S and obtaining outcome p, and playing a one-round experiment with strategy $(S, \boldsymbol{\rho})$ results in leftover state $\boldsymbol{\rho}'$, then applying $\mathsf{ValRepair}_p^S$ to $\boldsymbol{\rho}'$ outputs a *repaired* state $\boldsymbol{\rho}^*$ in the sense that $\omega_{\mathcal{G}}(S, \boldsymbol{\rho}^*) \approx p$.

We remark that our implementations of ValEst^S and $\mathsf{ValRepair}_p^S$ make black-box use of $U_S, U_S^\dagger$.

Given a strategy $(S, \boldsymbol{\rho})$ for the one-round experiment, our n-time strategy is as follows.

Repeat for $i \in [n]$:

(a) Apply $p_i \leftarrow \mathsf{ValEst}^S$.
(b) Receive $r_i \in R$; run $S(r_i)$ coherently to compute $\mathcal{Z}$ and send it to the referee.
(c) Receive $\mathcal{Z}$ and measurement result $b_i \in \{0,1\}$ from the referee.
(d) Apply $\mathsf{ValRepair}_{p_i}^S$.

The guarantee of ValEst implies that $\mathbb{E}[p_1] = \omega_{\mathcal{G}}(S, \boldsymbol{\rho})$, and that $\Pr[b_i = 1] \approx \mathbb{E}[p_i]$ for all i. The guarantee of $\mathsf{ValRepair}$ implies that $p_1 \approx p_2 \approx \cdots \approx p_n$ with high probability. Together these imply Theorem 4.3, by linearity of expectation.

Organization. In Section 4.1 we present general technical lemmas that are useful for analysing algorithms which consist of alternating applications of two binary projective measurements; both ValEst and $\mathsf{ValRepair}$ are of this type. In Section 4.2 we describe and analyze our ValEst procedure, which is a variant of procedures from [MW05, Zha20]. In Section 4.3 we describe and analyze $\mathsf{ValRepair}$. Finally, in Section 4.4 we prove Theorem 4.3.

4.1 Jordan subspaces and alternating measurements

We provide general tools for analysing *alternating projection algorithms*, which were introduced by Marriott and Watrous [MW05] for witness-preserving amplification of QMA. In more detail, given two binary-outcome projective measurements $\mathsf{A} = (\Pi_{\mathsf{A}}, \mathbf{I} - \Pi_{\mathsf{A}})$ and $\mathsf{B} = (\Pi_{\mathsf{B}}, \mathbf{I} - \Pi_{\mathsf{B}})$ on a Hilbert space $\mathcal{H}$, an alternating projection algorithm applies the measurements in alternating fashion $(\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B}, \ldots)$ until a stopping condition is met (e.g., a certain number of measurements have been performed or some outcome has been observed). We can describe the distribution of measurement outcomes using Jordan's lemma (Lemma 3.4).

Jordan decomposition. Applying Jordan's lemma (Lemma 3.4) to $(\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}})$ induces an orthogonal decomposition $\mathcal{H} = \bigoplus_j \mathcal{S}_j$ into one- and two-dimensional *Jordan subspaces* $\mathcal{S}_j$.

Within each two-dimensional Jordan subspace $\mathcal{S}_j$, we define four states $|v_{j,1}^{\mathsf{A}}\rangle, |v_{j,0}^{\mathsf{A}}\rangle, |v_{j,1}^{\mathsf{B}}\rangle, |v_{j,0}^{\mathsf{B}}\rangle$:

- $|v^{\mathsf{A}}_{j,1}\rangle$ is a state in $\mathcal{S}_j \cap \text{image}(\Pi_{\mathsf{A}})$.
- $|v^{\mathsf{B}}_{j,1}\rangle$ is a state in $\mathcal{S}_j \cap \text{image}(\Pi_{\mathsf{B}})$.
- $|v^{\mathsf{A}}_{j,0}\rangle$ is a state in $\mathcal{S}_j \cap \ker(\Pi_{\mathsf{A}})$ (orthogonal to $|v^{\mathsf{A}}_{j,1}\rangle$).
- $|v^{\mathsf{B}}_{j,0}\rangle$ is a state in $\mathcal{S}_j \cap \ker(\Pi_{\mathsf{B}})$ (orthogonal to $|v^{\mathsf{B}}_{j,1}\rangle$).

These states are unique up to phase. Let

$$p_j \coloneqq \left\| \langle v^{\mathsf{A}}_{j,1} | v^{\mathsf{B}}_{j,1} \rangle \right\|^2 = \left\| \langle v^{\mathsf{A}}_{j,0} | v^{\mathsf{B}}_{j,0} \rangle \right\|^2 \ .$$

We adopt the convention that the phases of these states are chosen to satisfy

$$|v^{\mathsf{A}}_{j,1}\rangle = \sqrt{p_j}\,|v^{\mathsf{B}}_{j,1}\rangle + \sqrt{1-p_j}\,|v^{\mathsf{B}}_{j,0}\rangle \quad \text{and} \quad |v^{\mathsf{B}}_{j,1}\rangle = \sqrt{p_j}\,|v^{\mathsf{A}}_{j,1}\rangle + \sqrt{1-p_j}\,|v^{\mathsf{A}}_{j,0}\rangle \ . \tag{1}$$

Notice that if $|\psi\rangle$ is the post-measurement state after A has returned 1, then $|\psi\rangle = \sum_j \alpha_j |v^{\mathsf{A}}_{j,1}\rangle$ for some choice of amplitudes $\{\alpha_j\}_j$. Likewise, if $|\psi\rangle$ is the post-measurement state after B has returned 1, then $|\psi\rangle = \sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle$ for some choice of amplitudes $\{\alpha_j\}_j$.

We can view each one-dimensional subspace $\mathcal{S}_j$ as a *degenerate* two-dimensional subspace. If Π_{A} acts as the identity on $\mathcal{S}_j$ then we label the vector spanning the subspace $|v^{\mathsf{A}}_{j,1}\rangle$; if Π_{A} is the zero projection on $\mathcal{S}_j$ then we label the vector $|v^{\mathsf{A}}_{j,0}\rangle$. We use a similar convention for Π_{B} (so the vector spanning a one-dimensional subspace has two labels). We set $p_j \coloneqq 1$ if both Π_{A} and Π_{B} act as the identity or both act as zero, and $p_j \coloneqq 0$ otherwise. One can verify that the discussion above for two-dimensional subspaces holds for one-dimensional subspaces under this convention.

Distribution of measurement outcomes. Consider the following (classical) probability distribution $\mathsf{MWDist}(T, p)$ (for "Marriott–Watrous distribution"), parameterized by a probability $p \in [0,1]$ and positive integer T.

$\mathsf{MWDist}(T, p)$:

1. For each $i \in [T]$, set $a_i := 1$ with probability p and $a_i := 0$ otherwise.
2. Let $b_0 \coloneqq 1$. For $i \in [T]$, define $b_i := b_{i-1} \oplus a_i$.
3. Output $b_1, b_2, \ldots, b_T$.

The following two lemmas characterize the distribution of measurement outcomes of an alternating measurement procedure. The analysis closely follows that of [MW05, Reg06].

Lemma 4.4. *The measurement outcomes that result from applying T alternating measurements $\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B} \ldots$ to $|v^{\mathsf{B}}_{j,1}\rangle$ are distributed according to $\mathsf{MWDist}(T, p_j)$.*

Proof. This is a consequence of two symmetric claims that follow directly from Eq. (1).

- If A is applied to $|v^{\mathsf{B}}_{j,b}\rangle$, then with probability p_j the outcome is b and the post-measurement state is $|v^{\mathsf{A}}_{j,b}\rangle$, and with probability $1-p_j$ the outcome is $1-b$ and the post measurement state is $|v^{\mathsf{A}}_{j,1-b}\rangle$.

- If B is applied to $|v^{\mathsf{A}}_{j,b}\rangle$, then with probability p_j the outcome is b and the post-measurement state is $|v^{\mathsf{B}}_{j,b}\rangle$, and with probability $1-p_j$ the outcome is $1-b$ and the post measurement state is $|v^{\mathsf{B}}_{j,1-b}\rangle$.

It is convenient to think of the initial state $|v^{\mathsf{B}}_{j,1}\rangle$ as the post-measurement state after B returns 1. Letting $b_0 \coloneqq 1$, for any $i \in [T]$ the i-th measurement outcome b_i is equal to b_{i-1} with probability p_j and equal to $1 - b_{i-1}$ with probability $1 - p_j$, giving the distribution $\mathsf{MWDist}(T, p_j)$. □

We can generalize Lemma 4.4 to characterize the measurement outcomes when we begin with any state in image(Π_{B}), which must be of the form $\sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle$.

Lemma 4.5. *The measurement outcomes that result from applying T alternating measurements $\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B}, \ldots$ to the state $\sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle$ have the following distribution:*
1. *sample p_j with probability $|\alpha_j|^2$;*
2. *output $\mathsf{MWDist}(T, p_j)$.*

Proof. Consider the *Jordan subspace measurement* $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}] \coloneqq (\Pi^{\mathrm{Jor}}_j)_j$ on $\mathcal{H}$, where

$$\Pi^{\mathrm{Jor}}_j \coloneqq |v^{\mathsf{A}}_{j,1}\rangle\langle v^{\mathsf{A}}_{j,1}| + |v^{\mathsf{A}}_{j,0}\rangle\langle v^{\mathsf{A}}_{j,0}| = |v^{\mathsf{B}}_{j,1}\rangle\langle v^{\mathsf{B}}_{j,1}| + |v^{\mathsf{B}}_{j,0}\rangle\langle v^{\mathsf{B}}_{j,0}| \ .$$

In words, $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ is the projective measurement onto the Jordan subspaces $\{S_j\}_j$ that outputs a Jordan subspace label j.

Suppose that we perform the measurement $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ on $\sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle$, and subsequently perform T alternating measurements $\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B}, \ldots$. The outcome of $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ is j with probability $|\alpha_j|^2$, and the subsequent alternating measurement outcomes are distributed according to $\mathsf{MWDist}(T, p_j)$ by Lemma 4.4. It remains to prove that the distribution of measurement outcomes is unchanged even if we skip the $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ measurement.

This is because $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ *commutes* with both A and B. To see that $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ commutes with A, observe that the corresponding measurement operators are diagonal in the basis $\{|v^{\mathsf{A}}_{j,b}\rangle\}_{j,b}$, since $\Pi_{\mathsf{A}} = \sum_j |v^{\mathsf{A}}_{j,1}\rangle\langle v^{\mathsf{A}}_{j,1}|$ by Jordan's lemma and $\Pi^{\mathrm{Jor}}_j = |v^{\mathsf{A}}_{j,0}\rangle\langle v^{\mathsf{A}}_{j,0}| + |v^{\mathsf{A}}_{j,1}\rangle\langle v^{\mathsf{A}}_{j,1}|$ for all j by definition. $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ commutes with B by an identical argument for the basis $\{|v^{\mathsf{B}}_{j,b}\rangle\}_{j,b}$.

As a consequence, we can commute $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ to occur *after* the T alternating measurements $\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B}, \ldots$, at which point $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$ has no effect on the measurement outcomes. □

Almost projective measurements. We state a property of general measurements due to [Zha20] that captures when a measurement is "close" to being projective, in the sense that sequential applications of the measurement yield similar outcomes.

Definition 4.6. A real-valued measurement M on $\mathcal{H}$ is (ε, δ)**-almost-projective** if applying M twice in a row to any state $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H})$ produces measurement outcomes p, p' where

$$\Pr[|p - p'| \leq \varepsilon] \geq 1 - \delta \ .$$

We briefly discuss how alternating measurements A, B constitutes a (ε, δ)-almost projective approximation of $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}]$. While we will not make use of this fact directly (we prove a variant of it in Lemma 4.9), we will introduce some concepts and notation that are useful later. For $\vec{b} \in \{0,1\}^{n+1}$, and letting $Q_n \coloneqq \{0, 1/n, 2/n, \ldots, 1\}$, define

$$\mathsf{NReps}(\vec{b}) \coloneqq \frac{|\{j \in \{1, \ldots, n\} : b_{j-1} = b_j\}|}{n} \in Q_n \ .$$

That is, $\mathsf{NReps}(\vec{b})$ is the number of pairs of consecutive repeated bits in $\vec{b}$, divided by n; for example $p(0,0,1,1,1,0) = 3/5$. The following proposition is immediate from the definition of MWDist:

Proposition 4.7. *If* $\vec{b} \sim \mathsf{MWDist}(T,p)$ *then* $\mathsf{NReps}(1,\vec{b}) \sim \mathsf{Bin}(T,p)/T$.

Consider the measurement procedure M that applies T measurements $\mathsf{A},\mathsf{B},\mathsf{A},\mathsf{B},\ldots$ in an alternating fashion, and outputs $\mathsf{NReps}(b_1,\ldots,b_T)$, where the b_i are the measurement outcomes. Then by Lemma 4.5, for $|\psi\rangle = \sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle$, $\mathbb{E}_{p\leftarrow\mathsf{M}(|\psi\rangle)}[p] = |\alpha_j|^2 p_j$. Moreover, if $T \approx \frac{1}{\varepsilon}\log\frac{1}{\delta}$ then M is (ε,δ)-almost projective. This is because sequential applications of M are equivalent to a single application of M of length $2T$; (ε,δ)-almost projectivity follows by a Chernoff bound.

4.2 Probability estimation

We describe a measurement procedure ValEst that *estimates* $\omega_{\mathcal{G}}(S,\boldsymbol{\rho})$, following techniques of [MW05, Zha20]. The procedure is a variation on the "approximate projective implementation" procedure of [Zha20], and we show that it is (ε,δ)-almost projective. We also show that if $\mathsf{ValEst}(\boldsymbol{\rho})$ produces an outcome $\geq p$ with high probability, then $\omega_{\mathcal{G}}(S,\boldsymbol{\rho})$ cannot be much smaller than p.

A player with unitary strategy U_S and initial state $\boldsymbol{\rho}$ in the game $\mathcal{G} = (R,Z,f)$ receives a random challenge $r \leftarrow R$, applies U_S to $|r\rangle\langle r|^{\mathcal{R}} \otimes \boldsymbol{\rho}^{\mathcal{Z},\mathcal{I}}$, and sends $\mathcal{Z}$ to the referee; here $\mathcal{R}$ is supported on $\{|r\rangle\}_{r\in R}$, $\mathcal{Z}$ is supported on $\{|z\rangle\}_{z\in Z}$, and $\mathcal{I}$ denotes the player's internal registers.

The procedure ValEst is parameterized by $\varepsilon,\delta \in [0,1]$ and a game $\mathcal{G}$, and has black-box access to the player's unitary U_S and its inverse $U_S^\dagger$, and operates on registers $(\mathcal{Z},\mathcal{I})$.

We set

$$t \coloneqq t(\varepsilon,\delta) \coloneqq \max\{\lceil n_{\varepsilon/2,\delta/4}/2\rceil, \log_{5/8}(\delta/2)\} = O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right) \quad ,$$

where $n_{\varepsilon,\delta}$ is a parameter defined in Proposition 3.1 for the Chernoff bound. Let $\mathcal{R}'$ be a register with basis $\{|r\rangle\}_{r\in R} \cup \{|\top\rangle, |\bot\rangle\}$. We define the state $|+_R\rangle$ on $\mathcal{R}',\mathcal{R}$ as

$$|+_R\rangle \coloneqq \frac{1}{2}|\top,0\rangle + \frac{1}{2}|\bot,0\rangle + \frac{1}{\sqrt{2|R|}}\sum_{r\in R}|r,r\rangle \quad ,$$

where $\top$ and $\bot$ are arbitrary symbols distinct from the elements of R, and $0 \in R$.

Remark 4.8. *We introduce the auxiliary (control) register $\mathcal{R}'$ for two reasons:*

(a) $\mathcal{R}'$ has two additional basis elements $|\top\rangle, |\bot\rangle$. These are special symbols which correspond to "automatically" winning or losing the game, respectively. This forces our probability estimates to be scaled within the range $[1/4, 3/4]$, which can easily be rescaled to $[0,1]$ before outputting a final value. This modification ensures that the procedure terminates within a polynomial number of steps except with negligible probability.

(b) Tracing out the $\mathcal{R}'$ register leaves the classical mixed state $\frac{1}{2}|0\rangle\langle 0| + \frac{1}{2|R|}\sum_r |r\rangle\langle r|$ on $\mathcal{R}$; this ensures that U_S behaves as if it were invoked on random $|r\rangle$ (or 0, with probability $1/2$).

We are now ready to define the procedure ValEst.

$\mathsf{ValEst}^{U}_{\mathcal{G},\varepsilon,\delta}$:

1. Initialize registers $(\mathcal{R}', \mathcal{R})$ to $|+_R\rangle$;
2. Define $\mathsf{M}_{\mathcal{G}} := (\Pi_{\mathcal{G}}, \mathbf{I} - \Pi_{\mathcal{G}})$ where $\Pi_{\mathcal{G}} := U_S^\dagger \Pi_f U_S$ for
$$\Pi_f := \sum_{r,z,f(r,z)=1} |r,z\rangle\langle r,z|^{\mathcal{R}',\mathcal{Z}} + |\top\rangle\langle\top|^{\mathcal{R}'} \otimes I^{\mathcal{Z}} \ .$$
3. For $i = 1, \ldots, t$:
   (a) Apply $\mathsf{M}_{\mathcal{G}}$, obtaining outcome $L_{2i-1} \in \{0,1\}$.
   (b) Apply $\mathsf{M}_{|+_R\rangle} := \left(|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}}, \mathbf{I} - |+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}}\right)$, obtaining outcome $L_{2i} \in \{0,1\}$.
4. If $L_{2t} = 1$, skip to Step 5. Otherwise, apply $\mathsf{M}_{\mathcal{G}}, \mathsf{M}_{|+_R\rangle}$ to $\mathcal{A}$ in an alternating fashion until $\mathsf{M}_{|+_R\rangle} \to 1$, or a further $2t$ measurements have been applied.
5. Discard $\mathcal{R}$ and $\mathcal{R}'$; output $\tilde{p} := 2 \cdot \mathsf{NReps}(1, L_1, \ldots, L_{2t}) - 1/2$.

Lemma 4.9. *The measurement* $\mathsf{ValEst} := \mathsf{ValEst}^{S}_{\mathcal{G},\varepsilon,\delta}$ *has the following properties:*

(i) ValEst *is an oracle circuit of size* $O(|f| \cdot \frac{1}{\varepsilon}\log\frac{1}{\delta})$ *that applies* U_S *and* $U_S^\dagger$ $O(\frac{1}{\varepsilon}\log\frac{1}{\delta})$ *times;*

(ii) for every $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{Z},\mathcal{I})$, $\mathbb{E}_{\tilde{p}\leftarrow\mathsf{ValEst}(\boldsymbol{\rho})}[\tilde{p}] = \omega_{\mathcal{G}}(S,\boldsymbol{\rho})$*;*

(iii) ValEst *is* (ε,δ)*-almost projective;*

(iv) for every $p \in \mathbb{R}$*, if* $\Pr_{p'\leftarrow\mathsf{ValEst}(\boldsymbol{\rho})}[p' \geq p] \geq 1-\gamma$ *then* $\omega_{\mathcal{G}}(S,\boldsymbol{\rho}) \geq p - \gamma - \varepsilon - \delta$*;*

(v) for every $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{Z},\mathcal{I})$, $\omega_{\mathcal{G}}(S, \mathsf{ValEst}(\boldsymbol{\rho})) \geq \omega_{\mathcal{G}}(S,\boldsymbol{\rho}) - \delta$.

Proof. Item (i) follows directly from the description; we proceed to prove Items (ii) to (v). It suffices to prove each property for pure states $|\psi\rangle \in \mathcal{Z}\otimes\mathcal{I}$, as the statement for mixed states follows by convexity.

Consider a decomposition of $\mathcal{R}'\otimes\mathcal{R}\otimes\mathcal{Z}\otimes\mathcal{I}$ into the Jordan subspaces for projectors $\Pi_{\mathcal{G}}$ and $|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}}\otimes\mathbf{I}^{\mathcal{Z},\mathcal{I}}$ (henceforth we will write the projector $|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}}\otimes\mathbf{I}^{\mathcal{Z},\mathcal{I}}$ as $|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}}$). Following our notation for Jordan subspaces in Section 4.1, we will associate $\mathsf{M}_{\mathcal{G}}$ with A and $\mathsf{M}_{|+_R\rangle}$ with B, so that in the j-th Jordan subspace:

- $\Pi_{\mathcal{G}}$ is a projection onto $|v^{\mathsf{A}}_{j,1}\rangle$,
- $|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}}$ is a projection onto $|v^{\mathsf{B}}_{j,1}\rangle$, and
- $p_j = \left\|\langle v^{\mathsf{A}}_{j,1}|v^{\mathsf{B}}_{j,1}\rangle\right\|^2$.

Write $|+_R\rangle^{\mathcal{R},\mathcal{R}'}\otimes|\psi\rangle^{\mathcal{Z},\mathcal{I}} = \sum_j \alpha_j |v^{\mathsf{B}}_{j,1}\rangle$. Note that
$$\sum_j |\alpha_j|^2 p_j = \|\Pi_{\mathcal{G}}|+_R\rangle|\psi\rangle\|^2 = \frac{\omega_{\mathcal{G}}(S,|\psi\rangle)}{2} + \frac{1}{4} \ .$$

By Lemma 4.5, $\tilde{p} \leftarrow \mathsf{ValEst}(|\psi\rangle)$ is distributed as:

1. Choose j with probability $|\alpha_j|^2$.
2. Sample $L_1, \ldots, L_{2t} \leftarrow \mathsf{MWDist}(p_j, 2t)$.
3. Output $\tilde{p} \coloneqq 2p(1, L_1, \ldots, L_{2t}) - 1/2$.

Hence in particular we have that

$$\mathbb{E}[\tilde{p}] = 2\sum_j |\alpha_j|^2 \, \mathbb{E}[\mathsf{NReps}(1, L_1, \ldots, L_{2t})] - 1/2 = 2\sum_j |\alpha_j|^2 p_j - 1/2 = \omega_{\mathcal{G}}(S, |\psi\rangle) \ ,$$

which establishes (ii).

We now prove (iv). Suppose that $\Pr_{p' \leftarrow \mathsf{ValEst}(|\psi\rangle)}[p' \geq p] \geq 1 - \gamma$. Then

$$\gamma \geq \Pr_{\tilde{p} \leftarrow \mathsf{ValEst}(|\psi\rangle)}[\tilde{p} < p] = \sum_j |\alpha_j|^2 \Pr_{\vec{L} \leftarrow \mathsf{MWDist}(p_j, 2t)}[p(1, \vec{L}) < p/2 + 1/4] \geq \sum_{j, p_j < p/2 + 1/4 - \varepsilon} |\alpha_j|^2 (1 - \delta) \ ,$$

by Proposition 4.7. Rearranging,

$$\sum_{j, p_j < p/2 + 1/4 - \varepsilon} |\alpha_j|^2 \leq \gamma + \delta \ .$$

Hence

$$\omega_{\mathcal{G}}(S, |\psi\rangle) = 2\sum_j |\alpha_j|^2 p_j - 1/2 \geq p - \gamma - \varepsilon - \delta \ .$$

Next we prove (iii). Let D be the distribution on $Q_{2t} \times Q_{2t}$ arising from two sequential applications of ValEst with initial state $|\psi\rangle$ (recall that $Q_{2t} = \{0, \frac{1}{2t}, \frac{2}{2t}, \ldots, 1\}$). Let D' be sampled as follows.

1. Choose j with probability $|\alpha_j|^2$.
2. Sample $L_1, \ldots, L_{4t} \leftarrow \mathsf{MWDist}(p_j, 4t)$.
3. Sample $L'_1, \ldots, L'_{2t} \leftarrow \mathsf{MWDist}(p_j, 2t)$.
4. Compute $\tilde{p} \coloneqq p(1, L_1, \ldots, L_{2t})$ and $\tilde{p}' \coloneqq p(1, L'_1, \ldots, L'_{2t})$.
5. Output $(\tilde{p}, \tilde{p}')$.

The statistical distance between D and D' is bounded by $\Pr[\forall i \in [t, 2t],\ L_{2i} = 0]$. This can be shown by coupling the outcomes of the first $4t$ measurements with $L_1, \ldots, L_{4t}$ drawn by D'. If this bad event does not occur, the first application of ValEst terminates in some state $|+_R\rangle^{\mathcal{R}', \mathcal{R}} |\phi\rangle^{\mathcal{Z}, \mathcal{I}}$, and so tracing out $(\mathcal{R}', \mathcal{R})$ and then reinitializing it to $|+_R\rangle$ at the beginning of the second application of ValEst has no overall effect on the state. In this case, therefore, we can view the two applications of ValEst as a single alternating measurement procedure of length $2i + 2t$ conditioned on the outcome of the $2i$-th measurement being 1. Then by Lemma 4.5, in this case D and D' are identically distributed.

We now bound $\Pr[\forall i \in [t, 2t],\ L_{2i} = 0]$. Suppose that j is sampled in the first step. For each $i \in [t+1, \ldots, 2t]$, the probability that $L_{2i} = 1$ given that $L_{2i-2} = 0$ is $2p_j(1 - p_j)$. Note that for every subspace j where $|v^{\mathsf{B}}_{j,1}\rangle$ is nonzero, $p_j = \left\| \Pi_{\mathcal{G}} |v^{\mathsf{B}}_{j,1}\rangle \right\|^2 \in [1/4, 3/4]$; in particular, this holds for all subspaces j such that $\alpha_j \neq 0$. Hence for any j sampled with positive probability, $2p_j(1 - p_j) \geq 3/8$. It follows that the probability that $L_{2i} = 0$ for all $i \in [t, 2t]$ is at most $(5/8)^t \leq \delta/2$.

Finally we show that

$$\Pr_{(\tilde{p}, \tilde{p}') \leftarrow D'}[|\tilde{p} - \tilde{p}'| > \varepsilon] < \delta/2 \ ,$$

which will complete the proof. Observe that for j sampled in the first step, $\tilde{p}, \tilde{p}' \sim \mathsf{Bin}(2t, p_j)/2t$. Hence by Proposition 3.1 (Chernoff bound), $\Pr[|\tilde{p} - p_j| > \varepsilon/2] < \delta/4$, and similarly for $\tilde{p}'$. The equation follows by a union bound.

It remains to prove Item (v). Recall that for any state $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{Z}, \mathcal{I})$, we have that

$$\omega_{\mathcal{G}}(S, \boldsymbol{\rho}) = 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}}(\,|+_R\rangle\langle+_R| \otimes \boldsymbol{\rho})\Big) - 1/2.$$

Let ValEst' be defined identically to ValEst except that it does not discard $\mathcal{R}, \mathcal{R}'$. Since for all j, Π_j^{Jor} commutes with $\mathsf{M}_{\mathcal{G}}$,$\mathsf{M}_{|+_R\rangle}$, we have

$$\operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \mathsf{ValEst}'(\boldsymbol{\rho})\Big) = \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}}(\,|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}} \otimes \boldsymbol{\rho})\Big)\ .$$

Then we have

$$\begin{aligned}
\omega_{\mathcal{G}}(S, \mathsf{ValEst}(\boldsymbol{\rho})) &= 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}}(\,|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}} \otimes \mathsf{ValEst}(\boldsymbol{\rho}))\Big) - 1/2 \\
&= 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}}(\,|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}} \otimes \operatorname{Tr}_{\mathcal{R},\mathcal{R}'}(\mathsf{ValEst}'(\boldsymbol{\rho})))\Big) - 1/2 \\
&\geq 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}}(\,|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}} \otimes \operatorname{Tr}_{\mathcal{R},\mathcal{R}'}(\,|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}} \cdot \mathsf{ValEst}'(\boldsymbol{\rho})))\Big) - 1/2 \\
&= 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}}\,|+_R\rangle\langle+_R|^{\mathcal{R}',\mathcal{R}} \cdot \mathsf{ValEst}'(\boldsymbol{\rho})\Big) - 1/2 \\
&\geq 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \mathsf{ValEst}'(\boldsymbol{\rho})\Big) - 1/2 - \delta \\
&= 2 \sum_j p_j \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \boldsymbol{\rho}\Big) - 1/2 - \delta = \omega_{\mathcal{G}}(S, \boldsymbol{\rho}) - \delta\ ,
\end{aligned}$$

where the final inequality follows because $\operatorname{Tr}(\,|+_R\rangle\langle+_R|\,\mathsf{ValEst}'(\boldsymbol{\rho}))$ is at least the probability that ValEst terminates with $\mathsf{M}_{|+_R\rangle} \to 1$, which is at least $1 - \delta$. □

4.3 A state repair procedure

We construct a procedure $\mathsf{Repair}^{\mathsf{M}}(p)$ parameterized by an almost-projective measurement M and with input $p \in \mathbb{R}$ that (under certain conditions) outputs a state $\boldsymbol{\rho}$ satisfying the guarantee: "applying M to $\boldsymbol{\rho}$ produces an outcome $\approx p$ with high probability". We then obtain $\mathsf{ValRepair}$ by plugging in the almost-projective measurement ValEst for M.

The procedure. Formally, our state repair procedure $\mathsf{Repair}_T^{\mathsf{M},\mathsf{P}}$ is a CPTP map on a register $\mathcal{H}$, parameterized by:

- a positive integer T,
- an oracle for an (ε, δ)-almost-projective measurement M on $\mathcal{H}$, and
- an oracle for an N-outcome projective measurement $\mathsf{P} = (\Pi_k)_{k=1}^N$ on $\mathcal{H}$,

and taking classical inputs (k, p) where $k \in [N]$ and $p \in \mathbb{R}$.

Recall that the measurement $\mathsf{M} = (M_q)_{q \in I}$, where $I \subseteq \mathbb{R}$ is the set of outcomes of M, can be implemented as a unitary U_{M} on $(\mathcal{H}, \mathcal{W})$ for some ancilla register $\mathcal{W}$, followed by some projective measurement $(\Pi_{\mathsf{M},q})_{q \in I}$ on $\mathcal{W}$. Formally, for each $q \in I$, the unitary U_{M} and projector $\Pi_{\mathsf{M},q}$ satisfy $M_q \boldsymbol{\rho} M_q^\dagger = \mathrm{Tr}_{\mathcal{W}}(\Pi_{\mathsf{M},q} U_{\mathsf{M}}(\boldsymbol{\rho} \otimes |0\rangle\langle 0|^{\mathcal{W}}) U_{\mathsf{M}}^\dagger)$ for all $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H})$. We are now ready to give the state repair procedure.

$\mathsf{Repair}_T^{\mathsf{M},\mathsf{P}}(k, p)$:

1. Define measurements

$$\mathsf{A}_p \coloneqq (\Pi_{\mathsf{A},p}, \mathbf{I} - \Pi_{\mathsf{A},p}) \text{ where } \Pi_{\mathsf{A},p} \coloneqq \sum_{q \in [p \pm \varepsilon]} U_{\mathsf{M}}^\dagger \Pi_{\mathsf{M},q} U_{\mathsf{M}} \ ,$$
$$\mathsf{B}_k \coloneqq (\Pi_{\mathsf{B},k}, \mathbf{I} - \Pi_{\mathsf{B},k}) \text{ where } \Pi_{\mathsf{B},k} \coloneqq \Pi_k \otimes |0\rangle\langle 0|^{\mathcal{W}} \ .$$

2. Initialize $\mathcal{W}$ to $|0\rangle$.
3. Apply the measurement A_p. If the outcome is 1, skip to Step 5.
4. Apply the measurements $\mathsf{B}_k, \mathsf{A}_p, \mathsf{B}_k, \mathsf{A}_p, \ldots$ in alternating fashion until either (1) $\mathsf{A}_p \to 1$ occurs or (2) T applications of $(\mathsf{B}_k, \mathsf{A}_p)$ have been applied (whichever comes first).
5. Apply U_{M} to $(\mathcal{H}, \mathcal{W})$, and discard the $\mathcal{W}$ registers.

The following lemma describes the effect of the repair procedure.

Lemma 4.10 (State repair)**.** *Let* M *be an* (ε, δ)*-almost projective measurement on* $\mathcal{H}$*,* $\mathsf{P} = (\Pi_k)_{k=1}^N$ *be a projective measurement on* $\mathcal{H}$ *with* N *outcomes,* T *be a positive integer. Consider the following quantum measurement procedure* RepairExpt *on* $\mathcal{H}$*:*

1. *Measure the initial state: apply* M*, obtaining outcome* p*;*
2. *Damage the state: apply* P*, obtaining outcome* k*;*
3. *Repair the state: run* $\mathsf{Repair}_T^{\mathsf{M},\mathsf{P}}(k, p)$ *and let* R *denote the total number of calls to* M *and* P*.*
4. *Output* p*.*

Then RepairExpt *is* $(2\varepsilon, N(\delta + 1/T) + 4\sqrt{\delta})$*-almost projective, and* $\mathbb{E}[R] \leq N + 4T\sqrt{\delta} + 1$*.*

Proof of Lemma 4.10. We write out in full the steps applied in RepairExpt:

RepairExpt:

1. Apply M, obtaining outcome p;
2. Apply P, obtaining outcome $k \in [N]$.
3. Initialize $\mathcal{W}$ to $|0\rangle$.
4. Apply the measurement A_p. If the outcome is 1, skip to Step 6.
5. Apply the measurements $\mathsf{B}_k, \mathsf{A}_p, \mathsf{B}_k, \mathsf{A}_p, \ldots$ in alternating fashion until either (1) $\mathsf{A}_p \to 1$ occurs or (2) T applications of $(\mathsf{B}_k, \mathsf{A}_p)$ have been applied (whichever comes first).
6. Apply U_{M} to $(\mathcal{H}, \mathcal{W})$, and discard the $\mathcal{W}$ registers.

From this point on, we refer to Steps 4 to 6 as $\mathsf{Repair}'(k, p)$, which maps $\mathcal{H} \otimes \mathcal{W} \to \mathcal{H}$. Define the $(N+1)$-outcome projective measurement $\mathsf{P}' \coloneqq ((\Pi_{\mathsf{B},k})_{k=1}^N, \Pi_\perp)$ on $\mathcal{H} \otimes \mathcal{W}$ where $\Pi_\perp \coloneqq \mathbf{I}^{\mathcal{H}} \otimes (\mathbf{I} - |0\rangle\langle 0|)^{\mathcal{W}}$. Next consider the following experiment Expt_1 (differences highlighted in red).

Expt_1:
1. Apply M, obtaining outcome p;
2. Initialize $\mathcal{W}$ to $|0\rangle$.
3. Apply P', obtaining outcome $k \in [N] \cup \{\perp\}$.
4. Apply $\mathsf{Repair}'(k, p)$.

Expt_1 and $\mathsf{RepairExpt}$ are equivalent, since Expt_1 can be obtained by performing the following changes to $\mathsf{RepairExpt}$:

- Swap the order of Step 2 and Step 3 in $\mathsf{RepairExpt}$. This does not change the resulting experiment since P acts trivially on $\mathcal{W}$.
- Then, apply P' instead of P to obtain k. This causes no change since $\mathsf{P}'(\boldsymbol{\sigma}^{\mathcal{H}} \otimes |0\rangle\langle 0|^{\mathcal{W}}) = \mathsf{P}(\boldsymbol{\sigma}^{\mathcal{H}}) \otimes |0\rangle\langle 0|^{\mathcal{W}}$ for all $\boldsymbol{\sigma} \in \mathbf{S}(\mathcal{H})$.

We now define another experiment Expt_2 as follows (differences from Expt_1 highlighted in red).

Expt_2:
1. Apply M, obtaining outcome p;
2. Initialize $\mathcal{W}$ to $|0\rangle$.
3. Apply A_p to $(\mathcal{H}, \mathcal{W})$ and postselect on obtaining outcome 1.
4. Apply P', obtaining outcome $k \in [N] \cup \{\perp\}$.
5. Apply $\mathsf{Repair}'(k, p)$.

It will be convenient hereafter to treat M and P' as CPTP maps that write their output to a new output register, i.e., $\mathsf{M} \colon \mathbf{S}(\mathcal{H}) \to \mathbf{S}(\mathcal{H} \otimes \mathcal{O}_1)$ and $\mathsf{P}' \colon \mathbf{S}(\mathcal{H}) \to \mathbf{S}(\mathcal{H} \otimes \mathcal{O}_2)$. For the remainder of the proof, fix an initial state $\boldsymbol{\rho} \in \mathbf{S}(\mathcal{H})$. Let $\boldsymbol{\rho}_1$ denote the state on $(\mathcal{H}, \mathcal{W}, \mathcal{O}_1)$ directly before Step 3 in Expt_1 applied to $\boldsymbol{\rho}$. Let $\boldsymbol{\rho}_2$ denote the state on the same registers directly before Step 4 in Expt_2 applied to $\boldsymbol{\rho}$. We show that these states are close in trace distance.

Claim 4.11. *The trace distance between $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ is at most $2\sqrt{\delta}$.*

Proof. Let $\Pi'_{\mathsf{A}} := \sum_{p \in I} |p\rangle\langle p|^{\mathcal{O}_1} \otimes \Pi_{\mathsf{A},p}^{\mathcal{H},\mathcal{W}}$. We have that

$$\boldsymbol{\rho}_2 = \frac{\Pi'_{\mathsf{A}} \boldsymbol{\rho}_1 \Pi'_{\mathsf{A}}}{\mathrm{Tr}(\Pi'_{\mathsf{A}} \boldsymbol{\rho}_1)} \ .$$

Note that $\mathrm{Tr}(\Pi'_{\mathsf{A}} \boldsymbol{\rho}_1) = \mathrm{Tr}\Big(\Pi'_{\mathsf{A}} (\mathsf{M}(\boldsymbol{\rho}) \otimes |0\rangle\langle 0|^{\mathcal{W}})\Big)$ is equal to the probability that applying M twice in succession to $\boldsymbol{\rho}$ yields outcomes p, p' such that $|p - p'| \leq \varepsilon$, and hence is at least $1 - \delta$. The claim follows by the gentle measurement lemma (Lemma 3.3). □

To complete the proof of the lemma, we make use of the following key claim about Expt_2. Roughly speaking, we show that in Expt_2, if we obtain outcome $p \in I$ in Step 1 and an outcome $k \neq \perp$ in Step 4 (which occurs with probability at least $1 - 2\sqrt{\delta}$ due to Claim 4.11) where the probability of obtaining k was β, then the final state $\boldsymbol{\rho}^*$ after Step 5 has the following guarantee: applying M to $\boldsymbol{\rho}^*$ produces an outcome p' within 2ε of p except with probability inversely proportional to β.

Claim 4.12. *Fix $p \in I, k \in [N]$; let $|\phi_{\mathsf{A}}\rangle$ be an arbitrary state in* $\mathrm{image}(\Pi_{\mathsf{A},p}) \subseteq \mathcal{H} \otimes \mathcal{W}$*, and define $|\phi_{\mathsf{B}}\rangle := \Pi_{\mathsf{B},k} |\phi_{\mathsf{A}}\rangle / \sqrt{\beta}$ where $\beta := \|\Pi_{\mathsf{B},k} |\phi_{\mathsf{A}}\rangle\|^2$. Applying $\mathsf{Repair}'(k,p)$ to $|\phi_{\mathsf{B}}\rangle \in \mathcal{H} \otimes \mathcal{W}$ yields the state $\boldsymbol{\rho}^* \in \mathbf{S}(\mathcal{H})$ where*

$$\Pr_{p' \leftarrow \mathsf{M}(\boldsymbol{\rho}^*)}[|p' - p| > 2\varepsilon] \leq (\delta + 1/T)/\beta,$$

and $\mathsf{Repair}'(k,p)$ applies $1 + 1/\beta$ measurements in expectation.

We show how Lemma 4.10 follows from Claim 4.12, and subsequently prove Claim 4.12.

Write $\boldsymbol{\rho}_2 = \sum_{p \in I} |p\rangle\langle p|^{\mathcal{O}_1} \otimes \boldsymbol{\rho}_p^{\mathcal{H},\mathcal{W}}$; note that $\mathrm{Tr}(\Pi_{\mathsf{A},p} \boldsymbol{\rho}_p) = \mathrm{Tr}(\boldsymbol{\rho}_p)$ due to the post-selection in Step 3. By the definition of $\mathsf{P}' \colon \mathbf{S}(\mathcal{H}) \to \mathbf{S}(\mathcal{H} \otimes \mathcal{O}_2)$,

$$\mathsf{P}'(\boldsymbol{\rho}_2) = \sum_{p \in I} |p\rangle\langle p|^{\mathcal{O}_1} \otimes \left(\Pi_\perp \boldsymbol{\rho}_p^{\mathcal{H},\mathcal{W}} \Pi_\perp \otimes |\perp\rangle\langle\perp|^{\mathcal{O}_2} + \sum_{k=1}^{N} \Pi_{\mathsf{B},k} \boldsymbol{\rho}_p^{\mathcal{H},\mathcal{W}} \Pi_{\mathsf{B},k} \otimes |k\rangle\langle k|^{\mathcal{O}_2} \right) .$$

By Claim 4.11, $\mathrm{Tr}(\Pi_\perp \boldsymbol{\rho}_2) \leq \mathrm{Tr}(\Pi_\perp \boldsymbol{\rho}_1) + 2\sqrt{\delta} = 2\sqrt{\delta}$. For $p \in I$, write $\boldsymbol{\rho}_p = \sum_i q_i |\psi_i\rangle\langle\psi_i|$ for unit states $|\psi_i\rangle \in \mathcal{H} \otimes \mathcal{W}$; note that $|\psi_i\rangle \in \mathrm{image}(\Pi_{\mathsf{A},p})$. For all i and any $k \in [N]$, we can define $|\psi_{i,k}\rangle := \Pi_{\mathsf{B},k} |\psi_i\rangle / \|\Pi_{\mathsf{B},k} |\psi_i\rangle\|$ and apply Claim 4.12 with $|\phi_{\mathsf{A}}\rangle$ set to $|\psi_i\rangle$ to obtain

$$\Pr_{p' \leftarrow \mathsf{M}(\boldsymbol{\rho}^*_{i,p,k})}[|p' - p| > 2\varepsilon] \leq (\delta + 1/T)/\|\Pi_{\mathsf{B},k} |\psi_i\rangle\|^2 ,$$

where $\boldsymbol{\rho}^*_{i,p,k} \in \mathbf{S}(\mathcal{H})$ is the state after applying $\mathsf{Repair}'(k,p)$ to $|\psi_{i,k}\rangle$.

To conclude, we show that Expt_2 is $(2\varepsilon, N(\delta + 1/T) + 2\sqrt{\delta})$-almost projective; the statement for $\mathsf{RepairExpt}$ will then follow by Claim 4.11. Let $\mathcal{O}_3$ be a new ancilla register that will store the outcome of the *second* application of Expt_2, and consider the projector $\Pi_{\mathsf{bad}} := \sum_p \sum_{p' \notin [p \pm 2\varepsilon]} |p, p'\rangle\langle p, p'|^{\mathcal{O}_1, \mathcal{O}_3}$ corresponding to the event that applying Expt_2 twice yields outcomes (p, p') more than 2ε apart. Since the outcome of Expt_2 is determined by the outcome of M, we have by convexity

$$\mathrm{Tr}(\Pi_{\mathsf{bad}} \cdot \mathsf{M}(\mathsf{Expt}_2(\boldsymbol{\rho}))) \leq N(\delta + 1/T) + 2\sqrt{\delta} .$$

Hence by Claim 4.11,

$$\mathrm{Tr}(\Pi_{\mathsf{bad}} \cdot \mathsf{M}(\mathsf{RepairExpt}(\boldsymbol{\rho}))) \leq N(\delta + 1/T) + 4\sqrt{\delta} ,$$

which completes the proof that $\mathsf{RepairExpt}$ is $(2\varepsilon, N(\delta + 1/T) + 4\sqrt{\delta})$-almost projective.

By Claim 4.11 and Claim 4.12, and law of total expectation, it holds that

$$\begin{aligned}
\mathbb{E}[R] &\leq d(\boldsymbol{\rho}'_1, \boldsymbol{\rho}'_2) \cdot T + \mathrm{Tr}(\Pi_\perp \boldsymbol{\rho}'_2) \cdot T + \sum_{p \in I, k \in [N]} \mathrm{Tr}(\Pi_{\mathsf{B},k} \boldsymbol{\rho}_p)(1 + \mathrm{Tr}(\boldsymbol{\rho}_p)/\mathrm{Tr}(\Pi_{\mathsf{B},k} \boldsymbol{\rho}_p)) \\
&\leq 2T\sqrt{\delta} + 2T\sqrt{\delta} + \sum_{p \in I, k \in [N]} (\mathrm{Tr}(\Pi_{\mathsf{B},k} \boldsymbol{\rho}_p) + \mathrm{Tr}(\boldsymbol{\rho}_p)) \\
&\leq N + 4T\sqrt{\delta} + 1 ,
\end{aligned}$$

which concludes the proof, given Claim 4.12. □

Proof of Claim 4.12. For this proof, we write A, B for $\mathsf{A}_p, \mathsf{B}_k$ and $\Pi_{\mathsf{A}}, \Pi_{\mathsf{B}}$ for $\Pi_{\mathsf{A},p}, \Pi_{\mathsf{B},k}$ respectively.

Consider a decomposition of $\mathcal{H} \otimes \mathcal{W}$ into the Jordan subspaces $\{\mathcal{S}_j\}_j$ for projectors Π_{A} and Π_{B}. Following our standard notation for Jordan subspaces, in the j-th Jordan subspace $\mathcal{S}_j$, Π_{A} is

a projection onto $|v^{\mathsf{A}}_{j,1}\rangle$ and Π_{B} is a projection onto $|v^{\mathsf{B}}_{j,1}\rangle$, and $p_j = |\langle v^{\mathsf{A}}_{j,1}|v^{\mathsf{B}}_{j,1}\rangle|^2$. Recall that we write Π^{Jor}_j for the projection onto $\mathcal{S}_j$.

Since $|\phi_{\mathsf{A}}\rangle = \sum_j \alpha_j |v^{\mathsf{A}}_{j,1}\rangle$ for some choice of $\{\alpha_j\}_j$, we can write $|\phi_{\mathsf{B}}\rangle$ as

$$|\phi_{\mathsf{B}}\rangle = \frac{1}{\sqrt{\beta}} \sum_j \alpha_j \Pi_{\mathsf{B}} |v^{\mathsf{A}}_{j,1}\rangle = \frac{1}{\sqrt{\beta}} \sum_j \alpha_j \sqrt{p_j} |v^{\mathsf{B}}_{j,1}\rangle$$

Let $\boldsymbol{\rho}' \in \mathbf{S}(\mathcal{H} \otimes \mathcal{W})$ be the state immediately before "Apply U_{M} to $(\mathcal{H}, \mathcal{W})$, and discard the $\mathcal{W}$ registers." in $\mathsf{Repair}'(k,p)$, so that $\boldsymbol{\rho}^* = \mathrm{Tr}_{\mathcal{W}}(U_{\mathsf{M}} \boldsymbol{\rho}' U^\dagger_{\mathsf{M}})$. We first bound $\mathrm{Tr}(\Pi_{\mathsf{A}} \boldsymbol{\rho}')$, i.e., the probability that $\mathsf{Repair}'(k,p)$ stops because $\mathsf{A} \to 1$, by analyzing the distribution of measurement outcomes that result from applying a total of $2T+1$ alternating measurements $\mathsf{A}, \mathsf{B}, \mathsf{A}, \mathsf{B}, \ldots, \mathsf{A}$. Note that the real Repair procedure terminates after obtaining a 1 outcome for A; we consider the distribution of a fixed number of measurements for the purpose of analysis.

Let $I(b_1, b_2, \ldots, b_{2T+1})$ be the smallest i such that $b_{2i+1} = 1$, or $T+1$ if there is no such i. Let D be denote the following distribution:

1. Sample j with probability $|\alpha_j|^2 p_j / \beta$
2. Sample $(b_1, b_2, \ldots, b_{2T+1}) \leftarrow \mathsf{MWDist}(2T+1, p_j)$.
3. Output $I(b_1, b_2, \ldots, b_{2T+1})$.

By Lemma 4.5, the expected number of measurements applied by Repair is $2\,\mathbb{E}[D] + 1$, and

$$\mathrm{Tr}(\Pi_{\mathsf{A}} \boldsymbol{\rho}') = 1 - \Pr_{i \leftarrow D}[i = T+1] \ .$$

We now analyse the distribution D. Suppose that j is sampled in Step 1. The probability that $b_1 = 1$ occurs is then p_j. Then for each $i \in [T]$, the probability that $b_{2i+1} = 1$ given that $b_{2i-1} = 0$ is $2p_j(1-p_j)$. Hence conditioned on j being sampled, D is dominated by the random variable D' which takes value 0 with probability p_j and is distributed as $\mathsf{Geo}(2p_j(1-p_j))$ with probability $1-p_j$, where $\mathsf{Geo}(q)$ is the geometric distribution with parameter q.

It follows that $\mathbb{E}[D] \le \frac{1}{\beta} \sum_j |\alpha_j|^2 p_j(1-p_j)\, \mathbb{E}[\mathsf{Geo}(2p_j(1-p_j))] = 1/(2\beta)$, and

$$\Pr_{i \leftarrow D}[i = T+1] \le \frac{1}{\beta} \sum_j |\alpha_j|^2 p_j(1-p_j)(1-2p_j(1-p_j))^T \le \frac{1}{\beta T} \ ,$$

since $x(1-2x)^T \le 1/T$ for all $x \in [0, 1/4]$. This establishes that $\mathrm{Tr}(\Pi_{\mathsf{A}} \boldsymbol{\rho}') \ge 1 - \frac{1}{\beta T}$.

To complete the proof of Claim 4.12, we prove that applying M to $\boldsymbol{\rho}^* = \mathrm{Tr}_{\mathcal{W}}(U_{\mathsf{M}} \boldsymbol{\rho}' U^\dagger_{\mathsf{M}})$ produces p' within 2ε of p with probability at least $1 - (\delta + 1/T)/\beta$.

Since A and B commute with Π^{Jor}_j, $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_j \boldsymbol{\rho}'\right) = \left\| \Pi^{\mathrm{Jor}}_j |\phi_{\mathsf{B}}\rangle \right\|^2 = |\alpha_j|^2 p_j / \beta$. In particular, η as defined in Claim 4.13 is equal to β, and $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_j \boldsymbol{\rho}'\right) = 0$ for all j with $p_j = 0$. By definition, the last measurement applied during Repair is A, and so since A is projective, $\boldsymbol{\rho}' = \mathsf{A}(\boldsymbol{\rho}') = \Pi_{\mathsf{A}} \boldsymbol{\rho}' \Pi_{\mathsf{A}} + (I - \Pi_{\mathsf{A}}) \boldsymbol{\rho}' (I - \Pi_{\mathsf{A}})$, which commutes with Π_{A}. The statement then follows by Claim 4.13. □

Claim 4.13. *Suppose $\boldsymbol{\rho}' \in \mathbf{S}(\mathcal{H} \otimes \mathcal{W})$ satisfies each of the following:*

- $\mathrm{Tr}(\Pi_{\mathsf{A}} \boldsymbol{\rho}') = 1 - \gamma$,
- *$\boldsymbol{\rho}'$ commutes with Π_{A}, and*
- $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_j \boldsymbol{\rho}'\right) = 0$ *for all j where $p_j = 0$.*

Let

$$\eta \coloneqq \frac{1}{\sum_{j,p_j>0} \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \boldsymbol{\rho}'\Big)/p_j} \ ,$$

and $\boldsymbol{\rho}^* \coloneqq \operatorname{Tr}_{\mathcal{W}}(U_{\mathsf{M}} \boldsymbol{\rho}' U_{\mathsf{M}}^\dagger)$. *Then*

$$\Pr_{p' \leftarrow \mathsf{M}(\boldsymbol{\rho}^*)}[|p' - p| > 2\varepsilon] \leq \delta/\eta + \gamma \ .$$

Proof. Since $\boldsymbol{\rho}'$ commutes with Π_{A}, we can write $\boldsymbol{\rho}' = \sum_i q_i \left|\phi_i\right\rangle\!\left\langle\phi_i\right|$, where the $\left|\phi_i\right\rangle$ are eigenstates of Π_{A}. Consider the unitary U on $\mathcal{H} \otimes \mathcal{W}$ that maps $\left|v_{j,b}^{\mathsf{A}}\right\rangle$ to $\left|v_{j,1-b}^{\mathsf{A}}\right\rangle$ for $b \in \{0,1\}$ for each 2-dimensional Jordan subspace $\mathcal{S}_j$, and acts as identity on each 1-dimensional subspace. Formally,

$$U \coloneqq \sum_{j,p_j \notin \{0,1\}} \left(\left|v_{j,1}^{\mathsf{A}}\right\rangle \left\langle v_{j,0}^{\mathsf{A}}\right| + \left|v_{j,0}^{\mathsf{A}}\right\rangle \left\langle v_{j,1}^{\mathsf{A}}\right|\right) + \sum_{j,p_j=1} \left|v_{j,1}^{\mathsf{A}}\right\rangle\!\left\langle v_{j,1}^{\mathsf{A}}\right| + \sum_{j,p_j=0} \left|v_{j,0}^{\mathsf{A}}\right\rangle\!\left\langle v_{j,0}^{\mathsf{A}}\right| \ .$$

In particular, if $\left|\phi_i\right\rangle = \sum_{j,p_j>0} \zeta_j \left|v_{j,0}^{\mathsf{A}}\right\rangle$, then $U\left|\phi_i\right\rangle = \sum_{j,p_j>0} \zeta_j \left|v_{j,1}^{\mathsf{A}}\right\rangle \in \operatorname{image}(\Pi_{\mathsf{A}})$. Moreover, Π_j^{Jor} commutes with U for all j.

Let $\boldsymbol{\sigma} \coloneqq \Pi_{\mathsf{A}} \boldsymbol{\rho}' + U(\mathbf{I} - \Pi_{\mathsf{A}}) \boldsymbol{\rho}' U^\dagger$. $\boldsymbol{\sigma}$ does not appear during the procedure; it is defined for the purpose of analysis. Intuitively, $\boldsymbol{\sigma}$ is the result of rotating, within each Jordan subspace, the part of $\boldsymbol{\rho}'$ in $\operatorname{image}(\mathbf{I} - \Pi_{\mathsf{A}})$ into $\operatorname{image}(\Pi_{\mathsf{A}})$. By unitary invariance of the trace,

$$\operatorname{Tr}(\boldsymbol{\sigma}) = \operatorname{Tr}(\Pi_{\mathsf{A}} \boldsymbol{\rho}') + \operatorname{Tr}\Big(U(\mathbf{I} - \Pi_{\mathsf{A}}) \boldsymbol{\rho}' U^\dagger\Big) = \operatorname{Tr}(\boldsymbol{\rho}') = 1 \ .$$

For all j, we have $\operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \boldsymbol{\sigma}\Big) = \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \boldsymbol{\rho}'\Big)$ since Π_j^{Jor} commutes with both U and Π_{A}. The trace distance between $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}'$ is at most $\operatorname{Tr}((\mathbf{I} - \Pi_{\mathsf{A}}) \boldsymbol{\rho}') = \gamma$. Finally, by definition of U, $\operatorname{Tr}(\Pi_{\mathsf{A}} \boldsymbol{\sigma}) = 1$.

We will now show that the outcome of $\mathsf{M}(\operatorname{Tr}_{\mathcal{W}}(U_{\mathsf{M}} \boldsymbol{\sigma} U_{\mathsf{M}}^\dagger))$ is in the range $p \pm 2\varepsilon$ with probability δ/η, which will complete the proof by contractivity of the trace distance. Define the linear operator $C \coloneqq \sum_{j,p_j>0} \frac{1}{\sqrt{p_j}} \left|v_{j,1}^{\mathsf{B}}\right\rangle \left\langle v_{j,1}^{\mathsf{A}}\right|$. Notice that $\Pi_{\mathsf{A}} C$ is the projection onto $\operatorname{image}(\Pi_{\mathsf{A}}) \cap (\bigoplus_{j,p_j>0} \mathcal{S}_j)$ since

$$\Pi_{\mathsf{A}} C = \sum_{j,p_j>0} \frac{1}{\sqrt{p_j}} \left|v_{j,1}^{\mathsf{A}}\right\rangle\!\left\langle v_{j,1}^{\mathsf{A}}\right| \left|v_{j,1}^{\mathsf{B}}\right\rangle \left\langle v_{j,1}^{\mathsf{A}}\right| = \sum_{j,p_j>0} \left|v_{j,1}^{\mathsf{A}}\right\rangle\!\left\langle v_{j,1}^{\mathsf{A}}\right| \ .$$

Let $\boldsymbol{\sigma}' \coloneqq C \boldsymbol{\sigma} C^\dagger / \operatorname{Tr}\Big(C \boldsymbol{\sigma} C^\dagger\Big)$. We have that

$$\operatorname{Tr}\Big(C \boldsymbol{\sigma} C^\dagger\Big) = \operatorname{Tr}\Big(C^\dagger C \boldsymbol{\sigma}\Big) = \sum_{j,p_j>0} \frac{1}{p_j} \operatorname{Tr}\Big(\left|v_{j,1}^{\mathsf{A}}\right\rangle\!\left\langle v_{j,1}^{\mathsf{A}}\right| \boldsymbol{\sigma}\Big) = \sum_{j,p_j>0} \frac{1}{p_j} \operatorname{Tr}\Big(\Pi_j^{\mathrm{Jor}} \boldsymbol{\rho}'\Big) = 1/\eta,$$

and $\operatorname{Tr}(\Pi_{\mathsf{B}} \boldsymbol{\sigma}') = 1$. By the definition of Π_{B}, this implies that $\boldsymbol{\sigma}' = \boldsymbol{\sigma}'' \otimes \left|0\right\rangle\!\left\langle 0\right|^{\mathcal{W}}$ for some $\boldsymbol{\sigma}'' \in \mathbf{S}(\mathcal{H})$. We also have that $\Pi_{\mathsf{A}} \boldsymbol{\sigma}' \Pi_{\mathsf{A}} = \eta \Pi_{\mathsf{A}} C \boldsymbol{\sigma} C^\dagger \Pi_{\mathsf{A}} = \eta \boldsymbol{\sigma}$, where the second equality follows from the fact that $\boldsymbol{\sigma} \in \mathbf{S}(\operatorname{image}(\Pi_{\mathsf{A}}) \cap (\bigoplus_{j,p_j>0} \mathcal{S}_j))$ by construction.

Recall that (1) applying U_{M} to a state of the form $\boldsymbol{\rho}'' \otimes \left|0\right\rangle\!\left\langle 0\right|^{\mathcal{W}}$, then applying the projective measurement $(\Pi_{\mathsf{M},q})_{q \in I}$ on $\mathcal{W}$ and tracing out $\mathcal{W}$ is equivalent to applying the (ε, δ)-almost-projective measurement $\mathsf{M} = (M_q)_{q \in I}$ to $\boldsymbol{\rho}''$ and (2) $\Pi_{\mathsf{A}} = \sum_{q \in [p \pm \varepsilon]} U_{\mathsf{M}}^\dagger \Pi_{\mathsf{M},q} U_{\mathsf{M}}$. So we have:

$$\boldsymbol{\rho}^* = \operatorname{Tr}_{\mathcal{W}}(U_{\mathsf{M}} \boldsymbol{\sigma} U_{\mathsf{M}}^\dagger) = \frac{1}{\eta} \operatorname{Tr}_{\mathcal{W}}(U_{\mathsf{M}} \Pi_{\mathsf{A}} \boldsymbol{\sigma}' \Pi_{\mathsf{A}} U_{\mathsf{M}}^\dagger)$$

$$= \frac{1}{\eta} \operatorname{Tr}_{\mathcal{W}} \left(\sum_{q,q' \in [p \pm \varepsilon]} \Pi_{\mathsf{M},q} U_{\mathsf{M}}(\boldsymbol{\sigma}'' \otimes |0\rangle\langle 0|^{\mathcal{W}}) U_{\mathsf{M}}^\dagger \Pi_{\mathsf{M},q'} \right)$$
$$= \frac{1}{\eta} \sum_{q \in [p \pm \varepsilon]} \operatorname{Tr}_{\mathcal{W}} \left(\Pi_{\mathsf{M},q} U_{\mathsf{M}}(\boldsymbol{\sigma}'' \otimes |0\rangle\langle 0|^{\mathcal{W}}) U_{\mathsf{M}}^\dagger \right)$$
$$= \frac{1}{\eta} \sum_{q \in [p \pm \varepsilon]} M_q \boldsymbol{\sigma}'' M_q^\dagger = \frac{\sum_{q \in [p \pm \varepsilon]} M_q \boldsymbol{\sigma}'' M_q^\dagger}{\operatorname{Tr}\left(\sum_{q \in [p \pm \varepsilon]} M_q \boldsymbol{\sigma}'' M_q^\dagger \right)} \ .$$

That is, $\boldsymbol{\rho}^*$ is the state after applying M to $\boldsymbol{\sigma}''$ conditioned on obtaining an outcome in the range $p \pm \varepsilon$, which occurs with probability η. But then by (ε, δ)-almost projectivity, the outcome of $\mathsf{M}(\boldsymbol{\rho}^*)$ is in the range $p \pm 2\varepsilon$ with probability $1 - \delta/\eta$. □

4.4 Proof of Theorem 4.3

We now prove Theorem 4.3. For $r \in R$, define $\mathsf{M}_{f,r} \coloneqq \left(\Pi_{f,r}^{\mathcal{Z},\mathcal{I}}, \mathbf{I} - \Pi_{f,r}^{\mathcal{Z},\mathcal{I}} \right)$ where

$$\Pi_{f,r}^{\mathcal{Z},\mathcal{I}} \coloneqq U_{S,r}^\dagger \sum_{z, f(r,z)=1} |z\rangle\langle z|^{\mathcal{Z}} U_{S,r} \ ,$$

where $U_{S,r}$ is a unitary implementation of the action of S on message r.

We set ValEst and $\mathsf{ValRepair}$ as follows:

- Let $\mathsf{ValEst}^S_{\mathcal{G},\varepsilon,\delta}$ be a CPTP map from $(\mathcal{Z},\mathcal{I})$ to $(\mathcal{Z},\mathcal{I})$ as in Lemma 4.9.
- Let $\mathsf{ValRepair}^S_{\mathcal{G},\varepsilon,\delta,T,r} \coloneqq \mathsf{Repair}_T^{\mathsf{ValEst}^S_{\mathcal{G},\varepsilon,\delta},\mathsf{M}_{f,r}}$ be a CPTP map from $(\mathcal{Z},\mathcal{I})$ to $(\mathcal{Z},\mathcal{I})$ as in Lemma 4.10 (that is, with $\mathcal{H} = (\mathcal{Z},\mathcal{I})$).

The algorithm A operates on registers $(\mathcal{Z},\mathcal{I})$ and works as follows.

$A^S_{\mathcal{G},n,\eta_0}$:
1. Let $\varepsilon \coloneqq \eta_0/(2n+2)$, $\delta \coloneqq \eta_0^2/cn^2$ for some universal constant c.
2. (Main loop.) For $i = 1, \ldots, n$,
   (a) Measure $p_i \leftarrow \mathsf{ValEst}^S_{\mathcal{G},\varepsilon,\delta}$ on registers $(\mathcal{Z},\mathcal{I})$.
   (b) Receive $r_i \in R$ from the referee and apply U_{S,r_i} to $(\mathcal{Z},\mathcal{I})$.
   (c) Send the register $\mathcal{Z}$ to the referee.
   (d) Receive the (partially measured) register $\mathcal{Z}$ from the referee, along with the outcome $b_i \in \{0,1\}$.
   (e) Apply $U_{S,r_i}^\dagger$ to $(\mathcal{Z},\mathcal{I})$.
   (f) Apply $\mathsf{ValRepair}^S_{\mathcal{G},\varepsilon,\delta,T,r_i}(p,b)$ to $(\mathcal{Z},\mathcal{I})$ with $T \coloneqq \lceil 1/\sqrt{\delta} \rceil$.

Claim 4.14. *For each $i \in [n]$, $p_{i+1} \geq p_i - 2\varepsilon$ with probability $1 - O(\sqrt{\delta})$.*

Proof. Steps 2(b) to 2(e) are equivalent to applying M_{f,r_i} to $(\mathcal{Z},\mathcal{I})$. Since $\mathsf{ValEst}^S_{\mathcal{G},\varepsilon,\delta}$ is (ε,δ)-almost projective (Lemma 4.9, Item (iii)), the claim follows from applying Lemma 4.10 with $\mathsf{M} = \mathsf{ValEst}^S_{\mathcal{G},\varepsilon,\delta}$, $\mathcal{H} = (\mathcal{Z},\mathcal{I})$, $\mathsf{P} = \mathsf{M}_{f,r_i}$, $N = 2$, $T = \lceil 1/\sqrt{\delta} \rceil$ and observing that the entire "Main loop" amounts to a single invocation of $\mathsf{RepairExpt}$, and is therefore a $(2\varepsilon, O(\sqrt{\delta}))$-almost-projective measurement.[13] □

[13] On the $(i+1)$-th invocation of the main loop the challenge r_{i+1} will generally be different than the challenge r_i used in the i-th invocation; however, almost projectivity still applies since p_{i+1} is clearly independent of r_{i+1}.

Let $\boldsymbol{\rho}_i$ be the state on $(\mathcal{Z}, \mathcal{I})$ at the beginning of the i-th iteration.

Claim 4.15. *For all $i \in [n]$, $\omega_{\mathcal{G}}(S, \boldsymbol{\rho}_i) \geq \omega_{\mathcal{G}}(S, \boldsymbol{\rho}) - 2i \cdot \varepsilon - O(i \cdot \sqrt{\delta})$.*

Proof. By Claim 4.14, with probability $1 - O(i \cdot \sqrt{\delta})$ it holds that

$$p_i \geq p_{i-1} - 2\varepsilon \geq p_{i-2} - 4\varepsilon \geq \cdots \geq p_1 - 2(i-1)\varepsilon \ .$$

Then by Lemma 4.9, Item (iv),

$$\omega_{\mathcal{G}}(S, \boldsymbol{\rho}_i) \geq \mathbb{E}[p_1] - 2i \cdot \varepsilon - O(i \cdot \sqrt{\delta}) \ .$$

Finally, by Lemma 4.9, Item (ii), $\mathbb{E}[p_1] = \omega_{\mathcal{G}}(S, \boldsymbol{\rho})$. □

Now since $\omega_{\mathcal{G}}(S, \mathsf{ValEst}(\boldsymbol{\sigma})) \geq \omega_{\mathcal{G}}(S, \boldsymbol{\sigma}) - \delta$ for all states $\boldsymbol{\sigma}$ by Lemma 4.9, Item (v), we have that $\Pr[b_i = 1] \geq \omega_{\mathcal{G}}(S, \boldsymbol{\rho}) - 2i \cdot \varepsilon - O(i \cdot \sqrt{\delta})$. Hence

$$\begin{aligned} \omega^n_{\mathcal{G}}(A^S_{\mathcal{G},n,\eta_0}, \boldsymbol{\rho}) = \mathbb{E}[\sum_{i \in [n]} b_i] &= \sum_{i \in [n]} \Pr[b_i = 1] \\ &\geq n \cdot (\omega_{\mathcal{G}}(S, \boldsymbol{\rho}) - (n+1)\varepsilon - O(n \cdot \sqrt{\delta})) \\ &\geq n \cdot (\omega_{\mathcal{G}}(S, \boldsymbol{\rho}) - \eta_0) \ , \end{aligned}$$

which completes the proof. The expected running time of this procedure is $\tilde{O}(|f| \cdot n/\eta_0)$.

5 A quantum rewinding lemma

We use Theorem 4.3 to prove a "quantum forking lemma" for collapsing protocols. We denote by $(\tau, \boldsymbol{\rho}) \leftarrow \langle \tilde{P}, V \rangle_{m-1}$ the partial transcript τ and intermediate state $\boldsymbol{\rho}$ of the malicious prover $\tilde{P}$ after running $m-1$ rounds of the interaction between $\tilde{P}$ and V. Recall that R_m denotes the set of random coins for round m of the protocol.

Theorem 5.1. *Let (P, V) be an m-round collapsing protocol. There exists an algorithm* Fork *running in expected polynomial time with black-box access to an adversary such that the following holds. Let $\tilde{P}$ be an efficient quantum adversary such that $\Pr\left[\langle \tilde{P}, V \rangle \to 1\right] \geq \eta$. Then for any $n \in \mathbb{N}$, $\eta_0 \in [0,1]$,*

$$\mathbb{E}\left[\, |W| \;\middle|\; \begin{array}{r} (\tau, \boldsymbol{\rho}) \leftarrow \langle \tilde{P}, V \rangle_{m-1} \\ \vec{r} = (r_1, \ldots, r_n) \leftarrow (R_m)^n \\ W \leftarrow \mathsf{Fork}^{\tilde{P}}(1^\lambda, 1^{1/\eta_0}, \tau, \vec{r}, \boldsymbol{\rho}) \end{array} \right] \geq n(\eta - \eta_0) - n^2/|R_m| - \mathrm{negl}(\lambda) \ .$$

Moreover, with probability 1, we have $\{(s_i, z_i)\}_i \leftarrow \mathsf{Fork}^{\tilde{P}}(1^\lambda, 1^{1/\eta_0}, \tau, \vec{r}, \boldsymbol{\rho})$ where:

- *$V(\tau, s_i, z_i) = 1$ holds for all $i \in [k]$,*
- *all s_i are distinct, and*
- *for each i there exists $j \in [n]$ such that $s_i = r_k$.*

Fork *runs in expected time* $\mathrm{poly}(\lambda) \cdot \tilde{O}(n/\eta_0)$.

Proof. We define an interactive quantum algorithm C that acts as the referee in an n-round single-player quantum game as in Section 4. For $r \in R$, define

$$\Pi_{V,r} \coloneqq \sum_{z, V(\tau,r,z)=1} |z\rangle\langle z| \ .$$

$C(\tau, \vec{r})$:

1. Set $W \coloneqq \emptyset$. For $j = 1, \ldots, n$,
   (a) Send $r_j \in R$ to the player.
   (b) Receive register $\mathcal{Z}$ from the player.
   (c) Apply the binary measurement $\mathsf{M}_{V,r_j} \coloneqq \left(\Pi_{V,r_j}, \mathbf{I} - \Pi_{V,r_j}\right)$ to register $\mathcal{Z}$, obtaining outcome b.
   (d) If $b = 1$, measure $\mathcal{Z}_m$ in the computational basis to obtain response z. If there is no z' such that $(r_j, z') \in W$, set $W \leftarrow W \cup \{(r_j, z)\}$.
   (e) Return register $\mathcal{Z}$ to the player.
2. Output W.

The extractor Fork is obtained by simulating $\langle A^{U^{(m)}}_{\mathcal{G},n,\eta_0}, C(\tau, \vec{r}) \rangle$, where $A_{\mathcal{G},n,\eta_0}$ is the algorithm guaranteed by Theorem 4.3 with $\mathcal{G} \coloneqq (R_m, Z_m, V(\tau, \cdot, \cdot))$, and $U^{(m)}$ is the unitary that the prover applies in the final round. The properties of the output of Fork (aside from the expected size of W) follow immediately from the definition.

By the collapsing property of (P, V), the measurement in Step 1(d) is undetectable to any efficient distinguisher; in particular, it is undetectable to A. We can therefore apply Theorem 4.3 to show that the expected number of successful iterations is at least $n(\eta - \eta_0) - \mathrm{negl}(\lambda)$. The expected number of repeated r_j is at most $n^2/|R|$, which yields the bound. □

Remark 5.2. *If the quantum prover $\tilde{P}$ has (non-uniform)* quantum *advice, then in general we can only run the extractor once.*

However, if the malicious quantum prover $\tilde{P}$ has (non-uniform) classical *advice, we can generate $(\tau, \boldsymbol{\rho}) \leftarrow \langle \tilde{P}, V \rangle_{m-1}$ as many times as we would like (obtaining a different $(\tau, \boldsymbol{\rho})$ each time). By running $\mathsf{Fork}^{\tilde{P}}$ on each $(\tau, \boldsymbol{\rho})$, we eventually obtain a set W of accepting transcripts with a shared prefix τ where $|W| \geq n(\eta - \eta_0) - n^2/|R_m|$ with probability arbitrarily close to 1.*

5.1 Special sound protocols

Theorem 5.1 immediately implies that any collapsing k-special sound protocol is an argument of knowledge. We first define k-special soundness, and then briefly explain how to apply Theorem 5.1 to obtain this result. Recall that a *sigma protocol* is a three-message protocol where the prover moves first.

Definition 5.3 (Special soundness)**.** A sigma protocol (P, V) is *k-special sound* if there exists an extractor Ext such that, given k accepting transcripts $(a, r_1, z_1), \ldots, (a, r_k, z_k)$ with all $r_i \in R$ distinct, $\mathsf{Ext}(x, a, (r_i, z_i)_{i=1}^k)$ outputs w such that $(x, w) \in \mathfrak{R}$.

Theorem 5.4. *Any collapsing k-special sound protocol is a post-quantum argument of knowledge with knowledge error $O(k/|R|)$.*

Proof sketch. Let $\tilde{P}$ be an adversary that convinces V with probability $\varepsilon > 4k/|R|$. The extractor E for (P, V) operates as follows, where Fork is as guaranteed by Theorem 5.1.

1. Obtain first message a from $\tilde{P}$; let $\boldsymbol{\rho}$ be the prover's state after sending a.
2. Sample $\vec{r} = (r_1, \ldots, r_n) \leftarrow R^n$ uniformly at random, where $n = 8k/\varepsilon$.
3. Run $W \leftarrow \mathsf{Fork}^{\tilde{P}}(1^\lambda, 1^{1/\eta_0}, a, \vec{r}, \boldsymbol{\rho})$, for $\eta_0 = \Theta(\varepsilon)$ to be chosen.
4. If $|W| \geq k$, output $w \leftarrow E_{\mathsf{ss}}(x, a, W)$.

η_0 can be chosen such that $\mathbb{E}[|W|] \geq n(\varepsilon - 2k/|R|)/2 \geq n\varepsilon/4$, and so the probability that $|W| \geq n\varepsilon/8 = k$ is at least $\varepsilon/8$ by Markov's inequality. The theorem follows by the definition of special soundness. □

For constant k, Theorem 5.4 states that the post-quantum knowledge error of any k-special sound collapsing sigma protocol is $O(1/|R|)$, which asymptotically matches the classical knowledge error. Previously, the post-quantum knowledge error of such protocols was only shown to be $O(1/\sqrt{|R|})$ via Unruh's rewinding lemma [Unr12].

Remark 5.5. *Theorem 5.4 alone is insufficient to imply post-quantum security of Kilian's protocol (when instantiated with a PCP of knowledge), since Kilian's protocol is* not *k-special sound for any $k = \mathrm{poly}(\lambda)$. In particular, k-special soundness requires successful extraction from* any *set of k accepting transcripts with distinct challenges r_i; the extractor for Kilian's protocol requires that the r_i are also "sufficiently random". We therefore prove post-quantum security of Kilian's protocol in Section 7 by directly applying Theorem 5.1 to obtain accepting transcripts for randomly sampled r_i.*

6 Collapsing vector commitments

We define collapsing vector commitments (Section 6.1), and then prove that Merkle trees are collapsing vector commitments when the underlying hash function is collapsing (Section 6.2). Later on, in Section 7, we will formulate Kilian's protocol in terms of vector commitments, and establish its post-quantum security when the vector commitment is collapsing.

6.1 Definition

A (static) vector commitment scheme VC [CF13] consists of the following algorithms.

- $\mathsf{VC.Gen}(1^\lambda, \Sigma, \ell)$ is a probabilistic algorithm that takes as input the security parameter 1^λ, an alphabet Σ, and a vector length $\ell \in \mathbb{N}$, and outputs a commitment key ck.
- $\mathsf{VC.Commit}(\mathsf{ck}, m)$ is a (possibly probabilistic) algorithm that takes as input a commitment key ck and a vector $m \in \Sigma^\ell$, and outputs a commitment string cm and auxiliary information aux.
- $\mathsf{VC.Open}(\mathsf{ck}, \mathsf{aux}, Q)$ is a deterministic algorithm that takes as input a commitment key ck, auxiliary information aux, and a subset $Q \subseteq [\ell]$, and outputs an opening proof pf.
- $\mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q, v, \mathsf{pf})$ is a deterministic algorithm that takes as input a commitment key ck, a commitment cm, a subset $Q \subseteq [\ell]$, alphabet symbols $v \in \Sigma^Q$, and an opening proof pf, and outputs a bit $b \in \{0,1\}$.

The vector commitment scheme VC is *complete* if for every security parameter λ, alphabet Σ, vector length $\ell \in \mathbb{N}$, and adversary Adv,

$$\Pr\left[\mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q, m[Q], \mathsf{pf}) = 1 \;\middle|\; \begin{array}{r} \mathsf{ck} \leftarrow \mathsf{VC.Gen}(1^\lambda, \Sigma, \ell) \\ (m \in \Sigma^\ell, Q \subseteq [\ell]) \leftarrow \mathsf{Adv}(\mathsf{ck}) \\ (\mathsf{cm}, \mathsf{aux}) \leftarrow \mathsf{VC.Commit}(\mathsf{ck}, m) \\ \mathsf{pf} \leftarrow \mathsf{VC.Open}(\mathsf{ck}, \mathsf{aux}, Q) \end{array}\right] = 1 \ .$$

The traditional definition of security for a vector commitment scheme is *position binding*, which states that no efficient attacker can open any location to two different values. In more detail, for every security parameter λ, alphabet Σ, vector length $\ell \in \mathbb{N}$, and polynomial-size (classical or quantum) adversary Adv,

$$\Pr\left[\begin{array}{c} \exists\, i \in Q_1 \cap Q_2 \text{ s.t. } v_1[i] \neq v_2[i] \\ \wedge\, \mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q_1, v_1, \mathsf{pf}_1) = 1 \\ \wedge\, \mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q_2, v_2, \mathsf{pf}_2) = 1 \end{array} \;\middle|\; \begin{array}{r} \mathsf{ck} \leftarrow \mathsf{VC.Gen}(1^\lambda, \Sigma, \ell) \\ \left(\mathsf{cm},\ \begin{array}{l} Q_1 \subseteq [\ell], v_1 \in \Sigma^{Q_1}, \mathsf{pf}_1 \\ Q_2 \subseteq [\ell], v_2 \in \Sigma^{Q_2}, \mathsf{pf}_2 \end{array}\right) \leftarrow \mathsf{Adv}(\mathsf{ck}) \end{array}\right] = \mathrm{negl}(\lambda) \ .$$

While position binding against classical adversaries suffices to prove security of Kilian's protocol against classical adversaries, it is not known whether position binding against quantum adversaries suffices to prove security of Kilian's protocol against quantum adversaries. (And, as discussed in Section 1, it is unlikely to.) Hence we rely on an additional *collapsing* property that we introduce.

Definition 6.1. VC is *collapsing* if for every security parameter λ, alphabet Σ, vector length $\ell \in \mathbb{N}$, and polynomial-size quantum adversary Adv,

$$\left|\Pr[\mathsf{VCCollapseExp}(0, \lambda, \Sigma, \ell, \mathsf{Adv}) = 1] - \Pr[\mathsf{VCCollapseExp}(1, \lambda, \Sigma, \ell, \mathsf{Adv}) = 1]\right| \leq \mathrm{negl}(\lambda) \ .$$

For $b \in \{0,1\}$ the experiment $\mathsf{VCCollapseExp}(b, \lambda, \Sigma, \ell, \mathsf{Adv})$ is defined as follows:

1. The challenger samples $\mathsf{ck} \leftarrow \mathsf{VC.Gen}(1^\lambda, \Sigma, \ell)$ and sends ck to Adv.
2. Adv replies with a classical message $(\mathsf{cm}, Q \subseteq [\ell])$, and a quantum state on registers $(\mathcal{V}, \mathcal{O})$, where the $\mathcal{V}$ registers contain strings $v \in \Sigma^Q$ and the $\mathcal{O}$ registers contain opening proofs pf.
3. The challenger computes into an ancilla register the bit $\mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q, \mathcal{V}, \mathcal{O})$ via some unitary U, measures the ancilla, and then applies $U^\dagger$ to uncompute. If the measured bit is 0 (verification fails), the challenger aborts and outputs $\perp$.
4. If $b = 0$, the challenger does nothing. If $b = 1$, the challenger measures the registers $(\mathcal{V}, \mathcal{O})$ in the standard basis to obtain a string v and opening proof pf, which it discards.
5. The challenger returns the contents of the (potentially measured) registers $(\mathcal{V}, \mathcal{O})$ to Adv.
6. Adv outputs a bit b, which is the output of the experiment.

Remark 6.2. *The definition of collapse binding for standard commitments implies (classical-style) binding [Unr16b]. However, we do not know whether our definition of collapsing for vector commitments implies position binding in general, without imposing additional structure on the vector commitment.*

6.2 Merkle trees are collapsing

We describe Merkle trees as an instance of vector commitments (Section 6.1), and then prove that they are collapsing when the underlying hash function is collapsing.

Construction 6.3. Let $\mathcal{H} = \{H_\lambda\}_{\lambda \in \mathbb{N}}$ be a function family with input size $n(\lambda)$ and output size $\ell(\lambda) = n(\lambda)/2$. Let $\mathsf{VC} := \mathsf{Merkle}[\mathcal{H}]$ be the vector commitment for messages over alphabet $\Sigma := \{0,1\}^{n(\lambda)}$ that is constructed as follows.

- $\mathsf{VC.Gen}(1^\lambda, \Sigma, \ell)$: sample a hash function $h \leftarrow H_\lambda$ and output the commitment key $\mathsf{ck} := (\ell, h)$.
- $\mathsf{VC.Commit}(\mathsf{ck}, m)$: use $h\colon \{0,1\}^{n(\lambda)} \to \{0,1\}^{n(\lambda)/2}$ to pairwise hash the message m to obtain a corresponding Merkle tree tr with root $\mathsf{rt} \in \{0,1\}^{n(\lambda)/2}$, and then output $\mathsf{cm} := \mathsf{rt}$ as a commitment and $\mathsf{aux} := (m, \mathsf{tr})$ as auxiliary information.
- $\mathsf{VC.Open}(\mathsf{ck}, \mathsf{aux}, Q)$: for each index $i \in Q$, deduce the authentication path path_i for index i in the Merkle tree tr, and then output the opening proof $\mathsf{pf} := (\mathsf{path}_i)_{i \in Q}$. (Some of the paths may have overlaps, in which case the opening proof pf can be compressed accordingly.)
- $\mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q, v, \mathsf{pf})$: for each index $i \in Q$, check that the authentication path path_i in pf is for messages of length ℓ, and that it authenticates the value v_i for location i in a Merkle tree with root cm.

It is well-known that Merkle trees satisfy the position binding property.

Claim 6.4. *If $\mathcal{H}$ is a collision-resistant hash function with input size $n(\lambda)$ and output size $\ell(\lambda) = \frac{n(\lambda)}{2}$ against classical (resp., quantum) adversaries then $\mathsf{VC} := \mathsf{Merkle}[\mathcal{H}]$ is a position-binding vector commitment scheme over alphabet $\Sigma := \{0,1\}^{n(\lambda)}$ against classical (resp., quantum) adversaries.*

We now show that if $\mathcal{H}$ is a collapsing hash function then $\mathsf{VC} := \mathsf{Merkle}[\mathcal{H}]$ is a collapsing vector commitment.

Claim 6.5. *If $\mathcal{H}$ is a collapsing hash function with input size $n(\lambda)$ and output size $\ell(\lambda) = \frac{n(\lambda)}{2}$ then $\mathsf{VC} := \mathsf{Merkle}[\mathcal{H}]$ is a collapsing vector commitment over alphabet $\Sigma := \{0,1\}^{n(\lambda)}$.*

Proof. The proof is a standard application of the collapsing hash function security property. We write the proof for the case of a singleton query set $Q = \{i\}$; extending to the general case is straightforward.

Fix a message length ℓ, and let $d := \lceil \log_2 \ell \rceil$ be the height of a Merkle tree for messages of length ℓ. For $j \in \{0, 1, \ldots, d\}$, we define a hybrid experiment $\mathbf{H}_j$ as follows:

1. The challenger samples $h \leftarrow H_\lambda$ and sends h to Adv.
2. Adv replies with a classical message $(\mathsf{rt}, i \in [\ell])$ (a Merkle root and a location) and a quantum state on registers $(\mathcal{V}, \mathcal{O}_1, \ldots, \mathcal{O}_d)$, where the register $\mathcal{V}$ corresponds to strings in Σ^Q and each register $\mathcal{O}_j$ corresponds to the j-th node in the Merkle opening proof ($j = 1$ is a leaf node). For convenience we set $\mathcal{Y}_1 := \mathcal{V}$.
3. The challenger coherently applies $\mathsf{VC.Verify}$ using d ancilla registers $\mathcal{Y}_2, \ldots, \mathcal{Y}_{d+1}$:
   (a) Let U_i be a unitary on the registers $(\mathcal{O}_1, \ldots, \mathcal{O}_d, \mathcal{Y}_1, \ldots, \mathcal{Y}_{d+1})$ that works as follows: for $k = 1, \ldots, d$, apply h to $(\mathcal{Y}_k, \mathcal{O}_k)$ or $(\mathcal{O}_k, \mathcal{Y}_k)$ (depending on the k-th bit of i) and XOR the result onto $\mathcal{Y}_k$.
   (b) Apply U_i and then measure the bit indicating whether $\mathcal{Y}_{d+1}$ equals rt (by applying the binary projective measurement $\left(|\mathsf{rt}\rangle\langle\mathsf{rt}|^{\mathcal{Y}_{d+1}}, \mathbf{I} - |\mathsf{rt}\rangle\langle\mathsf{rt}|^{\mathcal{Y}_{d+1}} \right)$). If the measured bit is 0 (verification fails), then the challenger aborts and outputs $\perp$.
4. The challenger measures registers $(\mathcal{O}_{d-j+1}, \ldots, \mathcal{O}_d)$ and $(\mathcal{Y}_{d-j+1}, \ldots, \mathcal{Y}_{d+1})$. (If $j = 0$ then the challenger does not measure any of the $\mathcal{O}$ registers.)
5. The challenger applies $U_i^\dagger$ to uncompute the $\mathcal{Y}_2, \ldots, \mathcal{Y}_{d+1}$ registers, and returns the registers $(\mathcal{V}, \mathcal{O}_1, \ldots, \mathcal{O}_d)$ to the adversary Adv.

Hybrid $\mathbf{H}_0$ corresponds to the experiment $\mathsf{VCCollapseExp}(0, \lambda, \Sigma, \ell, \mathsf{Adv})$ and hybrid $\mathbf{H}_d$ corresponds to the experiment $\mathsf{VCCollapseExp}(1, \lambda, \Sigma, \ell, \mathsf{Adv})$, for the vector commitment scheme $\mathsf{VC} := \mathsf{Merkle}[\mathcal{H}]$. (See Definition 6.1 for the definition of the collapsing experiment for VC.)

We are left to argue that, for each $j \in \{0, 1, \ldots, d-1\}$, $\mathbf{H}_j$ and $\mathbf{H}_{j+1}$ are indistinguishable. Suppose by way of contradiction that for some $j \in \{0, 1, \ldots, d-1\}$ the attacker Adv can distinguish $\mathbf{H}_j$ and $\mathbf{H}_{j+1}$ with advantage at least ϵ. We construct an adversary Adv_j that has distinguishing advantage at least ϵ for $\mathcal{H}$'s collapsing experiment $\mathsf{HCollapseExp}(b, \lambda, \mathsf{Adv}_j)$ (see Definition 3.9). The adversary Adv_j works as follows.

1. Receive a hash function h from the challenger.
2. Send h to Adv, and obtain the message (rt, i) and a quantum state on registers $(\mathcal{V}, \mathcal{O}_1, \ldots, \mathcal{O}_d)$.
3. Similarly to the challenger in the hybrids, set $\mathcal{V} := \mathcal{Y}_1$, prepare d internal ancilla registers $\mathcal{Y}_2, \ldots, \mathcal{Y}_{d+1}$, and apply the same unitary U_i on $(\mathcal{O}_1, \ldots, \mathcal{O}_d, \mathcal{Y}_1, \ldots, \mathcal{Y}_{d+1})$.
4. Measure the bit indicating whether $\mathcal{Y}_{d+1}$ equals the Merkle root rt, and aborts if this measurement does not return 1.
5. Measure $(\mathcal{O}_{d-j+1}, \ldots, \mathcal{O}_d)$ and $(\mathcal{Y}_{d-j+1}, \ldots, \mathcal{Y}_{d+1})$.
6. Forward the contents of $(\mathcal{O}_{d-j}, \mathcal{Y}_{d-j})$ to the challenger as the hash function input, and forward $\mathcal{Y}_{d-j}$ as the classical output. (If $b = 0$, the challenger in the collapsing experiment will not disturb the state on $(\mathcal{O}_{d-j}, \mathcal{Y}_{d-j})$; if instead $b = 1$, the challenger measures $(\mathcal{O}_{d-j}, \mathcal{Y}_{d-j})$ before returning these registers to Adv_j.)
7. Apply U_i again and return the registers $(\mathcal{V}, \mathcal{O}_1, \ldots, \mathcal{O}_d)$ to Adv.
8. Output whatever Adv outputs.

The proof is concluded by observing that Adv's view when inside the experiment $\mathsf{HCollapseExp}(0, \lambda, \mathsf{Adv}_j)$ corresponds to hybrid $\mathbf{H}_j$ and Adv's view when inside the experiment $\mathsf{HCollapseExp}(1, \lambda, \mathsf{Adv}_j)$ corresponds to hybrid $\mathbf{H}_{j+1}$. □

7 Post-quantum security of Kilian's protocol

Denote by $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ the instantiation of Kilian's protocol with PCP system PCP and vector commitment scheme VC (see Section 7.2 below). We prove the following theorem.

Theorem 7.1. *Let* PCP *be a PCP system for* $\mathfrak{R}$ *with negligible soundness error, and let* VC *be a collapsing vector commitment. Then* $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ *is a post-quantum succinct argument for* $\mathfrak{R}$*. Moreover, if* PCP *has negligible knowledge error, then* $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ *is also a post-quantum succinct argument of knowledge for* $\mathfrak{R}$*.*

Corollary 7.2. *Assuming the post-quantum hardness of LWE, there exist post-quantum succinct arguments for* NP*.*

Proof. Collapsing vector commitments can be obtained from collapsing hash functions (Claim 6.5), which in turn exist based on the post-quantum hardness of LWE [Unr16a]. The corollary follows from Theorem 7.1 applied to a PCP for NP with suitable efficiency (e.g., [BFLS91]). □

The rest of this section is organized as follows: in Section 7.1 we recall the definition of a PCP; in Section 7.2 we describe Kilian's protocol and prove that $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ is collapsing if VC is a collapsing vector commitment; in Section 7.3 we prove Theorem 7.1.

7.1 Probabilistically checkable proofs

A *probabilistically checkable proof* (PCP) for a relation $\mathfrak{R}$ with soundness error $\varepsilon_{\mathsf{PCP}}$, alphabet Σ, and proof length ℓ, is a pair of polynomial-time algorithms $\mathsf{PCP} = (\mathbf{P}_{\mathsf{PCP}}, \mathbf{V}_{\mathsf{PCP}})$ satisfying the following.

- **Completeness.** For every instance-witness pair $(x, w) \in \mathfrak{R}$, $\mathbf{P}_{\mathsf{PCP}}(x, w)$ outputs a proof string $\pi \colon [\ell] \to \Sigma$ such that $\Pr\left[\mathbf{V}^{\pi}_{\mathsf{PCP}}(1^{\lambda}, x) = 1\right] = 1$.
- **Soundness.** For every instance $x \notin \mathcal{L}(\mathfrak{R})$ and proof string $\pi \colon [\ell] \to \Sigma$, $\Pr\left[\mathbf{V}^{\pi}_{\mathsf{PCP}}(1^{\lambda}, x) = 1\right] \leq \varepsilon_{\mathsf{PCP}}$.

Probabilities are taken over the randomness r of $\mathbf{V}_{\mathsf{PCP}}$. The *randomness complexity* rc is the number of random bits used by $\mathbf{V}_{\mathsf{PCP}}$, and the *query complexity* qc is the number of locations of π read by $\mathbf{V}_{\mathsf{PCP}}$. The quantities $\varepsilon_{\mathsf{PCP}}, \ell, \Sigma, \mathsf{rc}, \mathsf{qc}$ can be functions of the instance size $|x|$.

We also consider PCPs that achieve a *proof of knowledge* property, which is a strengthening of the soundness property.

- **Proof of knowledge.** PCP has knowledge error κ_{PCP} if there exists a polynomial-time extractor algorithm $\mathbf{E}$ such that, for every instance x and proof string $\pi \colon [\ell] \to \Sigma$, if $\Pr[\mathbf{V}^{\pi}_{\mathsf{PCP}}(x) = 1] > \kappa_{\mathsf{PCP}}$ then $\mathbf{E}(x, \pi)$ outputs w such that $(x, w) \in \mathfrak{R}$.

7.2 Kilian's protocol

Kilian's protocol [Kil92] is a public-coin four-message interactive argument $\mathsf{ARG} = (P, V)$ obtained by combining two ingredients:

- a PCP system $\mathsf{PCP} = (\mathbf{P}_{\mathsf{PCP}}, \mathbf{V}_{\mathsf{PCP}})$ with alphabet Σ, proof length ℓ, randomness complexity rc, and query complexity qc; and

- a VC scheme $\mathsf{VC} = (\mathsf{Gen}, \mathsf{Commit}, \mathsf{Open}, \mathsf{Verify})$ over alphabet Σ.

The construction of the interactive argument, which we denote by $(P, V) \coloneqq \mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$, is specified below. The argument prover P and argument verifier V receive as input a security parameter λ (in unary) and an instance x, while P additionally receives as input a witness w for x.

1. V samples a commitment key $\mathsf{ck} \leftarrow \mathsf{VC.Gen}(\lambda, \ell)$ and sends ck to P.
2. P computes a PCP string $\pi \leftarrow \mathbf{P}_{\mathsf{PCP}}(x, w)$, computes a commitment to it $(\mathsf{cm}, \mathsf{aux}) \leftarrow \mathsf{VC.Commit}(\mathsf{ck}, \pi)$, and sends cm to V.
3. V samples PCP randomness $r \leftarrow \{0,1\}^{\mathsf{rc}}$ and sends r to P.
4. P runs the PCP verifier $\mathbf{V}^{\pi}_{\mathsf{PCP}}(x; r)$ to deduce a set $Q \subseteq [\ell]$ of queries made by $\mathbf{V}_{\mathsf{PCP}}$, computes an opening proof $\mathsf{pf} \leftarrow \mathsf{VC.Open}(\mathsf{ck}, \mathsf{aux}, Q)$, and sends $(\pi[Q], \mathsf{pf})$ to V.
5. V checks that $\mathbf{V}_{\mathsf{PCP}}(x; r)$ accepts when answering its PCP queries via $\pi[Q] \in \Sigma^Q$ and that $\mathsf{VC.Verify}(\mathsf{ck}, \mathsf{cm}, Q, \pi[Q], \mathsf{pf}) = 1$. (If $\mathbf{V}_{\mathsf{PCP}}$ makes any query outside of Q then reject.)

We show that $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ is a collapsing protocol when VC is collapsing.

Claim 7.3. *If* VC *is a collapsing vector commitment then for all* PCP*,* $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ *is a collapsing protocol.*

Proof. Consider an adversary Adv for $\mathsf{CollapseExp}$ for Kilian. We construct an Adv' for $\mathsf{VCCollapseExp}$ with the same advantage as follows:

1. Obtain ck from the challenger and send it to Adv. Measure the response cm.
2. Choose $r \leftarrow \{0,1\}^{\mathsf{rc}}$ and send it to Adv. Send (cm, Q) and the (unmeasured) state on $\mathcal{Z}_2$ to the challenger, where Q is the query set corresponding to r.
3. Receive a state on $\mathcal{Z}_2$ and pass it to Adv. Return the output of Adv. □

7.3 Proof of Theorem 7.1

Since Kilian's protocol instantiated with a collapsing vector commitment VC is collapsing (Claim 7.3), there exists an algorithm $\mathsf{Fork}^{\tilde{P}}$ making black-box queries to any malicious prover $\tilde{P}$ for Kilian's protocol that satisfies the guarantees of Theorem 5.1. We use $\mathsf{Fork}^{\tilde{P}}$ to implement an extractor $E^{\tilde{P}}$ that makes black-box queries to $\tilde{P}$ and outputs a PCP string $\pi \in \Sigma^{\ell}$.

$E^{\tilde{P}(x;|\psi\rangle)}(1^{\lambda}, x, 1^{1/\varepsilon})$:

1. Sample a commitment key $\mathsf{ck} \leftarrow \mathsf{VC.Gen}(\lambda, \ell)$ and query $\tilde{P}$ on ck to obtain a commitment cm. Let $\tau \coloneqq (\mathsf{ck}, \mathsf{cm})$, and let $\boldsymbol{\rho}$ denote the intermediate state of $\tilde{P}$.
2. Set $n \coloneqq 60\ell \cdot \log(2|\Sigma|)/\varepsilon$, sample $\vec{r} = (r_1, \ldots, r_n)$ uniformly at random from $(\{0,1\}^{\mathsf{rc}})^n$, and run $(\tau, (r_1, z_1), \ldots, (r_k, z_k)) \leftarrow \mathsf{Fork}^{\tilde{P}}(1^{\lambda}, 1^{3/\varepsilon}, \tau, \vec{r}, \boldsymbol{\rho})$. Abort if $k < 6\ell \cdot \log(2|\Sigma|)$.
3. Parse each z_i as $(\pi[Q_{r_i}], \mathsf{pf})$, where Q_{r_i} is defined to be the set of indices that $\mathbf{V}_{\mathsf{PCP}}(x)$ queries on random coins r_i.
4. Check that $\{(Q_{r_i}, \pi[Q_{r_i}]\}_{i \in [k]}$ are *consistent*, meaning that there does not exist a PCP index t with two different values. If this check fails, abort and output $\perp$.
5. Output a π obtained by combining the answers given in $\{(Q_{r_i}, \pi[Q_{r_i}])\}_{i \in [k]}$ and filling in any unanswered indices arbitrarily.

Claim 7.4. $\Pr\left[\perp \leftarrow E^{\tilde{P}}\right] \leq 1 - \Omega(\varepsilon) + \mathrm{negl}(\lambda)$

Proof. We first bound the probability that E aborts in Step 2. Define η_{ck} to be the probability that $\tilde{P}$ wins when ck is sampled in the first round; note that $\mathbb{E}_{\mathsf{ck}}[\eta_{ck}] \geq \varepsilon$. By Theorem 5.1, $\mathbb{E}[k|\mathsf{ck}] \geq \eta_{\mathsf{ck}} - \gamma \cdot \varepsilon$ for some $\gamma < 1$. Hence by Markov's inequality,

$$\Pr[k < 6\ell \cdot \log(2|\Sigma|)] = 1 - \Omega(\varepsilon) \ .$$

By the position-binding property of VC, the probability that $E^{\tilde{P}}$ aborts in Step 4 is $\mathrm{negl}(\lambda)$.

It follows that $\Pr\left[\pi \leftarrow E^{\tilde{P}}\right] \geq \Omega(\varepsilon) - \mathrm{negl}(\lambda)$. □

For a PCP π, let $\mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \coloneqq \Pr[\mathbf{V}^{\pi}_{\mathsf{PCP}}(x) = 1]$. We prove that conditioned on the event that $\pi \leftarrow E^{\tilde{P}}$, we have $\mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \geq k/(2n)$ with overwhelming probability.

Claim 7.5. $\Pr\left[(\pi \leftarrow E^{\tilde{P}}) \wedge (\pi \neq \bot) \wedge (\mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) < k/(2n))\right] \leq \mathrm{negl}(\lambda)$.

Proof. We first argue that for any fixed string $\pi^* \in \Sigma^{\ell}$ where $\mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi^*) < k/(2n)$, we have:

$$\Pr\left[\pi^* = \pi \wedge \pi \leftarrow E^{\tilde{P}}\right] \leq (2|\Sigma|)^{-\ell}.$$

The probability $E^{\tilde{P}}$ outputs such a π^* is upper bounded by the probability that for randomly sampled $(r_1, \ldots, r_n)$, there *exist* k distinct r_i such that $\mathbf{V}^{\pi^*}_{\mathsf{PCP}}(x; r_i) = 1$. For each r_i, the probability $\Pr\left[\mathbf{V}^{\pi^*}_{\mathsf{PCP}}(x; r_i) = 1\right] < k/(2n)$, so by a multiplicative Chernoff bound (Proposition 3.2) we have

$$\Pr_{r_1,\ldots,r_n}[\text{exists } k \text{ distinct } i \in [n] \text{ such that } \mathbf{V}^{\pi^*}_{\mathsf{PCP}}(x; r_i) = 1] \leq e^{-k/6} = (2|\Sigma|)^{-\ell} \ .$$

A union bound over all π^* completes the proof:

$$\begin{aligned}
&\Pr\left[(\pi \leftarrow E^{\tilde{P}}) \wedge (\pi \neq \bot) \wedge (\mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) < k/(2n))\right] \\
&= \sum_{\pi^*, \mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi^*) < k/(2n)} \Pr\left[\pi^* = \pi \wedge \pi \leftarrow E^{\tilde{P}}\right] \\
&\leq |\Sigma|^{\ell}/(2|\Sigma|)^{\ell} = \mathrm{negl}(\lambda).
\end{aligned}$$ □

By combining Claims 7.4 and 7.5 with the fact that $k/(2n) \geq \varepsilon/20$, we obtain

$$\Pr\left[(\pi \leftarrow E^{\tilde{P}}) \wedge (\mathsf{win}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \geq \varepsilon/20)\right] \geq \Omega(\varepsilon) - \mathrm{negl}(\lambda).$$

If PCP has negligible soundness error, then this implies that $\varepsilon = \mathrm{negl}(\lambda)$.

If PCP is a proof of knowledge with negligible knowledge error $\kappa_{\mathsf{PCP}} = \mathrm{negl}(\lambda)$ with witness extractor $\mathbf{E}$, then the following extractor achieves knowledge error $\kappa = \mathrm{negl}(\lambda)$: run $\pi \leftarrow E^{\tilde{P}}$ and output $w \leftarrow \mathbf{E}(x, \pi)$.

Acknowledgements

Part of this work was done while FM was visiting UC Berkeley and the Simons Institute for the Theory of Computing from Fall 2019 to Spring 2020. AC is supported by the Ethereum Foundation. FM thanks Justin Holmgren for helpful discussions. NS is supported by DARPA under Agreement No. HR00112020023. NS thanks Dominique Unruh for helpful discussions.

References

[AGKZ20] Ryan Amos, Marios Georgiou, Aggelos Kiayias, and Mark Zhandry. One-shot signatures and applications to hybrid quantum/classical authentication. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 255–268, 2020.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998. Preliminary version in FOCS '92.

[ARU14] Andris Ambainis, Ansis Rosmanis, and Dominique Unruh. Quantum attacks on classical proof systems: The hardness of quantum rewinding. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 474–483, 2014.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: a new characterization of NP. *Journal of the ACM*, 45(1):70–122, 1998. Preliminary version in FOCS '92.

[BBC+18] Carsten Baum, Jonathan Bootle, Andrea Cerulli, Rafaël del Pino, Jens Groth, and Vadim Lyubashevsky. Sub-linear lattice-based zero-knowledge arguments for arithmetic circuits. In *Proceedings of the 38th Annual International Cryptology Conference*, CRYPTO '18, pages 669–699, 2018.

[BDF+11] Dan Boneh, Özgür Dagdelen, Marc Fischlin, Anja Lehmann, Christian Schaffner, and Mark Zhandry. Random oracles in a quantum world. In *Proceedings of the 17th International Conference on the Theory and Application of Cryptology and Information Security*, ASIACRYPT '11, pages 41–69, 2011.

[BFLS91] László Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy. Checking computations in polylogarithmic time. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing*, STOC '91, pages 21–32, 1991.

[BISW17] Dan Boneh, Yuval Ishai, Amit Sahai, and David J. Wu. Lattice-based SNARGs and their application to more efficient obfuscation. In *Proceedings of the 36th Annual International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT '17, pages 247–277, 2017.

[BISW18] Dan Boneh, Yuval Ishai, Amit Sahai, and David J. Wu. Quasi-optimal SNARGs via linear multi-prover interactive proofs. In *Proceedings of the 37th Annual International Conference on Theory and Application of Cryptographic Techniques*, EUROCRYPT '18, pages 222–255, 2018.

[BLNS20] Jonathan Bootle, Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. A non-PCP approach to succinct quantum-safe zero-knowledge. In *Proceedings of the 40th Annual International Cryptology Conference*, CRYPTO '20, pages 441–469, 2020.

[CF13] Dario Catalano and Dario Fiore. Vector commitments and their applications. In *Proceedings of the 16th International Conference on Practice and Theory in Public-Key Cryptography*, PKC '13, pages 55–72, 2013.

[CMS19] Alessandro Chiesa, Peter Manohar, and Nicholas Spooner. Succinct arguments in the quantum random oracle model. In *Proceedings of the 17th Theory of Cryptography Conference*, TCC '19, pages 1–29, 2019.

[DFMS19] Jelle Don, Serge Fehr, Christian Majenz, and Christian Schaffner. Security of the Fiat–Shamir transformation in the quantum random-oracle model. In *Proceedings of the 39th Annual International Cryptology Conference*, CRYPTO '19, pages 356–383, 2019.

[FGL+91] Uriel Feige, Shafi Goldwasser, László Lovász, Shmuel Safra, and Mario Szegedy. Approximating clique is almost NP-complete (preliminary version). In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science*, SFCS '91, pages 2–12, 1991.

[GMNO18] Rosario Gennaro, Michele Minelli, Anca Nitulescu, and Michele Orrù. Lattice-based zk-SNARKs from square span programs. In *Proceedings of the 25th ACM Conference on Computer and Communications Security*, CCS '18, pages 556–573, 2018.

[GW11] Craig Gentry and Daniel Wichs. Separating succinct non-interactive arguments from all falsifiable assumptions. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, STOC '11, pages 99–108, 2011.

[Jor75] Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société mathématique de France*, 3:103–174, 1875.

[Kil92] Joe Kilian. A note on efficient zero-knowledge proofs and arguments. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, STOC '92, pages 723–732, 1992.

[LZ19] Qipeng Liu and Mark Zhandry. Revisiting post-quantum Fiat–Shamir. In *Proceedings of the 39th Annual International Cryptology Conference*, CRYPTO '19, pages 326–355, 2019.

[MW05] Chris Marriott and John Watrous. Quantum Arthur–Merlin games. *Computational Complexity*, 14(2):122–152, 2005.

[Nao03] Moni Naor. On cryptographic assumptions and challenges. In *Proceedings of the 23rd Annual International Cryptology Conference*, CRYPTO '03, pages 96–109, 2003.

[Reg05] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, STOC '05, pages 84–93, 2005.

[Reg06] Oded Regev. Fast amplification of QMA (lecture notes), Spring 2006.

[Sho94] Peter W. Shor. Algorithms for quantum computation: Discrete logarithms and factoring. In *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '94, pages 124–134, 1994.

[Unr12] Dominique Unruh. Quantum proofs of knowledge. In *Proceedings of the 31st Annual International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT '12, pages 135–152, 2012.

[Unr16a] Dominique Unruh. Collapse-binding quantum commitments without random oracles. In *Proceedings of the 22nd International Conference on the Theory and Applications of Cryptology and Information Security*, ASIACRYPT '16, pages 166–195, 2016.

[Unr16b] Dominique Unruh. Computationally binding quantum commitments. In *Proceedings of the 35th Annual International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT '16, pages 497–527, 2016.

[VW16] Thomas Vidick and John Watrous. Quantum proofs. *Found. Trends Theor. Comput. Sci.*, 11(1-2):1–215, 2016.

[VZ21] Thomas Vidick and Tina Zhang. Classical proofs of quantum knowledge. arXiv quant-ph/2005.01691, 2021.

[Wat06] John Watrous. Zero-knowledge against quantum attacks. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, STOC '06, pages 296–305, 2006.

[Win99] Andreas J. Winter. Coding theorem and strong converse for quantum channels. *CoRR*, abs/1409.2536, 1999.

[Zha19] Mark Zhandry. Quantum lightning never strikes the same state twice. In *Proceedings of the 38th Annual International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT '19, pages 408–438, 2019.

[Zha20] Mark Zhandry. Schrödinger's pirate: How to trace a quantum decoder. In *Proceedings of the 18th Theory of Cryptography Conference*, TCC '20, pages 61–91, 2020.