Post-Quantum Succinct Arguments

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

March 15, 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).

At the heart of our proof is a new "measure-and-repair" quantum rewinding procedure that achieves asymptotically optimal knowledge error.

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

Contents

1 Introduction **3**
- 1.1 Our results . . . 3

2 Technical overview **5**
- 2.1 Kilian's protocol . . . 5
- 2.2 Our approach to post-quantum security of Kilian's protocol . . . 6
- 2.3 Prior quantum techniques . . . 7
- 2.4 A closer look at Unruh's lemma . . . 8
- 2.5 State recovery . . . 9
- 2.6 State repair . . . 12
- 2.7 Approximate state repair . . . 15
- 2.8 Sub-sampling and probabilistic special soundness . . . 17

3 Preliminaries **18**
- 3.1 Distributions and concentration inequalities . . . 18
- 3.2 Quantum preliminaries and notation . . . 19
- 3.3 Jordan's lemma . . . 20
- 3.4 Probabilistically checkable proofs . . . 21
- 3.5 Collapsing hash functions . . . 22
- 3.6 Interactive arguments . . . 23
- 3.7 Collapsing protocols . . . 24

4 Probabilistic special soundness **24**
- 4.1 Definition . . . 24
- 4.2 Examples . . . 26

5 Alternating projector algorithms **27**
- 5.1 The "approximate Jordan" unitary . . . 27
- 5.2 A "measure-and-repair" lemma . . . 29
- 5.3 Amplification . . . 31

6 Quantum extraction **32**
- 6.1 A quantum rewinding procedure . . . 32
- 6.2 Quantum rewinding lemma . . . 33
- 6.3 Proof of Theorem 6.1 . . . 35

7 Collapsing vector commitments **35**
- 7.1 Definition . . . 36
- 7.2 Merkle trees are collapsing . . . 37

8 Post-quantum security of Kilian's protocol **39**
- 8.1 Protocol description . . . 39
- 8.2 Kilian's protocol is probabilistically special sound . . . 40
- 8.3 Proof of Theorem 8.1 . . . 41

Acknowledgements **42**

References **42**

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. 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*. In turn, collapsing hash functions [Unr16b] are

[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.

implied by post-quantum lossy functions, which exist assuming the quantum hardness of Learning with Errors [Unr16a].

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

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 a *constant* number of responses [Unr16b, DFMS19].

Our extraction procedure enables rewinding not only Kilian's protocol, but any *collapsing protocol*. A collapsing protocol refers to any public-coin interactive argument with the guarantee 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.

Probabilistic special soundness. Rather than proving security of Kilian's protocol directly, we identify a concrete *classical* security property satisfied by a broad class of interactive arguments that includes Kilian's protocol, and apply our extraction procedure to prove that *any* collapsing protocol that satisfies this property is a quantum argument of knowledge.[3]

Many interactive arguments have a *special soundness* property. A public-coin interactive argument is k-special sound if an efficient extractor can output a witness when given any k accepting transcripts $(\tau, r_1, z_1), \ldots, (\tau, r_k, z_k)$ with common prefix τ, *distinct* verifier challenges r_i, and prover responses z_i. However, Kilian's protocol on a PCP with negligible soundness error is *not* k-special sound for any polynomial k. To extract a PCP that inherits the attacker's success probability, we need the additional restriction that the transcripts $(\tau, r_1, z_1), \ldots, (\tau, r_k, z_k)$ are generated by an efficient attacker subject to the condition that the r_i are sufficiently random.[4]

We call this notion *probabilistic special soundness*, and consider it to be of independent interest. For instance, the λ-fold parallel repetition of any k-special sound protocol is not k-special sound for any constant $k \geq 3$, but *is* probabilistically k-special sound. In fact, to our knowledge, all known interactive arguments of knowledge whose classical security relies on rewinding only the last round satisfy probabilistic k-special soundness for some $k = \text{poly}(\lambda)$.

Since Kilian's protocol is collapsing and probabilistically special sound, Theorem 1.1 is an immediate corollary of our general quantum extraction theorem:

Theorem 1.2 (Informal)**.** *Any collapsing protocol that is probabilistically special sound is a post-quantum argument of knowledge.*

Optimal knowledge error. Our theorem achieves asymptotically optimal knowledge error. As an additional application, we improve 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

[3] Kilian's protocol can be viewed as an argument of knowledge of a PCP proof accepted with probability above a certain threshold. This implies that Kilian's protocol is sound for any sound PCP, and moreover that if the PCP is a proof of knowledge for an NP relation $\mathfrak{R}$, then Kilian's protocol is an argument of knowledge for $\mathfrak{R}$.

[4] Concretely, the security reduction needs the attacker to pick the r_i from a $\text{poly}(\lambda)$-size challenge set $S \subseteq \{0,1\}^\lambda$ chosen at random, where λ is a security parameter. If the extractor is only required to output a witness with high probability over S, Kilian's protocol is in fact "special sound" when the number of transcripts k is (slightly larger than) the PCP length.

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 is no guarantee for $1/C \leq \varepsilon \leq 1/\sqrt{C}$. An immediate consequence of our techniques is that there is an extractor running in time $\mathrm{poly}(\lambda, 1/\varepsilon)$ that outputs a witness with probability $\Omega(\varepsilon)$ provided that $\varepsilon \geq (1+\delta)/C$ for any constant $\delta > 0$.

Discussion: is collapsing necessary? 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. Since we prove that post-quantum security of Kilian's protocol can be based on collapsing hash functions, a negative answer to this question would imply that there exist CRHFs that are not collapsing.

This question has been studied before, and in particular Zhandry [Zha19] proved that a CRHF that is not collapsing can be used to construct "quantum lightning," a strong cryptographic object with no known instantiation under well-studied assumptions.[5] Therefore, an immediate corollary of our result is that any post-quantum CRHF for which Kilian's protocol is not sound can be used to construct quantum lightning.

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 $z := (\pi[Q], \mathsf{pf})$ to V.[6]

Once the interaction is complete, V accepts if (1) pf is 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 $(h_{\mathsf{CRHF}}, \mathsf{cm}, r, z)$.

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

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

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.

If the PCP has alphabet Σ and proof length ℓ, one can show (see Section 8.2) 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, plugging in a PCP system with negligible soundness error yields an interactive argument with 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 enable 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.

3. **Probabilistic special sound collapsing protocols are arguments of knowledge.** For maximal generality, we introduce a new notion of *probabilistic special soundness*, a relaxation of special soundness that captures a broad class of interactive protocols such as Kilian. Using our new extraction procedure, we prove that any collapsing protocol satisfying probabilistic special soundness is a post-quantum argument of knowledge. Post-quantum soundness of Kilian's protocol follows immediately.

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 develop our extraction procedure over Sections 2.5 to 2.7. Finally, in Section 2.8, we define probabilistic special soundness and explain why this notion is compatible with our approach to quantum extraction.

2.3 Prior quantum techniques

We begin with a discussion of existing 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 $\tilde{P}(\mathsf{state}_\tau, \cdot)$ successfully 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 queries $r \leftarrow R$ 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. An interactive argument is a *collapsing protocol* [Unr16b, DFMS19, LZ19] if for any r, if an efficient prover outputs a superposition $|\phi\rangle$ of *accepting responses*,[7] i.e. $|\phi\rangle = \sum_z \alpha_z |z\rangle$ where each $|z\rangle$ in the superposition satisfies $V(\tau, r, z) = 1$, then it cannot distinguish between $|\phi\rangle$ and the state that results after measuring $|\phi\rangle$ in the computational basis.[8]

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 $O(\varepsilon^{2k-1})$.

[7] We write $|\phi\rangle$ as a pure state for clarity, but the collapsing property also holds for states entangled with another subsystem.

[8] 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.

This $O(\varepsilon^{2k-1})$ probability — which does not appear classically — 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, $k = \Omega(\ell)$ where ℓ is the PCP length. Thus, Unruh's lemma is insufficient since the guarantee only holds with probability $\varepsilon^{\Omega(\ell)}$.

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 $\{\mathsf{M}_r\}_{r\in R}$. We write binary-outcome projective measurements as $\mathsf{M}_r = (\Pi_r, \mathbf{I} - \Pi_r)$ where Π_r is associated with outcome 1, and $\mathbf{I} - \Pi_r$ with outcome 0.

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

> **Unruh's lemma**: For any state $|\psi\rangle$ and any collection of binary-outcome projective measurements $\{\mathsf{M}_r\}_{r\in R}$, if applying $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$ to $|\psi\rangle$ returns 1 with probability ε, then starting from $|\psi\rangle$ and applying $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$ 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 we define $\mathsf{M}_r = (\Pi_r, \mathbf{I} - \Pi_r)$ 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, M_r measures *whether* $\tilde{P}$ causes V to accept on challenge r. Therefore, the probability ε in Unruh's lemma (the probability $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$ applied to $|\psi\rangle$ returns 1) is the probability that $\tilde{P}(|\psi\rangle, \cdot)$ successfully answers a random challenge $r \leftarrow R$ in the interactive protocol. We will 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.[9] This is a strong requirement: even in the classical setting, k consecutive trials succeed with probability ε^k. Classically this can easily be resolved by repeating $N = k/\epsilon$ times 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 in total.

[9] 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.

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 $\mathsf{M}_r = (\Pi_r, \mathbf{I} - \Pi_r)$ 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 it turns out 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.

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.

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.

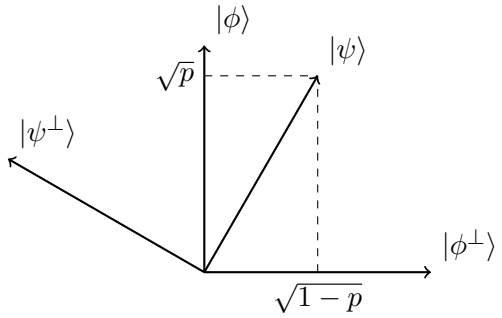

Figure 1: The quantum states $|\psi\rangle$ and $|\psi^\perp\rangle$ correspond to outcomes 1 and 0 of $\mathsf{Equals}_{|\psi\rangle} = (\,|\psi\rangle\langle\psi|\,, \mathbf{I} - |\psi\rangle\langle\psi|)$, respectively. The quantum states $|\phi\rangle$ and $|\phi^\perp\rangle$ correspond to outcomes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

Observe 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

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

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 technique for QMA amplification [MW05].[10] 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 $|\psi^\perp\rangle$.
- 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 $\mathsf{MW}(p)$ (for "Marriott-Watrous"):

1. Initialize $b_0 = 1$ (representing the fact that 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

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

[10] We remark that the goal of [MW05] was not to reconstruct a particular quantum state, but to estimate the probability p.

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.[11] 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 Π_v, Π_w induce a decomposition of the ambient Hilbert space into two-dimensional subspaces S_j such both Π_v and Π_w act as rank-one projectors within each subspace.[12]

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

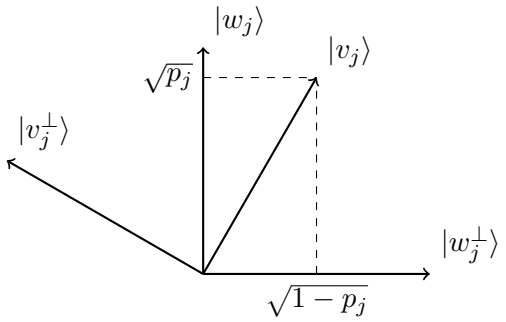

Figure 2: The states $|v_j\rangle$ and $|v_j^{\perp}\rangle$ correspond to 1 and 0 outcomes of $(\Pi_v, \mathbf{I} - \Pi_v)$, respectively; $|w_j\rangle$ and $|w_j^{\perp}\rangle$ correspond to 1 and 0 outcomes of $(\Pi_w, \mathbf{I} - \Pi_w)$, respectively.

By Jordan's lemma, a quantum state $|\phi\rangle$ satisfying $\Pi_w |\phi\rangle = |\phi\rangle$ can be written as

$$|\phi\rangle = \textstyle\sum_j \alpha_j |w_j\rangle \quad ,$$

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

To see why this distribution arises, consider the projective measurement $\mathsf{M}_{\mathrm{Jor}} = (\Pi_j^{\mathrm{Jor}})_j$ that projects onto the Jordan subspaces $\{S_j\}_j$ and returns j as the outcome, i.e., each Π_j^{Jor} is a projection onto the S_j subspace. Since $\mathsf{M}_{\mathrm{Jor}}$ acts as the identity within every Jordan subspace S_j, a

[11] 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.

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

consequence of Jordan's lemma is that $\mathsf{M}_{\mathrm{Jor}}$ commutes with both $(\Pi_v, \mathbf{I} - \Pi_v)$ and $(\Pi_w, \mathbf{I} - \Pi_w)$. 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 mind, our analysis of the "state recovery" procedure in the two-dimensional setting extends to higher dimensions by associating $(\Pi_v, \mathbf{I} - \Pi_v)$ with $\mathsf{Equals}_{|\psi\rangle}$ and $(\Pi_w, \mathbf{I} - \Pi_w)$ 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

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

To summarize, the takeaway of our discussion so far is the following lemma.

> **Setup:** Fix measurements $\mathsf{M}_v = (\Pi_v, \mathbf{I} - \Pi_v)$ and $\mathsf{M}_w = (\Pi_w, \mathbf{I} - \Pi_w)$ and a state $|\psi\rangle$ in the span of Π_v. Apply M_w to $|\psi\rangle$ and let $|\phi\rangle$ be the post-measurement state.
>
> **State recovery lemma:** Starting from $|\phi\rangle$, apply $\mathsf{M}_v, \mathsf{M}_w, \mathsf{M}_v, \mathsf{M}_w, \ldots$ until M_v returns 1. The procedure requires at most $2T - 1$ measurements with probability $1 - 1/T$.

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 $|\psi\rangle$. However, our goal is to efficiently extract successful attacker responses, which "only" requires that the probability M_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 adapt the "state recovery" procedure from the previous subsection to 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 will 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 will 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 a random M_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 $\mathsf{MixM}(\{\mathsf{M}_r\}_{r\in R})$, 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 $|\mathbb{1}_R\rangle$ 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 $|\mathbb{1}_R\rangle^{\mathcal{R}} \otimes |\psi\rangle$ 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.

$$E \coloneqq |\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}} \cdot \Pi_{\mathsf{CProj}} \cdot |\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}} \text{ where } |\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}} \text{ denotes } |\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}} \otimes \mathbf{I}^{\mathcal{H}} \ .$$

As desired, any eigenstate of E with positive eigenvalue p is of the form $|\mathbb{1}_R\rangle |\chi\rangle$ where $|\chi\rangle \in \mathcal{H}$ has success probability p:

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

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 $|\mathbb{1}_R\rangle |\psi\rangle$ of E with eigenvalue p. The precise relationship is summarized in the following observation:

> **Key fact:** For any state $|\psi\rangle$ with success probability p, $|\mathbb{1}_R\rangle |\psi\rangle$ can be written as a *linear combination of eigenstates of* E
>
> $$|\mathbb{1}_R\rangle |\psi\rangle = \textstyle\sum_j \alpha_j |\mathbb{1}_R\rangle |\chi_j\rangle$$
>
> where each $|\mathbb{1}_R\rangle |\chi_j\rangle$ 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 $\mathsf{Test}_\varepsilon \coloneqq (\Pi_\varepsilon, \mathbf{I} - \Pi_\varepsilon)$. 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 $|\mathbb{1}_R\rangle |\psi\rangle$ where $|\psi\rangle$ has success probability 2ε is a linear combination of eigenstates $\sum_j \alpha_j |\mathbb{1}_R\rangle |\chi_j\rangle$ 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 $|\mathbb{1}_R\rangle |\chi_j\rangle$ 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 $|\mathbb{1}_R\rangle |\psi\rangle \in (\mathcal{R}, \mathcal{H})$ 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) Measure $\mathcal{R}$ in the computational basis to obtain r, and then apply M_r to $\mathcal{H}$ to obtain an outcome b. Call this step "successful" if $b = 1$.
   (b) (Repair step) Using measurement outcome b, define

$$\mathsf{M}_{r,b} \coloneqq \begin{cases} |r\rangle\langle r| \otimes \Pi_r & \text{if } b = 1 \\ |r\rangle\langle r| \otimes (I - \Pi_r) & \text{if } b = 0 \end{cases}.$$

   Repair the state by applying $\mathsf{Test}_\varepsilon, \mathsf{M}_{r,b}, \mathsf{Test}_\varepsilon, \mathsf{M}_{r,b}, \ldots$ 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 $|\mathbb{1}_R\rangle |\chi\rangle$ where $|\chi\rangle$ has success probability ε, the Measure step is equivalent to an application of $\mathsf{MixM}(\{\mathsf{M}_r\}_r)$ that succeeds with at least ε probability.

We are left to show that the Repair step does not run for too long. This requires a more sophisticated "state repair" lemma than the one we argued previously, since by introducing the ancilla registers $\mathcal{R}$, the total disturbance caused by the Measure step corresponds to a $2|R|$-outcome measurement on $(\mathcal{R}, \mathcal{H})$ that returns (r, b). Previously, we had argued that state repair is possible after disturbance from a *binary-outcome* measurement. It turns out that we can strengthen our previous lemma to handle this case, but we obtain a failure probability that depends on the number of possible outcomes (r, b). We state this stronger lemma below.

> **Setup:** Fix a binary measurement $\mathsf{B} = (\Pi, \mathbf{I} - \Pi)$ and an N-outcome measurement $(\Pi_i)_{i\in[N]}$, along with a state $|\psi\rangle$ in the span of Π. Apply $(\Pi_i)_{i\in[N]}$ to $|\psi\rangle$ to obtain outcome $i \in [N]$, and let $|\phi\rangle$ be the post-measurement state.
>
> **Generalized state repair lemma:** Define the *binary* measurement $\mathsf{M}_i = (\Pi_i, \mathbf{I} - \Pi_i)$. Starting from $|\phi\rangle$, apply $\mathsf{B}, \mathsf{M}_i, \mathsf{B}, \mathsf{M}_i, \ldots$ until B returns 1. The procedure terminates after at most $2T - 1$ measurements with probability at least $1 - N/T$.

Similar to the binary case, this lemma can be proved (with slightly more effort) by appealing to the Jordan subspace decomposition for Π and Π_i.

Recap. To summarize our progress so far, we briefly discuss what our state repair procedure means 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 $c \in \mathbb{N}$ and parameter $0 < \delta < 1$, 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
- with probability $1 - \delta$, the total number of measurements we perform is at most $2c^2|R|/\delta$.

Accounting for the initialization step, this procedure has total success probability $\varepsilon - 1/\mathrm{poly}(\lambda)$. Moreover, if $|R|$, c, and $1/\delta$ are all polynomially bounded, the total number of measurements performed is $\mathrm{poly}(\lambda)$.

While this is promising, we are far from done.

1. We do not know of a way to efficiently implement $\mathsf{Test}_\varepsilon$. Hence, in Section 2.7, we will show how to replace the $\mathsf{Test}_\varepsilon$ measurement 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. The assumption that R has size $\mathrm{poly}(\lambda)$ appears to limit this technique to protocols with inverse polynomial soundness. For Kilian's protocol in particular, we would hope to achieve soundness error $\mathrm{negl}(\lambda)$ when the underlying PCP has soundness error $\mathrm{negl}(\lambda)$. We handle this issue (classically!) for a broad class of protocols, including Kilian, in Section 2.8.

2.7 Approximate state repair

Approximating $\mathsf{Test}_\varepsilon$. While we do not know of a way to implement $\mathsf{Test}_\varepsilon$, it turns out that 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 $|\mathbb{1}_R\rangle|\chi_j\rangle$ be an eigenstate of $E = |\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}} \Pi_{\mathsf{CProj}} |\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}}$ 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 $|\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}}$: any state in the span of $|\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}}$ that is in the Jordan subspace S_j must be an eigenstate $|\mathbb{1}_R\rangle|\chi_j\rangle$ of E with eigenvalue p_j.

Then, by the analysis in Section 2.5, if we start from $|\mathbb{1}_R\rangle|\chi_j\rangle$ and apply the binary projective measurements $\mathsf{CProj} = (\Pi_{\mathsf{CProj}}, \mathbf{I} - \Pi_{\mathsf{CProj}})$ and $\mathsf{M}_{|\mathbb{1}_R\rangle} = (|\mathbb{1}_R\rangle\langle\mathbb{1}_R|, \mathbf{I} - |\mathbb{1}_R\rangle\langle\mathbb{1}_R|)$ in an alternating fashion:

$$\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}, \mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}, \ldots,$$

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: $|\mathbb{1}_R\rangle|\psi\rangle$ for state $|\psi\rangle$ with success probability at least 2ε.

1. Apply $2t$ measurements $\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}, \ldots, \mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}$. 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 $|\mathbb{1}_R\rangle|\psi\rangle$, we employ the method from Section 2.5 of projecting onto the Jordan subspaces $\{S_j\}_j$ for the projectors Π_{CProj} and $|\mathbb{1}_R\rangle\langle\mathbb{1}_R|$. Since any state $|\mathbb{1}_R\rangle|\psi\rangle$ can be written as a linear combination $\sum_j \alpha_j |\mathbb{1}_R\rangle|\chi_j\rangle$ of eigenstates of E, the result of applying $\mathsf{ApproxTest}_{\varepsilon,t}$ to $|\mathbb{1}_R\rangle|\psi\rangle$ 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.

Technical challenge: $\mathsf{ApproxTest}_{\varepsilon,t}$ is not projective. In Section 2.6 we claimed that if a state $|\psi\rangle$ initially in the span of some projector Π is disturbed by an N-outcome measurement $(\Pi_i)_{i\in[N]}$, then by performing alternating measurements, we can return our state to the span of Π in $2T$ measurements except with probability N/T. It is not clear that such a statement holds if $(\Pi, \mathbf{I}-\Pi)$ is replaced by a *non-projective* measurement.

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

$$\mathsf{ApproxTest}_{\varepsilon,t}, \mathsf{M}_{r,b}, \mathsf{ApproxTest}_{\varepsilon,t}, \mathsf{M}_{r,b}, \ldots$$

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

$$\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}, \mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}, \ldots \ .$$

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 $\mathsf{Test}_\varepsilon, \mathsf{M}_{r,b}, \mathsf{Test}_\varepsilon, \mathsf{M}_{r,b}, \ldots$ 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 $|\mathbb{1}_R\rangle|\chi_j\rangle$ 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 $\mathsf{M}_{|\mathbb{1}_R\rangle}$, 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 $\mathsf{M}_{|\mathbb{1}_R\rangle}$ 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 $\mathsf{M}_{r,b}$ (which originally acts as identity on the $\mathcal{L}$ registers) to additionally project $\mathcal{L}$ onto $|0^{2t}\rangle$. This modified measurement $\mathsf{M}'_{r,b}$ returns 1 if and only if the original $\mathsf{M}_{r,b}$ returns 1 *and* the binary projective measurement of $\mathcal{L}$ onto $|0^{2t}\rangle$ returns 1; notice that $\mathsf{M}'_{r,b}$ is still a binary projective measurement.

For this modification to work, we must change our stopping condition. We cannot simply consider our state to be "repaired" the moment $\mathsf{ApproxTest}_{\varepsilon,t}$ first returns 1, since this measurement outcome may be meaningless if the preceding $\mathsf{M}'_{r,b}$ measurement returned 0 (as this indicates the registers $\mathcal{L}$ may have been non-zero when $\mathsf{ApproxTest}_{\varepsilon,t}$ was applied). Thus, we have to use a more restrictive stopping condition: the state is only repaired when *consecutive* $\mathsf{M}'_{r,b}$ and $\mathsf{ApproxTest}_{\varepsilon,t}$ measurements return 1. We therefore have to extend our previous state repair lemma, and for that extension the probability that the procedure takes more than T steps is $N/\sqrt{T}$.

Technical challenge: handling the approximation. We must address the fact that $\mathsf{ApproxTest}_{\varepsilon,t}$ returns answers that are only "approximately correct" (even after we ensure the ancilla registers $\mathcal{L}$ are $|0^{2t}\rangle$). Even if a state is entirely concentrated on Jordan subspaces with eigenvalue p smaller than our threshold ε, there is still a reasonable probability that performing $2t$ Bernoulli trials will produce an estimate larger than ε, and $\mathsf{ApproxTest}_{\varepsilon,t}$ would output 1 in this case.

Resolution: gradually reduce the ε threshold. By setting t appropriately large, we can ensure for any desired "slack" parameter γ that the probability $\mathsf{ApproxTest}_{\varepsilon,t}$ returns 1 on an eigenstate with eigenvalue $\varepsilon - \gamma$ is negligible. As a result, however, we no longer have the guarantee that the state has success probability ε, but rather $\varepsilon - \gamma$. This means that the next time we want to use $\mathsf{ApproxTest}_{\varepsilon,t}$, we actually have to use $\mathsf{ApproxTest}_{\varepsilon-\gamma,t}$, and with each subsequent measure-and-repair loop we need to subtract another γ. Fortunately this suffices: for any desired number c of measurement-and-repair loops, we can set $\gamma = \varepsilon/c$ so that the threshold never drops below $\varepsilon/2$.

This concludes the discussion of the main technical difficulties, and we now have an *efficient* extraction procedure applicable to any collapsing protocol with a $\mathrm{poly}(\lambda)$-size challenge space. Next we explain how to apply this procedure to protocols with larger challenge spaces.

2.8 Sub-sampling and probabilistic special soundness

We employ a generic classical "sub-sampling" strategy to reduce the challenge space size. Given a malicious prover $\tilde{P}(|\psi\rangle, \cdot)$ with success probability ε in a protocol with super-polynomial challenge space R, we sample a random set $S \subset R$ of size $|S| = \mathrm{poly}(\lambda, 1/\varepsilon)$. The sub-sampling step turns an attacker with success probability ε on a random challenge $r \leftarrow R$ into an attacker with success probability roughly ε on a random challenge $r \leftarrow S$; this can be formalized with standard concentration bounds. We then perform our extraction procedure over S rather than R.

The sub-sampling guarantee. Precisely stating the guarantee of the sub-sampling step requires some care. Given an attacker that succeeds with probability ε, if we sample a random set of challenges S, our extraction theorem only guarantees that the the attacker answers a (roughly) ε fraction of the challenges in S — in particular, this does not rule out adversaries that can view S and then *adaptively* decide which $\varepsilon \cdot |S|$ challenges to answer correctly on.

Such a "strategy" is impossible in the classical setting, since each extraction attempt is an independent instance. In our quantum extraction procedure, however, in the very first step we

(approximately) project the adversary on to certain eigenspaces of an operator that depends on the whole of S. At this point the adversary's behavior may depend on S in ways we do not understand.

At first glance, this could be a problem: consider Kilian's protocol instantiated with a $\mathrm{negl}(\lambda)$-soundness PCP. The security reduction fails completely if we can only extract $\mathrm{poly}(\lambda)$ accepting transcripts for randomnesses chosen arbitrarily from R, since the union of the corresponding query sets may only cover a tiny fraction of the PCP.

However, the sub-sampling procedure allows us to provide a stronger guarantee than this. Even if the attacker "adaptively" chooses the challenges r on which it answers correctly, it must still answer an ε fraction of the challenges in the randomly sampled set S. We show that for Kilian's protocol applied to a PCP with negligible soundness error, the probability that an adversary can do this is negligible over the choice of S, *even when the adversary knows* S.

Probabilistic special soundness. That the randomness of S is enough to show soundness for Kilian's protocol does not appear to be an accident. In a typical classical security proof for an interactive argument, the extractor repeatedly rewinds the malicious prover, eventually collecting a number of accepting transcripts for different challenges. These transcripts must be enough to extract a witness (or otherwise derive a contradiction to soundness). In typical security reductions, this is argued by choosing challenges uniformly at random, which implies that the prover must answer a "large enough" fraction of those challenges correctly.

We define a notion called "probabilistic special soundness" which aims to explicitly capture this type of reduction. This is a relaxation of k-special soundness where the extraction guarantee must hold only when the transcripts $\{(r_i, z_i)\}_{i=1}^{k}$ have $r_i \in S$ for a randomly polynomial-size set $S \subseteq R$ provided to the attacker. It turns out this notion is satisfied by a wide variety of interactive protocols including Kilian's protocol, any (standard) special sound protocol, and any λ-fold parallel repetition of a special sound protocol; see Section 4 for further details.

We note that in the classical setting, *any* interactive argument that is probabilistically $\mathrm{poly}(\lambda)$-special sound is an argument of knowledge via the proof strategy outlined above: repeatedly rewind the attacker on random challenges until it produces enough valid transcripts to extract a witness. We therefore find consider notion interesting even outside the context of post-quantum security.

3 Preliminaries

Notation. 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 some 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 Distributions and concentration inequalities

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

Proposition 3.1 (Additive Chernoff bound)**.** *For $\delta, \epsilon > 0$, define $n_{\epsilon,\delta} \coloneqq \log(1/2\delta)/2\epsilon^2$. For $X \sim \mathsf{Bin}(n,p)$, if $n \geq n_{\epsilon,\delta}$, then*

$$\Pr[p - \epsilon \leq X/n \leq p + \epsilon] \geq 1 - \delta \ .$$

Proposition 3.2 (Multiplicative Chernoff bound without replacement)**.** *Let $x_1, \ldots, x_N \in \{0,1\}$. Let $S \subseteq [N]$ be uniformly random of size K. Then*

$$\Pr\left[\sum_{i \in S} x_i \geq (1+\delta)\mu\right] \leq e^{-\delta^2 \mu/3} \ ,$$

where $\mu \coloneqq \frac{K}{N} \sum_{i=1}^{N} x_i$.

We will also use a concentration inequality due to Hoeffding [Hoe63] for sampling without replacement from a finite set of probabilities.

Proposition 3.3 (Hoeffding's inequality)**.** *Let $z_1, \ldots, z_N \in [0,1]$. Let $S \subseteq [N]$ be uniformly random of size K. Then for all $\varepsilon > 0$,*

$$\Pr\left[\mu - \frac{1}{K}\sum_{i \in S} z_i > \varepsilon\right] \leq e^{-2K\varepsilon^2}$$

where $\mu \coloneqq \frac{1}{N} \sum_{i=1}^{N} z_i$.

3.2 Quantum preliminaries and notation

A (pure) *quantum state* is a vector $|\psi\rangle$ in a complex Hilbert space $\mathcal{H}$ with $\||\psi\rangle\| = 1$; in this work, $\mathcal{H}$ will always be finite-dimensional. A *density matrix* is a Hermitian operator $\rho \in S(\mathcal{H})$, the space of Hermitian operators on $\mathcal{H}$, with $\mathrm{Tr}(\rho) = 1$. A density matrix represents a probabilistic mixture of pure states (a mixed state). The density matrix corresponding to the pure state $|\psi\rangle$ is $|\psi\rangle\langle\psi|$.

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 ρ to the density matrix $U\rho U^\dagger$.

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 $\sum_i \Pi_i = \mathbf{I}$. This implies that $\Pi_i \Pi_j = 0$ for $i \neq j$. 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 $|\psi_i\rangle = \Pi_i |\psi\rangle / \sqrt{p_i}$.

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 $\mathbf{I} - \Pi$ is associated with the outcome 0.

General (non-unitary) evolution of a quantum state can be represented via a *completely-positive trace-preserving (CPTP)* map $T \colon S(\mathcal{H}) \to S(\mathcal{H})$. For example, we can model the application of a projective measurement $\mathsf{P} = (\Pi_i)_{i \in S}$ to a mixed state via the CPTP map $\rho \mapsto \sum_i |i\rangle\langle i| \otimes \Pi_i \rho \Pi_i^\dagger$, which introduces a classical ancilla to record the measurement outcome. We omit the precise definition of these maps in this work; we will only use that they are trace-preserving (for all $\rho \in S(\mathcal{H})$, $\mathrm{Tr}(T(\rho)) = \mathrm{Tr}(\rho)$) and linear.

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 $\lambda > \lambda_0$, $|\mathsf{Adv}_\lambda| \leq p(\lambda)$.

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_v and Π_w on a Hilbert space $\mathcal{H}$, there exists an orthogonal decomposition of $\mathcal{H}$ into one-dimensional and two-dimensional subspaces (the* Jordan subspaces*) that are invariant under both Π_v and Π_w. Moreover:*

- *in each one-dimensional space, Π_v and Π_w act as identity or rank-zero projectors; and*
- *in each two-dimensional subspace S_j, Π_v and Π_w are rank-one projectors: there exist $|v_j\rangle, |w_j\rangle \in S_j$ such that Π_v projects onto $|v_j\rangle$ and Π_w projects onto $|w_j\rangle$.*

Proof. The proof follows by considering the eigenvectors of the matrix $\Pi_v + \Pi_w$, which span $\mathcal{H}$ since $\Pi_v + \Pi_w$ is Hermitian. Let $|\psi\rangle$ be an eigenvector with eigenvalue p (i.e., $\Pi_v |\psi\rangle + \Pi_w |\psi\rangle = p |\psi\rangle$). There are two cases to consider.

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

If $\Pi_v |\psi\rangle$ does not lie in $\mathrm{span}(|\psi\rangle)$, then $\mathrm{span}(|\psi\rangle, \Pi_v |\psi\rangle)$ is a two-dimensional subspace. This subspace is invariant under Π_v, which acts as a projector onto $|v_j\rangle \coloneqq \Pi_v |\psi\rangle$. Moreover, this subspace can be written as $\mathrm{span}(|\psi\rangle, \Pi_w |\psi\rangle)$, and by an identical argument, Π_w projects this subspace onto $|w_j\rangle \coloneqq \Pi_w |\psi\rangle$. □

For each two-dimensional subspace S_j, we call $p_j \coloneqq |\langle v_j | w_j \rangle|^2$ the *eigenvalue* of the j-th subspace. We choose phases for $|v_j\rangle$ and $|w_j\rangle$ such that

$$|w_j\rangle = \sqrt{p_j}\,|v_j\rangle + \sqrt{1-p_j}\,|v_j^\perp\rangle \quad ,$$
$$|v_j\rangle = \sqrt{p_j}\,|w_j\rangle + \sqrt{1-p_j}\,|w_j^\perp\rangle \quad ,$$

where $|v_j^\perp\rangle \in S_j$ is orthogonal to $|v_j\rangle$ and $|w_j^\perp\rangle \in S_j$ is orthogonal to $|w_j\rangle$. In particular, it holds that $\Pi_v |w_j\rangle = \sqrt{p_i}\,|v_j\rangle$ and $\Pi_w |v_j\rangle = \sqrt{p_i}\,|w_j\rangle$. Thus applying the measurement $(\Pi_v, \mathbf{I} - \Pi_v)$ to the state $|w_j\rangle$ yields outcome 1 and post-measurement state $|v_j\rangle$ with probability p_i, and outcome 0 and post-measurement state $|v_i^\perp\rangle$ with probability $1 - p_i$. A symmetric statement holds for applying $(\Pi_w, \mathbf{I} - \Pi_w)$ to the state $|v_j\rangle$.

We treat one-dimensional subspaces as a "degenerate" version of the two-dimensional case. If Π_v acts as the identity on the j-th subspace then we label the vector spanning the subspace $|v_j\rangle$; if Π_v is the zero projection on this subspace then we label the vector $|v_j^\perp\rangle$. We use a similar convention for Π_w (so the vector spanning a one-dimensional subspace has two labels). We set $p_j \coloneqq 1$ if both Π_v and Π_w 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.

We define a measurement projecting on to the Jordan subspaces that, while it cannot be efficiently realized in general, will be a useful analysis tool.

Definition 3.5. For Hermitian projectors Π_v and Π_w, the *Jordan subspace measurement* is the measurement $\mathsf{M}_{\mathrm{Jor}}[\Pi_v, \Pi_w] \coloneqq (\Pi_j^{\mathrm{Jor}})_j$, where

$$\Pi_j^{\mathrm{Jor}} \coloneqq |v_j\rangle\langle v_j| + |v_j^\perp\rangle\langle v_j^\perp| = |w_j\rangle\langle w_j| + |w_j^\perp\rangle\langle w_j^\perp|$$

for vectors $|v_j\rangle, |v_j^\perp\rangle, |w_j\rangle, |w_j^\perp\rangle$ as in Lemma 3.4.

We also define projections on to Jordan subspaces with certain ranges of eigenvalues:

- $\Pi^{\mathrm{Jor}}_{p\pm\epsilon} \coloneqq \sum_{j,p_j\in[p\pm\epsilon]} \Pi^{\mathrm{Jor}}_j$.
- $\Pi^{\mathrm{Jor}}_{\geq p} \coloneqq \sum_{j,p_j\geq p} \Pi^{\mathrm{Jor}}_j$.

Since the Jordan subspaces are invariant under both Π_v and Π_w, the following useful proposition holds for the Jordan subspace measurement.

Proposition 3.6. *Let* $(\Pi^{\mathrm{Jor}}_j)_j \coloneqq \mathsf{M}_{\mathrm{Jor}}[\Pi_v, \Pi_w]$. *For all* j, Π^{Jor}_j *commutes with both* Π_v *and* Π_w.

Finally, we give a simple fact about states concentrated around Jordan subspaces of value at least p, which are eigenstates of one of the two projectors.

Claim 3.7. *Suppose that* ρ *is such that* $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p}\rho\right) \geq 1-\beta$, *where* $\Pi^{\mathrm{Jor}}_{\geq p} \coloneqq \Pi^{\mathrm{Jor}}_{\geq p}[\Pi_v, \Pi_w]$. *Then if* $\mathrm{Tr}(\Pi_v\rho) = 1$, $\mathrm{Tr}(\Pi_w\rho) \geq p-\beta$; *the same implication holds with* v *and* w *switched.*

Proof. Write $\rho = \sum_i q_i |\phi_i\rangle\langle\phi_i|$, with $|\phi_i\rangle = \sum_j \alpha_j^{(i)} |v_j\rangle$. Then

$$\mathrm{Tr}(\Pi_w\rho) = \sum_i q_i \sum_j |\alpha_j^{(i)}|^2 p_j \geq p \sum_i q_i \sum_{j,p_j\geq p} |\alpha_j^{(i)}|^2 \geq p(1-\beta) \ ,$$

which completes the proof. The other implication is symmetrical. □

3.4 Probabilistically checkable proofs

A *probabilistically checkable proof* (PCP) for an NP 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 $\Pr[\mathbf{V}^\pi_{\mathsf{PCP}}(x) = 1] = 1$.
- **Soundness.** For every instance $x \notin \mathcal{L}(\mathfrak{R})$ and proof string $\pi\colon [\ell] \to \Sigma$, $\Pr[\mathbf{V}^\pi_{\mathsf{PCP}}(x) = 1] \leq \varepsilon_{\mathsf{PCP}}$.

The quantities $\varepsilon_{\mathsf{PCP}}, \ell, \Sigma$ can be functions of the instance size $|x|$. 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 $\mathbf{V}_{\mathsf{PCP}}$. (Both can be functions of $|x|$.)

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 $(x, w) \in \mathfrak{R}$.

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,

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

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 Interactive arguments

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

Interactive quantum circuits. A m-round interactive quantum circuit is a sequence of unitary quantum circuits $U^{(1)}, \ldots, U^{(m)}$, where $U^{(i)} \colon R_i \to U(\mathcal{I}, \mathcal{Z}_i)$ is a classical circuit mapping each challenge to (a classical description of) a unitary circuit. 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 the following game:

1. Initialize the register $\mathcal{I}$ to $|\psi\rangle$, and $\tau \coloneqq ()$.
2. For $i = 1, \ldots, m$,
   (a) Sample $r_i \leftarrow R_i$.
   (b) Apply unitary $U^{(i)}(r_i)$ to $(\mathcal{I}, \mathcal{Z}_i)$.
   (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 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. This 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. The *size* of an interactive quantum circuit is the sum of the sizes of the (classical) circuits implementing $U^{(1)}, \ldots, U^{(m)}$.

Definition 3.11. A (post-quantum) *interactive argument* for a relation $\mathfrak{R}$ with soundness s is a pair of interactive classical polynomial-time algorithms (P, V) such that the following holds.

Completeness. For all $(x, w) \in \mathfrak{R}$,

$$\Pr[\langle P(x, w), V(x)\rangle = 1] = 1 \ .$$

Soundness. For all $x \notin \mathcal{L}(\mathfrak{R})$, and all polynomial-size interactive quantum circuits $\tilde{P}$,

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

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

We will also consider interactive arguments that satisfy the stronger property of *knowledge soundness*. Below we write $\mathsf{Ext}^{\tilde{P}}$ for an extractor with "black-box" access to $\tilde{P} = (U^{(1)}, \ldots, U^{(m)})$; this means that Ext is a quantum circuit with special gates corresponding to $U^{(i)}(r)$ and $(U^{(i)}(r))^\dagger$ for $i \in [m]$.

Definition 3.12. An interactive argument satisfies (post-quantum) *knowledge soundness* with knowledge error κ if there exists a quantum extractor Ext such that for all polynomial-size quantum adversaries $\tilde{P}$, states $|\psi\rangle$ and instances x, and all $\varepsilon \leq \Pr\Big[\langle \tilde{P}(|\psi\rangle), V(x)\rangle = 1\Big]$, $\mathsf{Ext}^{\tilde{P}}(x, \varepsilon)$ has size $\mathrm{poly}(\lambda, 1/\varepsilon)$ and if $\varepsilon \geq \kappa(\lambda)$ then

$$\Pr\Big[(x, w) \in \mathfrak{R} \mid w \leftarrow \mathsf{Ext}^{\tilde{P}}(x, \varepsilon)\Big] = \Omega(\varepsilon) - \mathrm{negl}(\lambda) \ .$$

Remark 3.13. *An unusual feature of our definition is that we provide the extractor with a lower bound on the success probability of the adversary on x. This arises from a technical requirement in our security proof.*

If the extractor has access to arbitrarily many copies of the adversary's initial *state $|\psi\rangle$, we can generically remove this requirement by running* Ext *repeatedly for decreasing values of ε. This is possible, for example, if the adversary's initial state is classical, or can otherwise be generated efficiently.*

3.7 Collapsing protocols

We define the following experiment $\mathsf{CollapseExp}(b, \tilde{P}, \mathsf{Adv})$.

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.

Definition 3.14. We say that a protocol is collapsing if for every polynomial-size interactive quantum adversary $\tilde{P}$ and polynomial-size quantum distinguisher Adv,

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

4 Probabilistic special soundness

In this section we define a new notion we call *probabilistic special soundness*, a relaxation of the standard special soundness notion. Importantly, while Kilian's protocol is not known to be special sound, we will show in Section 8 that it *is* probabilistically special sound. Looking ahead to Section 6, we will show in Theorem 6.1 that any collapsing interactive protocol (Definition 3.14) satisfying probabilistic special soundness is a post-quantum argument of knowledge. Since probabilistic special soundness applies to a broad class of protocols, Theorem 6.1 will apply not just to Kilian's protocol, but a wide class of interactive protocols (e.g., special sound collapsing protocols).

4.1 Definition

To provide context for our new definition, we first recall the definition of (standard) special soundness.

Definition 4.1 (k-special soundness)**.** An interactive argument (P, V) satisfies k-special soundness for a relation $\mathfrak{R}$ if for any instance x, given any k protocol transcripts $(\tau, r_1, z_1), \ldots, (\tau, r_k, z_k)$ with shared prefix τ where (1) each r_i is distinct and (2) for all i, $V(\tau, r_i, z_i) = 1$ an efficient extractor E can output a witness w such that $\mathfrak{R}(x, w) = 1$.

Probabilistic special soundness relaxes the definition of special soundness in two ways. Specifically, we only require that the extractor outputs a witness with high probability when:

1. the transcripts are generated by a computationally bounded adversary, and
2. the challenges r_i are contained within a set S of polynomial size chosen uniformly at random.

Concretely, we introduce an extraction experiment for an interactive argument $\mathsf{ARG} = (P, V)$ parameterized by a natural number k, an interactive (quantum) adversary Adv, and a classical extractor algorithm Ext. Let $R = R_m$ denote the space of challenges used in the last round of ARG. Let $\mathsf{Sampler}(R)$ be an oracle which, when queried with an integer $N \leq |R|$, samples and outputs a uniformly random subset $S \subseteq R$ of size N. (We force the adversary to read the response it gets from $\mathsf{Sampler}$, and so if Adv queries $\mathsf{Sampler}$ with N, its running time is at least N.) The extraction experiment PSSExp is defined as follows:

$\mathsf{PSSExp}(k, \mathsf{Adv}, \mathsf{Ext})$:

1. Run $\langle \mathsf{Adv}, V \rangle$ until just before the last verifier message. Let τ be the resulting partial transcript, and let x be the corresponding instance.

2. Adv is now given oracle access to $\mathsf{Sampler}(R)$, and generates k continuations

$$(r_1, z_1), \ldots, (r_k, z_k) \leftarrow \mathsf{Adv}^{\mathsf{Sampler}(R)}.$$

   Let $S \subseteq R$ denote the set of challenges sampled by $\mathsf{Sampler}$ across all of Adv's queries.

3. The extractor Ext is given the transcripts and extracts a witness

$$w \leftarrow E(\tau, (r_1, z_1), \ldots, (r_k, z_k)).$$

4. The output of the experiment is 1 if and only if all the following are satisfied:

   - each r_i is unique and $r_i \in S$
   - each transcript is accepting, i.e., $V(\tau, r_i, z_i) = 1$ for each i
   - w is *not* a valid witness, i.e., $\mathfrak{R}(x, w) = 0$.

Definition 4.2 (probabilistic k-special soundness)**.** An interactive argument is *probabilistically k-special sound* for a relation $\mathfrak{R}$ if there exists an efficient classical extractor Ext such that for any efficient quantum polynomial-time Adv,

$$\Pr[\mathsf{PSSExp}(k(\lambda), \mathsf{Adv}, \mathsf{Ext}) = 1] \leq \mathrm{negl}(\lambda).$$

For comparison, we re-state the standard k-special soundness definition as follows.

Definition 4.3 (k-special soundness, alternative formulation)**.** An interactive argument is *k-special sound* for a relation $\mathfrak{R}$ if there exists an efficient classical extractor Ext such that for any *unbounded time* Adv,

$$\Pr[\mathsf{PSSExp}(k(\lambda), \mathsf{Adv}, \mathsf{Ext}) = 1] = 0.$$

The reason why this alternative formulation is equivalent to the standard formulation of k-special soundness (Definition 4.1) is that an unbounded time Adv can ask $\mathsf{Sampler}(R)$ for $|R|$ challenges, at which point it is free to choose the challenges r_i arbitrarily as in the standard definition.

We can drop the explicit k and say that a protocol is probabilistically special sound if there exists a $k = \text{poly}(\lambda)$ such that the protocol is probabilistically k-special sound.

Claim 4.4. *Any protocol that is probabilistically special sound is a classical argument of knowledge.*

Proof. Let $\tilde{P}$ be an adversary against the argument system. We define an adversary for the PSS experiment as follows:

1. Run $\tilde{P}$ until just before the last verifier message, interacting with the challenger, to obtain a transcript τ.
2. Query $\mathsf{Sampler}(R)$ to obtain a set $S \subseteq R$ of size $\max(\min(2\lambda(1/\varepsilon)^2, |R|), 4k/\varepsilon)$.
3. Run $\tilde{P}$ on every $r \in S$; let $(r_1, z_1), \ldots, (r_t, z_t)$ be the valid continuations obtained.
4. If $t \geq k$, output $(r_1, z_1), \ldots, (r_k, z_k)$; otherwise output $\perp$.

Suppose $\tilde{P}$ succeeds with inverse polynomial probability $\varepsilon(\lambda)$. Then with probability at least $\varepsilon/2$, τ is such that the cheating prover answers a random challenge $r \in R$ with probability $\varepsilon/2$. By Hoeffding's inequality, with probability $1 - e^{-\lambda}$, $\tilde{P}$ answers a random challenge $r \in S$ with probability at least $\varepsilon/4$. Hence with all but negligible probability, $t \geq k$, and so by the PSS guarantee the extractor outputs a valid witness with all but negligible probability. □

4.2 Examples

Probabilistic special soundness is a versatile definition that captures not just Kilian's protocol, but a large class of interactive arguments. In this section, we highlight several other classes of probabilistically special sound interactive protocols.

Claim 4.5 (Trivial)**.** *Any protocol satisfying k-special soundness is probabilistically k-special sound.*

Proof. This is immediate since an extractor that can produce a witness given an arbitrary collection of valid transcripts $(\tau, r_1, z_1), \ldots, (\tau, r_k, z_k)$ can of course produce a witness when the challenges r_i must be sampled randomly by a challenger. □

Claim 4.6. *If an interactive argument is probabilistically k-special sound for a relation $\mathfrak{R}$ for some constant k, then its λ-fold parallel repetition is also probabilistically k-special sound for $\mathfrak{R}$.*

Proof. Without loss of generality, we assume the challenge set to of the base interactive protocol to be at least size k; otherwise k can be re-defined to be the same size as the challenge set without sacrificing the k-special soundness guarantee.

We write a challenge $\vec{r}$ in the λ-fold parallel repetition as $\vec{r} = (r^{(1)}, \ldots, r^{(\lambda)})$ where $r^{(q)}$ is the challenge corresponding to the qth parallel repetition. For any size-k subset $K \subset S$, say that $\{\vec{r}_i\}_{i \in K}$ is *extractable* if there exists an index $q \in [\lambda]$ such that all elements in $\{r_i^{(q)}\}_{i \in [K]}$ are distinct. For any fixed $q \in [\lambda]$, the probability that $\{r_i^{(q)}\}_{i \in [K]}$ does not contain k distinct elements is at most $1 - k!/k^k$. Since each repetition is independent, the probability that $\{\vec{r}_i\}_{i \in K}$ is not extractable is at most $(1 - k!/k^k)^\lambda$, which arises from the case where the challenge space of the base protocol is exactly k. By a union bound over all $\binom{|S|}{k} = \text{poly}(\lambda)$ choices of K, the probability there exists a size-k subset $K \subset S$ for which $\{r_i\}_{i \in [K]}$ is not extractable is at most $\text{poly}(\lambda) \cdot (1 - k!/k^k)^\lambda = \text{negl}(\lambda)$ since $1 - k!/k^k < 1$ is a constant. □

5 Alternating projector algorithms

In this section we define three algorithms which we will make use of in Section 6 to construct our quantum rewinding procedure. At the core of each one is the idea of alternating binary projective measurements. In Section 5.1, we define a unitary procedure which coherently computes an approximation of the Jordan spectrum of a state. In Section 5.2 we define an algorithm which "repairs" a state after measurement. In Section 5.3 we define a generic "amplification" algorithm which projects a state on to a given subspace under certain conditions.

5.1 The "approximate Jordan" unitary

We begin by defining the "coherent" application of a binary projective measurement.

Definition 5.1. For a binary projective measurement $\mathsf{M} = (\Pi, \mathbf{I} - \Pi)$ on $\mathcal{A}$, denote by U_{M} the unitary

$$P_s \otimes \mathbf{X}^{\mathcal{Z}} + (I - P_s) \otimes \mathbf{I}^{\mathcal{Z}}$$

on $(\mathcal{A}, \mathcal{Z})$ for one-qubit ancilla register $\mathcal{Z}$, where $\mathbf{X}$ is the Pauli bit-flip operator.

For binary projective measurements M, N on $\mathcal{A}$ and real numbers $0 < \epsilon, \delta < 1$, we define the *"approximate Jordan" unitary* $\mathsf{AJor}_{\epsilon,\delta}[\mathsf{M}, \mathsf{N}]$, acting on registers $(\mathcal{A}, \mathcal{L})$ where $\mathcal{L}$ is a $(2t+1)$-qubit register for $t \coloneqq n_{\epsilon,\delta}/2$ (where $n_{\epsilon,\delta} \coloneqq \log(1/2\delta)/2\epsilon^2$ is the parameter defined in Proposition 3.1 for our statement of the Chernoff bound). Let $\mathcal{L}_i$ denote the i-th qubit of $\mathcal{L}$, numbered from zero. $\mathsf{AJor}_{\epsilon,\delta}[\mathsf{M}, \mathsf{N}]$ denotes the following unitary operation:

$\underline{\mathsf{AJor}_{\epsilon,\delta}[\mathsf{M}, \mathsf{N}]}$:

1. Apply U_{N} to the registers $(\mathcal{R}, \mathcal{H}, \mathcal{L}_0)$.
2. For $i = 1, \ldots, t = n_{\epsilon,\delta}/2$:
   (a) Apply U_{M} to the registers $(\mathcal{R}, \mathcal{H}, \mathcal{L}_{2i-1})$.
   (b) Apply U_{N} to the registers $(\mathcal{R}, \mathcal{H}, \mathcal{L}_{2i})$.

For $L \in \{0,1\}^{n+1}$, define the *estimate*

$$p(L) \coloneqq |\{j \in \{1, \ldots, n\} : L_{j-1} = L_j\}|/n \in Z_n$$

where $Z_n \coloneqq \{0, 1/n, 2/n, \ldots, 1\}$. That is, $p(L)$ is the number of pairs of consecutive repeated bits in L, divided by n; for example $p(0,0,1,1,1,0) = 3/5$. We define projections on to lists with a given estimate.

- For $p^* \in Z_{2t}$, define $\Pi^{\mathcal{L}}_{p^*} \coloneqq \sum_{L, p(L)=p^*} |L\rangle\langle L|^{\mathcal{L}}$.
- For $p^* \in [0,1]$, define $\Pi^{\mathcal{L}}_{\geq p^*} \coloneqq \sum_{p \geq p^*} \Pi^{\mathcal{L}}_p$.
- For $p^* \in [0,1]$, define $\Pi^{\mathcal{L}}_{p^* \pm \epsilon} \coloneqq \sum_{p^* - \epsilon \leq p \leq p^* + \epsilon} \Pi^{\mathcal{L}}_p$.

To analyse the above unitary, we will consider the following similar (non-unitary) procedure $\mathsf{AJor}'_{\epsilon,\delta}[\mathsf{M}, \mathsf{N}]$:

1. Apply N to the register $\mathcal{A}$, obtaining outcome $L_0 \in \{0,1\}$.

2. For $i = 1, \ldots, t = n_{\epsilon,\delta}/2$:
   (a) Apply M to the register $\mathcal{A}$, obtaining outcome $L_{2i-1} \in \{0,1\}$.
   (b) Apply N to the register $\mathcal{A}$, obtaining outcome $L_{2i} \in \{0,1\}$.
3. Output $p := p(L_0, L_1, \ldots, L_{2T+1})$.

Let $\mathsf{AJor}'_{\epsilon,\delta}[\mathsf{M},\mathsf{N}](|\phi\rangle)$ be a random variable corresponding to the outcome p of this procedure applied to $|\phi\rangle$. It is easy to see that $\left\|\Pi^{\mathcal{L}}_{p^*}\mathsf{AJor}_{\epsilon,\delta}[\mathsf{M},\mathsf{N}]\,|\phi\rangle\,|0\rangle\right\|^2 = \Pr\left[\mathsf{AJor}'_{\epsilon,\delta}[\mathsf{M},\mathsf{N}](|\phi\rangle) = p^*\right]$ for $p^* \in Z_{2T}$.

Definition 5.2. For a binary projective measurements M,N, we say a state $|\phi\rangle$ is *aligned* with (M,N) if it is an eigenstate of M or N. The *Jordan spectrum* of a state $|\phi\rangle$ with respect to (M,N) is $(s_j = \left\|\Pi^{\mathrm{Jor}}_j\,|\phi\rangle\right\|^2)_j$, for $\mathsf{M}_{\mathrm{Jor}}[\mathsf{M},\mathsf{N}] = (\Pi^{\mathrm{Jor}}_j)_j$.

We now proceed to show some key properties of AJor. Throughout we will consider a fixed choice of M,N and ϵ,δ, and we write AJor for $\mathsf{AJor}_{\epsilon,\delta}[\mathsf{M},\mathsf{N}]$. We will denote by $|v_j\rangle, |v_j^\perp\rangle, |w_j\rangle, |w_j^\perp\rangle$ the vectors obtained by applying Lemma 3.4 to $(\Pi_{\mathsf{M}},\Pi_{\mathsf{N}})$, and by $\mathsf{M}_{\mathrm{Jor}} = (\Pi^{\mathrm{Jor}}_j)_j$ the corresponding Jordan subspace measurement. The following lemma relates the action of $\mathsf{AJor}\,|\phi\rangle\,|0\rangle$ with the Jordan spectrum of $|\phi\rangle$. This is a variant of lemmas appearing in [MW05, Zha20].

Lemma 5.3. *Let $X_j \sim \mathsf{Bin}(2T,p_j)$ be independent binomial random variables. Let $|\phi\rangle$ be a state aligned with (M,N) with Jordan spectrum $(s_j)_j$. Then $\left\|\Pi^{\mathcal{L}}_p\mathsf{AJor}\,|\phi\rangle\,|0\rangle\right\|^2 = \sum_j s_j \Pr[X_j/2T = p]$.*

Proof. We first consider the case where $\|\Pi_{\mathsf{N}}\,|\phi\rangle\|^2 = 1$. Since $\mathsf{M}_{\mathrm{Jor}}$ commutes with AJor,

$$\left\|\Pi^{\mathcal{L}}_p\mathsf{AJor}\,|\phi\rangle\,|0\rangle\right\|^2 = \sum_j s_j \Pr[\mathsf{AJor}'(|w_j\rangle) = p] \ .$$

Consider now the procedure $\mathsf{AJor}'(|w_j\rangle)$. Let $L' \in \{0,1\}^{2T}$ be defined as follows:

$$L'_i = \begin{cases} 1 & \text{if } L_{i-1} = L_i \\ 0 & \text{if } L_{i-1} \neq L_i. \end{cases}$$

Notice that $p(L)$ is equal to the normalised Hamming weight of L'. We now show that L' is distributed as $2T$ independent Bernoulli trials with parameter p_j, which completes the proof.

Observe that at the beginning of the i-th iteration of step 2, if $L_{2i-2} = 1$ then the state is $|w_j\rangle$ and if $L_{2i-2} = 0$ then the state is $|w_j^\perp\rangle$. In the former case, $L_{2i-1} = 1$ with probability p_j; in the latter case, $L_{2i-1} = 1$ with probability $1 - p_j$. Hence in both cases $L_{2i-2} = L_{2i-1}$ with probability p_j. A similar argument for $|v_j\rangle, |v_j^\perp\rangle$ shows that $L_{2i-1} = L_{2i}$ with probability p_j. Thus each L'_i is Bernoulli with parameter p_j.

For the remaining cases: $\|\Pi_{\mathsf{N}}\,|\phi\rangle\|^2 = 0$, $\|\Pi_{\mathsf{M}}\,|\phi\rangle\|^2 = 1$ and $\|\Pi_{\mathsf{M}}\,|\phi\rangle\|^2 = 0$, it suffices to observe that $\mathsf{AJor}'(|w_j^\perp\rangle)$, $\mathsf{AJor}'(|v_j\rangle)$, $\mathsf{AJor}'(|v_j^\perp\rangle)$ differ from $\mathsf{AJor}'(|w_j\rangle)$ only in the probability that $L_0 = 1$, which does not affect the distribution of L'. □

We now show that the above implies that the state after obtaining outcome p from AJor is concentrated on Jordan subspaces with eigenvalue close to p.

Corollary 5.4 (Post-measurement state)**.** *For any state $|\phi\rangle$ aligned with* M, N*,*

$$\sum_p \left\| \Pi^{\mathrm{Jor}}_{p\pm\epsilon} \Pi^{\mathcal{L}}_p \mathsf{AJor} \, |\phi\rangle \, |0\rangle \right\|^2 \geq 1 - \delta \ .$$

Proof. Since $\Pi^{\mathrm{Jor}}_{p\pm\epsilon}$ commutes with AJor, we may equivalently bound

$$\sum_p \left\| \Pi^{\mathcal{L}}_p \Pi^{\mathrm{Jor}}_{p\pm\epsilon} \mathsf{AJor} \, |\phi\rangle \, |0\rangle \right\|^2 = \sum_{p \in Z_{2t}} \sum_{\substack{j \\ p_j \in [p\pm\epsilon]}} s_j \Pr[X_j/2t = p] = \sum_j s_j \sum_{p \in [p_j \pm \epsilon] \cap Z_{2t}} \Pr[X_j/2t = p] \geq 1 - \delta \ ,$$

where the inequality follows by a Chernoff bound (Proposition 3.1). □

We also have the following approximate converse to the above; we state it in terms of Jordan subspaces with value *at least* p since this will be more useful for us later.

Corollary 5.5. *Let $|\phi\rangle$ be a state aligned with* M *or* N *such that* $\left\| \Pi^{\mathrm{Jor}}_{\geq p} |\phi\rangle \right\|^2 \geq 1 - \beta$*. Then* $\left\| \Pi^{\mathcal{L}}_{\geq p-\epsilon} \mathsf{AJor} \, |\phi\rangle \, |0\rangle \right\|^2 \geq 1 - \beta - \delta$*.*

Proof. $\left\| \Pi^{\mathcal{L}}_{\geq p-\epsilon} \mathsf{AJor} \, |\phi\rangle \, |0\rangle \right\|^2 = \sum_j q_j \Pr[X_j \geq p - \epsilon] \geq (1 - \delta) \sum_{j, p_j \geq p} q_j \geq 1 - \beta - \delta.$ □

The next claim shows that *postselecting* on a particular outcome p forces the post-measurement state to be concentrated on Jordan subspaces with value close to p, so long as the outcome was not too unlikely.

Claim 5.6. *Let $|\phi\rangle$ be a state and define $\gamma := \left\| \Pi^{\mathcal{L}}_{\geq p^*} \mathsf{AJor}_{\epsilon,\delta} \, |\phi\rangle \, |0\rangle \right\|^2$. Let $|\psi\rangle := \Pi^{\mathcal{L}}_{\geq p^*} \mathsf{AJor}_{\epsilon,\delta} \, |\phi\rangle \, |0\rangle \, / \sqrt{\gamma}$, and $\rho := \mathrm{Tr}_{\mathcal{L}}(\, |\psi\rangle\langle\psi|)$. Then $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p^* - \epsilon} \rho\right) \geq 1 - \delta/\gamma$.*

Proof. Let $|\phi_p\rangle := \Pi^{\mathcal{L}}_p \mathsf{AJor}_{\epsilon,\delta} \, |\phi\rangle \, |0\rangle$, $\rho_p := \mathrm{Tr}_{\mathcal{L}}(\, |\phi_p\rangle\langle\phi_p|)$, so that $\rho = \sum_{p \geq p^*} \rho_p / \gamma$. By Corollary 5.4, $\sum_p \mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p-\epsilon} \rho_p\right) \geq 1 - \delta$. Then since $\sum_{p \geq p^*} \mathrm{Tr}(\rho_p) = \gamma$,

$$1 - \delta \leq 1 - \gamma + \mathrm{Tr}\left(\sum_{p \geq p^*} \Pi^{\mathrm{Jor}}_{\geq p-\epsilon} \rho_p \right) \leq 1 - \gamma + \gamma \, \mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p^* - \epsilon} \rho\right) \ .$$

Rearranging completes the proof. □

5.2 A "measure-and-repair" lemma

We now introduce a generic *state repair* procedure $\mathsf{MeasRep}$. Let M, N be binary-outcome projective measurements. Define $\mathsf{MeasRep}(\mathsf{M}, \mathsf{N})$ as follows.

$\underline{\mathsf{MeasRep}(\mathsf{M}, \mathsf{N}) :}$

1. Apply the measurement M. If the outcome is 1, stop.
2. Repeat the following until $b \wedge b' = 1$:
   (a) Apply N, obtaining an outcome b.
   (b) Apply M, obtaining an outcome b'.

The following lemma gives a guarantee on the running time of MeasRep when preceded by an N-outcome projective measurement.

Lemma 5.7 (Measure-and-Repair Lemma)**.** *Let $|\phi\rangle$ be a state such that $\|\Pi_{\mathsf{M}}|\phi\rangle\|^2 = 1-\beta$ for some constant $\beta > 0$. Let $\mathsf{P} \coloneqq (\Pi_k)_{k=1}^N$ be an N-outcome projective measurement. Consider the following procedure:*

1. *Apply P, obtaining an outcome k. Let $\mathsf{N} \coloneqq (\Pi_k, \mathbf{I}-\Pi_k)$.*
2. *Run $\mathsf{MeasRep}(\mathsf{M},\mathsf{N})$.*

Let X be a random variable denoting the number of iterations of Step 2 of MeasRep *in the above procedure when applied to $|\phi\rangle$. Then for all $\gamma > 0$ and any positive integer T,*

$$\Pr\Big[X > T \text{ or } \|\Pi_{\mathsf{M}}|\psi\rangle\|^2 < \gamma\Big] \le T\gamma + \sqrt{\beta} + \frac{Ne^{-1/3}}{\sqrt{2T+1}},$$

where $|\psi\rangle$ is the state just before applying the measurement M in the final iteration of MeasRep *(if* MeasRep *terminates in Step 1, $|\psi\rangle$ is the state just before running* MeasRep*).*

Remark 5.8. *We stress that the procedure is not guaranteed to recover the original state $|\phi\rangle$, even if $\|\Pi_{\mathsf{B}}|\phi\rangle\|^2 = 1$, since applying P to $|\phi\rangle$ may destroy information about $|\phi\rangle$.*

We now prove Lemma 5.7.

Proof. We first prove that the probability the $\mathsf{MeasRep}(\mathsf{M},\mathsf{N})$ procedure takes more than T iterations is at most $\sqrt{\beta} + \frac{Ne^{-1/3}}{\sqrt{2T+1}}$.

Let $|\phi'\rangle \coloneqq \Pi_{\mathsf{M}}|\phi\rangle / \sqrt{1-\beta}$; we will analyse the random variable X' which is as X but with respect to $|\phi'\rangle$, and then show that X is close to X'.

Fix some outcome $k \in [N]$, and let E_k be the event that outcome k is obtained when applying P. Let $|\psi_k\rangle$ be the state after measuring P conditioned on obtaining outcome k, i.e. $|\psi_k\rangle \coloneqq \Pi_k|\phi'\rangle / \sqrt{q_k}$, where $q_k \coloneqq \|\Pi_k|\phi'\rangle\|^2$. For each j, let $(|v_j\rangle, |v_j^\perp\rangle)$ and $(|w_j\rangle, |w_j^\perp\rangle)$ be the bases for the Jordan subspaces (Lemma 3.4) for Π_{B}, Π_k (respectively), and write $|\phi'\rangle = \sum_j \alpha_j |v_j\rangle$. Then $|\psi_k\rangle = \sum_j \alpha_j \sqrt{p_j}|w_j\rangle / \sqrt{q_k}$.

We first argue the following must hold:

$$\Pr[X > T \mid E_k] = \frac{1}{q_k}\sum_j |\alpha_j|^2 p_j \Pr[\mathsf{Time}(\mathsf{MeasRep}(\mathsf{M},\mathsf{N}), |w_j\rangle) > T \mid E_k], \tag{1}$$

where $\mathsf{Time}(\mathsf{MeasRep}(\mathsf{M},\mathsf{N}), |w_j\rangle)$ denotes the number of iterations of Step 2 when MeasRep is applied to $|w_j\rangle$.

Eq. (1) can be established by analyzing the effect of applying the projective measurement $\mathrm{Jor}[\Pi_{\mathsf{B}}, \Pi_k]$ (Definition 3.5) to the state. Conditioned on E_k, every measurement applied in the $\mathsf{MeasRep}(\mathsf{M},\mathsf{N})$ procedure commutes with $\mathrm{Jor}[\Pi_{\mathsf{B}}, \Pi_k]$ (Proposition 3.6). Therefore, conditioned on E_k, applying $\mathrm{Jor}[\Pi_{\mathsf{B}}, \Pi_k]$ *before* applying $\mathsf{MeasRep}(\mathsf{M},\mathsf{N})$ does not affect the number of times Step 2 is repeated, and Eq. (1) follows.

We now prove that

$$\Pr[\mathsf{Time}(\mathsf{MeasRep}(\mathsf{M},\mathsf{N}), |w_j\rangle) > T \mid E_k] \le (1-p_j)(1-p_j^2(1-p_j))^T.$$

We can immediately observe that the probability Step 1 (i.e., applying measurement B to $|w_j\rangle = \Pi_k |w_j\rangle$) returns 1 is p_j, meaning we repeat Step 2 zero times with probability p_j. So with probability $1 - p_j$, the $\mathsf{MeasRep}(\mathsf{M}, \mathsf{N})$ will proceed to Step 2. Immediately before any iteration of Step 2, the state is either $|v_j\rangle$ or $|v_j^{\perp}\rangle$. In the former case, the probability of obtaining 1 outcomes for M_k followed by B is p_j^2. In the latter case, the probability is $p_j(1-p_j)$. Hence in both cases, the probability is at least $p_j^2(1-p_j)$, and so $\Pr[\mathsf{Time}(\mathsf{MeasRep}(\mathsf{M}, \mathsf{N}), |w_j\rangle) > T] \leq (1-p_j)(1-p_j^2(1-p_j))^T$. Then

$$\begin{aligned}\Pr[X > T \mid E_k] &\leq \frac{1}{q_k} \sum_j |\alpha_j|^2 p_j(1-p_j)(1-p_j^2(1-p_j))^T \\ &\leq \frac{e^{-1/3}}{q_k\sqrt{2T+1}}\end{aligned}$$

since for all $p_j \in [0,1]$, $p_j(1-p_j)(1-p_j^2(1-p_j))^T \leq e^{-1/3}/\sqrt{2T+1}$, and $\sum_j |\alpha_j|^2 = 1$.

Now $\Pr[E_k] = q_k$, and so

$$\Pr[X' > T] = \sum_{k=1}^{N} \Pr[X' > T \mid E_k] \Pr[E_k] \leq \frac{Ne^{-1/3}}{\sqrt{2T+1}} \ .$$

The fidelity of $|\phi\rangle$ and $|\phi'\rangle$ is $|\langle\phi|\phi'\rangle|^2 = |\langle\phi| P |\phi\rangle|^2/(1-\beta) = 1-\beta$. Hence the trace distance between $|\phi\rangle$ and $|\phi'\rangle$ is at most $\sqrt{\beta}$, from which we obtain that

$$\Pr[X > T] \leq \frac{Ne^{-1/3}}{\sqrt{2T+1}} + \sqrt{\beta} \ .$$

Let E_i be the event that $\mathsf{MeasRep}(\mathsf{M}, \mathsf{N})$ terminates at the i-th iteration; let $|\phi_i\rangle$ be (a random variable describing) the state in this iteration after applying M_k. Let $E_i^{\geq}$ be the subevent of E_i where $\|P|\phi_i\rangle\|^2 \geq \gamma$, and $E_i^{<} \coloneqq E_i \setminus E_i^{\geq}$. Observe that $\Pr[X \leq T] = \Pr[\cup_i E_i] \leq \Pr\left[\cup_i E_i^{\geq}\right] + \sum_i \Pr[E_i^{<}]$. Hence $\Pr\left[\cup_i E_i^{\geq}\right] \geq \Pr[X \leq T] - T\gamma$, which completes the proof. □

5.3 Amplification

Let $\mathsf{Amp}(\mathsf{M}, \mathsf{N})$ be the procedure which applies the measurements M, N in an alternating fashion until a 1 outcome for M occurs.

Lemma 5.9. *Let* $\mathsf{M} = (\Pi_{\mathsf{M}}, \mathbf{I} - \Pi_{\mathsf{M}}), \mathsf{N} = (\Pi_{\mathsf{N}}, \mathbf{I} - \Pi_{\mathsf{N}})$ *be binary projective measurements such that the maximum eigenvalue of* $\Pi_{\mathsf{N}}\Pi_{\mathsf{M}}\Pi_{\mathsf{N}}$ *is at most* $3/4$. *Let* ρ *be a state aligned with* (M, N) *such that* $\mathrm{Tr}\left(\Pi_{\geq p}^{\mathrm{Jor}}\rho\right) \geq 1-\beta$ *for* $p \in [0, 3/4], \beta \in [0,1]$. *Then*

$$\Pr[\mathsf{Time}(\mathsf{Amp}(\mathsf{M}, \mathsf{N}), \rho) > S] \leq (1-p)^S + \beta \ .$$

Moreover, letting ρ' *be the resulting state, it holds that* $\mathrm{Tr}\left(\Pi_{\geq p}^{\mathrm{Jor}}\rho'\right) \geq 1-\beta$.

Proof. Since M, N commute with $\mathsf{M}_{\mathrm{Jor}}[\Pi_{\mathsf{M}}, \Pi_{\mathsf{N}}]$, the "moreover" part holds.

Let $\rho = \sum_i q_i \ketbra{\phi_i}{\phi_i}$, and let $(s_j^{(i)})_j$ be the Jordan spectrum of $\ketbra{\phi_i}{\phi_i}$ with respect to $(\mathsf{M}_P, \mathsf{M}_Q)$. It holds that

$$\begin{aligned}\Pr[\mathsf{Time}(\mathsf{Amp}(P,Q),\rho) > S] &= \sum_i q_i \sum_j s_j^{(i)} \cdot \Pr\Big[\mathsf{Time}(\mathsf{Amp}(P,Q), |w_j^{\perp}\rangle) > S\Big] \\ &\leq \sum_i q_i \sum_{j, p_j \geq p} s_j^{(i)} \cdot \Pr\Big[\mathsf{Time}(\mathsf{Amp}(P,Q), |w_j^{\perp}\rangle) > S\Big] + \beta\end{aligned}$$

since $\beta \geq \mathrm{Tr}\Big((I - \Pi^{\mathrm{Jor}}_{\geq p})\rho\Big) = \sum_i q_i \sum_{j, p_j < p} |\alpha_j^{(i)}|^2$.

Consider now the procedure $\mathsf{Amp}(P,Q)$ applied to $|w_j^{\perp}\rangle$ where $p_j \geq p$. The probability that, after applying M_P, applying M_Q yields 0 is $1 - 2p_j(1-p_j) \leq 1 - p/2$, since $p_j \leq 3/4$; if this occurs, the post-measurement state is $|w_j^{\perp}\rangle$. Thus the probability that, in S iterations, applying M_Q always yields 0 is at most $(1-p/2)^S$. □

6 Quantum extraction

In this section we prove our main technical theorem: if an interactive argument $\mathsf{ARG} = (P, V)$ is collapsing and probabilistically special sound then ARG is a post-quantum proof of knowledge.

Theorem 6.1. *Let $\mathsf{ARG} = (P, V)$ be a m-round interactive argument system. Suppose that ARG is collapsing and probabilistically k-special sound for some $k = \mathrm{poly}(\lambda)$. Then for every constant $\delta > 0$, ARG is a post-quantum proof of knowledge with knowledge error $(1+\delta)k/|R_m| + \mathrm{negl}(\lambda)$.*

In Section 6.1 we present our quantum rewinding procedure $\mathsf{QRewind}$ for general projective measurements. In Section 6.2 we prove a guarantee on the behavior of $\mathsf{QRewind}$ which is sufficient for our application. Finally, in Section 6.3 we prove Theorem 6.1.

6.1 A quantum rewinding procedure

The quantum rewinding procedure takes as input a list of projectors $(P_r)_{r \in R}$ acting on a register $\mathcal{H}$ and $\delta_0 \in [0,1]$, and operates on registers $\mathcal{R}, \mathcal{H}, \mathcal{L}$ where $\mathcal{R}$ is a register taking values in the set $R \cup \{\top, \bot\}$ and $\mathcal{L}$ is a register on $2t+1$ qubits for t to be fixed later.

Before presenting the procedure we provide some definitions. Let $\mathsf{CProj} \coloneqq (\Pi_{\mathsf{CProj}}, \mathbf{I} - \Pi_{\mathsf{CProj}})$, where

$$\Pi_{\mathsf{CProj}} \coloneqq \sum_{r \in R} \ketbra{r}{r} \otimes P_r + \ketbra{\top}{\top} \otimes \mathbf{I} \ .$$

Define the state $|\mathbb{1}_R\rangle \coloneqq \frac{1}{\sqrt{2|R|}} \sum_{r \in R} |r\rangle + \frac{1}{2}|\top\rangle + \frac{1}{2}|\bot\rangle$. We define a projective measurement $\mathsf{M}_{|\mathbb{1}_R\rangle} \coloneqq (\ketbra{\mathbb{1}_R}{\mathbb{1}_R}^{\mathcal{R}}, \mathbf{I} - \ketbra{\mathbb{1}_R}{\mathbb{1}_R}^{\mathcal{R}})$ on to this state. We define $N \coloneqq \lceil 2|R| \ln(1/2\delta_0) \rceil$. We define $\epsilon \coloneqq \delta_0/10N$ and $\delta \coloneqq (\delta_0/(10N|R|))^8$, and $t \coloneqq n_{\epsilon,\delta}/2$ for $n_{\epsilon,\delta}$ as defined in Proposition 3.1.

Finally, for $p \in (0,1)$, we define the projection $M_{p,\epsilon,\delta}$ on $\mathcal{H}, \mathcal{R}, \mathcal{L}$ as

$$M_{p,\epsilon,\delta} \coloneqq \mathsf{AJor}_{\epsilon,\delta}[\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}]^{\dagger} \cdot \Pi^{\mathcal{L}}_{\geq p-\epsilon} \cdot \mathsf{AJor}_{\epsilon,\delta}[\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}] \ .$$

Remark 6.2. *Notice that we have defined the state $|\mathbb{1}_R\rangle$ and projection Π_{CProj} such that every eigenstate of $E = \ketbra{\mathbb{1}_R}{\mathbb{1}_R} \Pi_{\mathsf{CProj}} \ketbra{\mathbb{1}_R}{\mathbb{1}_R}$ of the form $|\mathbb{1}_R\rangle|\psi\rangle$ has eigenvalue between $1/4$ and $3/4$. In particular E has no eigenvalue larger than $3/4$.*

$\underline{\mathsf{QRewind}((P_r)_R, \delta_0) :}$

Step 1. a) Initialize the $\mathcal{R}$ register to the state $|\mathbb{1}_R\rangle$.

b) Apply $\mathsf{AJor}'_{\epsilon,\delta}[\mathsf{CProj}, |\mathbb{1}_R\rangle]$ to registers $(\mathcal{R}, \mathcal{H})$, obtaining outcome p.

c) Run $\mathsf{Amp}^{\mathcal{R},\mathcal{H}}(\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_\mathcal{R}\rangle})$ for at most $\lceil 20 \log(1/\delta) \rceil$ steps.

Step 2. Let $W \coloneqq \emptyset$. Repeat N times:

a) Measure the $\mathcal{R}$ register, obtaining an outcome $r \in R \cup \{\top, \bot\}$.

b) Apply CProj to registers $(\mathcal{R}, \mathcal{H})$, obtaining an outcome $b \in \{0, 1\}$.

c) If $b = 1$, and $r \neq \top$, set $W \leftarrow W \cup \{r\}$.

d) Set the $\mathcal{L}$ register to $|0\rangle$. Set $\mathsf{M} \coloneqq (\Pi, \mathbf{I} - \Pi)$, where

$$\Pi \coloneqq \begin{cases} |r\rangle\langle r| \otimes P_r \otimes |0\rangle\langle 0|^{\mathcal{L}} & \text{if } b = 1, \text{ and} \\ |r\rangle\langle r| \otimes (\mathbf{I} - P_r) \otimes |0\rangle\langle 0|^{\mathcal{L}} & \text{if } b = 0, \end{cases}$$

where $P_\top \coloneqq \mathbf{I}$ and $P_\bot \coloneqq 0$.

e) Run $\mathsf{MeasRep}(M_{p,\epsilon,\delta}, \mathsf{M})$ for at most $(10N|R|/\delta_0)^2 = 1/\sqrt[4]{\delta}$ iterations.

f) Apply $\mathsf{AJor}_{\epsilon,\delta}[\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_R\rangle}]$ to registers $(\mathcal{R}, \mathcal{H}, \mathcal{L})$. Trace out the $\mathcal{L}$ register.

g) Update $p \leftarrow p - \epsilon$.

h) Run $\mathsf{Amp}^{\mathcal{R},\mathcal{H}}(\mathsf{CProj}, \mathsf{M}_{|\mathbb{1}_\mathcal{R}\rangle})$ for at most $\lceil 20 \log(1/\delta) \rceil$ steps.

Step 3. Output W.

We call an iteration of Step 2 *successful* if $b = 1$. The *outcome* of an iteration is r.

6.2 Quantum rewinding lemma

Lemma 6.3 (Correctness of Quantum Rewinding)**.** *Let $|\psi\rangle$ be a state for which*

$$\frac{1}{R} \sum_{r \in R} \|P_r |\psi\rangle\|^2 = \eta.$$

Consider the random variable $W \leftarrow \mathsf{QRewind}((P_r)_{r \in R}, \delta_0)$, where the $\mathcal{H}$ register is in state $|\psi\rangle$. The expected size of W is at least $(\eta - \delta_0)|R|$. The rewinding procedure runs in time $\mathrm{poly}(|R|, 1/\delta_0)$.

We begin with lemma about the effect of a single step of the rewinding procedure.

Claim 6.4. *Let T be the CPTP map on $\mathcal{R}, \mathcal{H}$ implemented by a single iteration of Step 2. Suppose that $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p} \rho\right) \geq 1 - \sqrt{\delta}$ for the value of p at the beginning of the iteration, and $p \geq 1/8 + \epsilon$. Then $T(\rho) = \zeta \rho_{\mathsf{good}} + (1 - \zeta)\rho_{\mathsf{err}}$ where $\rho_{\mathsf{good}}, \rho_{\mathsf{err}}$ are density matrices, $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p - \epsilon} \rho_{\mathsf{good}}\right) \geq 1 - \sqrt{\delta}$, $\mathrm{Tr}\left(|\mathbb{1}_R\rangle\langle\mathbb{1}_R|^{\mathcal{R}} \rho_{\mathsf{good}} \right) = 1$ and $\zeta \geq 1 - \delta_0/10N$.*

Proof. Write $\rho \coloneqq \sum_i q_i \ket{\phi_i}\bra{\phi_i}$. Let $\beta_i \coloneqq 1 - \left\|\Pi^{\mathrm{Jor}}_{\geq p} \ket{\phi_i}\right\|^2$; it holds that $\sum_i q_i\beta_i = \sqrt{\delta}$.

We now consider the effect of Step 2e) on the state $\ket{\phi_i}\ket{0}$. Since $\ket{\phi_i}$ is aligned with $\mathsf{M}_{\ket{\mathbb{1}_R}}$, $\|M_{p,\epsilon,\delta}\ket{\phi_i}\|^2 \geq 1-\beta_i-\delta$. By Lemma 5.7, setting $\gamma \coloneqq \sqrt{\delta}$, with probability $\zeta_i \geq 1-\sqrt[4]{\delta}-\sqrt{\beta_i}-\sqrt{\delta}-\delta_0/20N$ it holds that the (normalized) state $\ket{\psi_i}$ after Step 2f) is equal to $\mathsf{AJor} M_{p,\epsilon,\delta}\ket{\phi_i'}\ket{0}/\sqrt{\gamma'}$ for some state $\ket{\phi_i'}$ aligned with CProj and $\gamma' \geq \sqrt{\delta}$. In this case, by Claim 5.6, it holds that $\left\|\Pi^{\mathrm{Jor}}_{\geq p-\epsilon}\ket{\psi_i}\right\|^2 \geq 1-\delta/\gamma' \geq 1-\sqrt{\delta}$.

Let ρ' be the state on $\mathcal{R},\mathcal{H}$ after Step 2f). Then $\rho' = \sum_i q_i(\zeta_i\rho^{(i)}_{\mathsf{good}} + (1-\zeta_i)\rho^{(i)}_{\mathsf{err}})$, for density matrices $\rho^{(i)}_{\mathsf{good}}$, $\rho^{(i)}_{\mathsf{err}}$ such that for all i, $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq p-\epsilon}\rho^{(i)}_{\mathsf{good}}\right) \geq 1-\sqrt{\delta}$. By Jensen's inequality, $\sum_i q_i\zeta_i \geq 1-2\sqrt[4]{\delta}-\sqrt{\delta}-\delta_0/20N$.

Finally by Claim 5.6, since $\ket{\mathbb{1}_R}\bra{\mathbb{1}_R}\cdot\Pi_{\mathsf{CProj}}\cdot\ket{\mathbb{1}_R}\bra{\mathbb{1}_R}$ and $p \geq 1/8$, with probability at least $1-\delta-\sqrt{\delta}$, Step 2h) terminates within the prescribed number of steps. Conditioned on this event, the resulting state is in the image of $\ket{\mathbb{1}_R}\bra{\mathbb{1}_R}$, from which the claim follows. □

Proof of Lemma 6.3. The running time of the procedure is clear from the description. We now prove correctness.

We divide the range $[0,1]$ into $m = \lceil 1/\sqrt[4]{\delta}\rceil$ intervals

$$U = ([0,\sqrt[4]{\delta}),[\sqrt[4]{\delta},2\sqrt[4]{\delta}),\cdots,[(m-1)\sqrt[4]{\delta},1]).$$

Let $u_\ell = (\ell-1)\sqrt[4]{\delta}$ denote the lower bound on the ℓ-th interval in U. We write the state ρ on $\mathcal{R},\mathcal{H}$ at the beginning of the first iteration of Step 2 as $\rho = \sum_{\ell\in[m]} a_\ell\rho_\ell$, where ρ_ℓ is the state conditioned on the estimate p obtained in Step 1b) lying in the ℓ-th interval in U. By Corollary 5.4, $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq u_\ell-\epsilon}\rho_\ell\right) \geq 1-\delta/a_\ell$ and by Lemma 5.3, $\sum_\ell u_\ell\cdot a_\ell \geq E[p] - \sqrt[4]{\delta} = 1/4+\eta/2-\sqrt[4]{\delta}$.

Note that for any state $\ket{\psi}$, the Jordan spectrum $(s_j)_j$ of $\ket{\mathbb{1}_R}\ket{\psi}$ with respect to $(\mathsf{CProj},\mathsf{M}_{\ket{\mathbb{1}_R}})$ is supported entirely on j such that $1/4 \leq p_j \leq 3/4$, and so the probability that $p \leq 1/4-\epsilon$ is at most δ. Hence by Claim 5.6, since the largest eigenvalue of $\ket{\mathbb{1}_R}\bra{\mathbb{1}_R}\cdot\Pi_{\mathsf{CProj}}\cdot\ket{\mathbb{1}_R}\bra{\mathbb{1}_R}$ is at most $3/4$, the probability that Step 1c) succeeds is at at least $1-2\delta$.

We now show by induction that, if the state at the beginning of the first iteration of Step 2 is $\rho^{(1)}_\ell \coloneqq \rho_\ell$ for ℓ such that $a_\ell \geq \sqrt{\delta}$ (in particular, $p \geq 1/4-\epsilon$), then for all $i \leq N$, there exist $\rho^{(i)}_{\mathsf{good}}, \rho^{(i)}_{\mathsf{err}}$, $q_i \in [0,1]$ such that the state at the beginning of the i-th iteration is $\rho^{(i)}_\ell \coloneqq T^{i-1}(\rho^{(1)}_\ell) = q_i\rho^{(i)}_{\mathsf{good}} + (1-q_i)\rho^{(i)}_{\mathsf{err}}$, where $q_i \geq 1-i\cdot(\delta_0/10N)$, $\mathrm{Tr}\left(\Pi^{\mathrm{Jor}}_{\geq u_\ell - i\cdot\epsilon}\rho^{(i)}_{\mathsf{good}}\right) \geq 1-\sqrt{\delta}$ and $\mathrm{Tr}\left(\Pi^{\mathcal{R}}_{\ket{\mathbb{1}_R}}\rho^{(i)}_{\mathsf{good}}\right) = 1$. Clearly this holds for $i=1$ (with $q_i \leq \delta$ by Claim 5.6); now assume that this holds for some $i \in \mathbb{N}$. By applying Claim 6.4, and by linearity, there exist density matrices $\rho'_{\mathsf{good}}, \rho'_{\mathsf{err}}, \rho''_{\mathsf{err}}$ and $\beta \in [0,1]$ such that $\rho^{(i+1)}_\ell = q_i(\beta\rho'_{\mathsf{good}} + (1-\beta)\rho'_{\mathsf{err}}) + (1-q_i)\rho''_{\mathsf{err}}$, $\beta \geq 1-\delta_0/10N$. Then we set $\rho^{(i+1)}_{\mathsf{good}} \coloneqq \rho'_{\mathsf{good}}$ and $\rho^{(i+1)}_{\mathsf{err}} = (q_i(1-\beta)\rho'_{\mathsf{err}} + (1-q_i)\rho''_{\mathsf{err}})/(1-q_i\beta)$, and $q_{i+1} \coloneqq q_i\beta$.

By linearity, the state at iteration i is $\rho^{(i)} = \sum_\ell a_\ell\rho^{(i)}_\ell$. Observe that $\sum_{\ell,a_\ell\geq\sqrt{\delta}} a_\ell u_\ell \geq \eta/2 - 2\sqrt[4]{\delta}$. It holds by Claim 3.7 that the probability that the i-th outcome is successful is $\mathrm{Tr}\left(\Pi_{\mathsf{CProj}}\rho^{(i)}\right) = \sum_\ell a_\ell \mathrm{Tr}\left(\Pi_{\mathsf{CProj}}\rho^{(i)}_\ell\right) \geq (\sum_{\ell,a_\ell\geq\sqrt{\delta}} a_\ell u_\ell) - i\cdot(\epsilon+\delta_0/10N) - \sqrt{\delta} \geq 1/4+\eta/2-\delta_0/4$ for $i \leq N$.

Let X_i denote the size of W (i.e., number of distinct successful outcomes that are not $\top$) after i iterations. We have that

$$E[X_i] \geq E[X_{i-1}] + 1/4 + \eta/2 - \delta_0/4 - E[X_{i-1}]/2|R| - 1/4 \ .$$

Solving the recurrence, $E[X_N] \geq 2|R|(\eta/2-\delta_0/4)\cdot(1-(1-\frac{1}{2|R|})^N) \geq (\eta-\delta_0)|R|$. □

6.3 Proof of Theorem 6.1

Let $\{\tilde{P}_\lambda\}_{\lambda\in\mathbb{N}}$ be a polynomial-size interactive quantum circuit family, and let $\lambda \in \mathbb{N}$ be such that $\tilde{P}_\lambda = (U^{(1)}, \ldots, U^{(m)})$ convinces V with probability at least $\varepsilon_\lambda \geq (1+\delta)k(\lambda)/|R(\lambda)|$; otherwise the proof of knowledge property holds trivially. For the remainder of the proof we drop λ from the notation. Without loss of generality we will assume $\delta \leq 1/4$.

Setup. Denote by $R = R_m$ the verifier's challenge space in the last round of the protocol. For a transcript τ of the first $m-1$ rounds and $r \in R$, define $V_{\tau,r} \coloneqq \sum_{z,V(\tau,r,z)=1} |z\rangle\langle z|^{\mathcal{Z}}$ where $\mathcal{Z} = \mathcal{Z}_m$ is the prover's response register in the last round. Define the projection Π_r on $(\mathcal{I}, \mathcal{Z})$ as

$$\Pi_r \coloneqq (U^{(m)}(r))^\dagger \cdot V_{\tau,r} \cdot U^{(m)}(r)$$

Modified rewinding procedure. Let $\mathsf{QRewind}'$ denote the procedure $\mathsf{QRewind}$ with the following modification. Let $\mathcal{H} = (\mathcal{I}, \mathcal{Z})$ as above. We replace Step 2c) with:

Step 2c') If $b = 1$, then measure the prover's response register $\mathcal{Z}$, obtaining outcome z.
If there is no z' such that $(r, z') \in W$, set $W \leftarrow W \cup \{(r, z)\}$.

Observe that the output of $\mathsf{QRewind}'((\Pi_r)_r, \delta_0)$ is then a set W of pairs (r, z) such that every pair in W has distinct r and for all $(r, z) \in W$, $V(\tau, r, z) = 1$. Moreover, because ARG is collapsing, no efficient adversary can (black-box) distinguish $\mathsf{QRewind}$ from the procedure which runs $\mathsf{QRewind}'$ and outputs the set $\{r : \exists z, (r, z) \in W\}$.

We define an adversary Adv for PSSExp as follows.

1. Run $\langle \tilde{P}, V\rangle$ until round $m-1$, obtaining a transcript prefix τ.
2. Query $\mathsf{Sampler}$ to obtain a uniformly random subset $S \subseteq R$ of size
$$\max(\min(2\lambda(1/\delta\varepsilon)^2, |R|), (1+\delta)k/\varepsilon) \ .$$
3. Run $((r_1, a_1), \ldots, (r_t, a_t)) \leftarrow \mathsf{QRewind}'((\Pi_{\tau,r})_{r\in S}, \delta\varepsilon/4)$.
4. If $t > k$, output $(\tau, r_1, a_1), \ldots, (\tau, r_k, a_k)$, otherwise output $\perp$.

The size of Adv is $\mathrm{poly}(\lambda, 1/\varepsilon)$.

Let p_τ be the probability that $\tilde{P}$ causes the verifier to accept given transcript τ, and let $|\psi\rangle$ be a random variable denoting the state after Step 1. Observe that $p_\tau = \frac{1}{|R|}\sum_{r\in R} \|P_r |\psi\rangle\|^2$, and $\mathbb{E}_\tau(p_\tau) \geq \varepsilon$. By Hoeffding's inequality, with probability $1 - e^{-\lambda}$ it holds that $\frac{1}{|S|}\sum_{r\in S} \|P_r |\psi\rangle\|^2 \geq p_\tau - \delta\varepsilon/2$. Hence by Lemma 6.3, the expected number t of successful iterations of $\mathsf{QRewind}$ with distinct outcomes $\mathbb{E}[t] \geq \mathbb{E}_\tau[(p_\tau - 3\delta\varepsilon/4 - e^{-\lambda})|S|] \geq (1 - 3\delta/4)\varepsilon|S| - \mathrm{negl}(\lambda)$. Since $(1 - 3\delta/4)(1+\delta) \geq 1 + \delta/16$ is a constant larger than 1 for $\delta \leq 1/4$, by Markov's inequality it holds that $t \geq \varepsilon|S|/(1+\delta) \geq k$ with probability $\Omega(\varepsilon) - \mathrm{negl}(\lambda)$.

By collapsing, this also holds for $\mathsf{QRewind}'$. Then by the guarantee of the PSS experiment, the extractor succeeds with probability $\Omega(\varepsilon) - \mathrm{negl}(\lambda)$.

7 Collapsing vector commitments

Looking ahead to Section 8, we will formally instantiate Kilian's protocol with a *vector commitment* [CF13], a cryptographic primitive that generalizes a Merkle tree built from a collision-resistant hash function.

In this section, we recall the standard *position-binding* definition of vector commitments, and also introduce a new notion of *collapsing* for vector commitments (Section 7.1). Moreover, we show that both position-binding and collapsing are satisfied by a Merkle tree built from any collapsing hash function (Section 7.2).

We define collapsing vector commitments (Section 7.1), and then prove that Merkle trees are collapsing vector commitments when the underlying hash function is collapsing (Section 7.2).

7.1 Definition

A (static) vector commitment scheme VC 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, an index $Q \in [\ell]$, 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.[13] In more detail, for every security parameter λ, alphabet Σ, vector length $\ell \in \mathbb{N}$, and polynomial-size 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 suffices to prove security of Kilian's protocol against classical adversaries, it is not known to suffice to prove security against quantum adversaries. (And, as discussed in Section 1, it is unlikely to.) We will therefore rely on an additional *collapsing* property that we introduce.

[13] We remark that unlike traditional cryptographic commitments, *vector* commitments are typically not required to satisfy an explicit hiding property.

Definition 7.1. 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 $v \in \Sigma^S$ 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 7.2. *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.*

7.2 Merkle trees are collapsing

We describe Merkle trees as an instance of vector commitments (Section 7.1), and then prove that they are collapsing when the underlying hash function is collapsing.

Construction 7.3. 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 7.4. *If $\mathcal{H}$ is a post-quantum secure collision resistant hash function with input size $n(\lambda)$ and output size $\ell(\lambda) = n(\lambda)/2$ then $\mathsf{VC} := \mathsf{Merkle}[\mathcal{H}]$ is a position binding over alphabet $\Sigma := \{0,1\}^{n(\lambda)}$.*

We now show that if $\mathcal{H}$ is a collapsing hash function then Merkle

Claim 7.5. *If $\mathcal{H}$ is a collapsing hash function with input size $n(\lambda)$ and output size $\ell(\lambda) = n(\lambda)/2$ 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 $j \in \{0, 1, 2, \ldots, d\}$, 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 $\mathcal{Y}_2, \ldots, \mathcal{Y}_{d+1}$. Specifically:
   (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 XORs the result onto $\mathcal{Y}_k$.
   (b) The challenger applies U_i and then measures the bit indicating whether $\mathcal{Y}_{d+1}$ equals rt (it applies the binary projective measurement $(|\mathsf{rt}\rangle\langle\mathsf{rt}|^{\mathcal{Y}_{d+1}}, \mathbf{I} - |\mathsf{rt}\rangle\langle\mathsf{rt}|^{\mathcal{Y}_{d+1}})$). 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 $(\mathcal{Y}_{d-j+1}, \ldots, \mathcal{Y}_{d+1})$; in the case that $j = 0$, 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.

Observe that 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 7.1 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 $\epsilon(\lambda)$. We construct an adversary Adv_j that has distinguishing advantage at least $\epsilon(\lambda)$ for the hash 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, we set $\mathcal{V} := \mathcal{Y}_1$ and 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 forwards $\mathcal{Y}_{d-j}$ 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 $\mathbf{H}_{j+1}$. □

8 Post-quantum security of Kilian's protocol

In this section we prove our main theorem. 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 8.1 below).

Theorem 8.1. *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* $\mathfrak{R}$*.*

Combined with our construction of collapsing vector commitments from collapsing hash functions (Claim 7.5), and the instantiation of collapsing hash functions from post-quantum hardness of LWE [Unr16a], this implies that there exist post-quantum succinct non-interactive arguments for NP assuming post-quantum hardness of LWE.

In Section 8.1 we describe Kilian's protocol in detail. In Section 8.2 we prove that Kilian is probabilistically special sound. In Section 8.3 we prove Theorem 8.1.

8.1 Protocol description

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 $(P, V) := \mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$, is specified below. Note that the argument prover P receives as input an instance x and witness w, while the argument verifier V receives as input just the instance 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 the PCP verifier makes any query outside of Q then reject.)

8.2 Kilian's protocol is probabilistically special sound

Notation. Let $\mathsf{Ext}^{\mathsf{Kilian}}$ denote the standard PCP extractor for Kilian's protocol that takes as input $(\mathsf{ck}, \mathsf{cm}, \{(r_i, z_i)\}_{i\in[k]})$, to be interpreted as a collection of k protocol transcripts with shared prefix $\tau = (\mathsf{ck}, \mathsf{cm})$, and performs the following steps:

1. Verify that all the transcripts are valid, i.e., $V^{\mathsf{Kilian}}(\tau, r_i, z_i) = 1$ for each i. Moreover, it verifies that each r_i is unique. If any of these checks fail, abort and output $\perp$.
2. 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.
3. 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 $\perp$.
4. Output π constructed by "stitching together" the answers given in $\{(Q_{r_i}, \pi[Q_{r_i}])\}_{i\in[k]}$, and filling in any unanswered indices arbitrarily.

For a $\mathsf{PCP} = (\mathbf{P}_{\mathsf{PCP}}, \mathbf{V}_{\mathsf{PCP}})$, instance x, and PCP π, we denote its *acceptance probability* as

$$\mathsf{WIN}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \coloneqq \Pr[\mathbf{V}^{\pi}_{\mathsf{PCP}}(x) = 1].$$

Let V^{Kilian} denote the verifier algorithm in $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ that takes as input an instance x and transcript $(\mathsf{ck}, \mathsf{cm}, r, z)$ and outputs 0 or 1.

A PCP extraction lemma. We now prove that if an efficient Adv outputs k accepting transcripts for Kilian's protocol with shared prefix τ whose challenges must be chosen from a set of randomly sampled challenges of size $N = \mathrm{poly}(\lambda)$, then running $\mathsf{Ext}^{\mathsf{Kilian}}$ on these transcripts yields a π satisfying $\mathsf{WIN}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \geq k/(2N)$ except with $\mathrm{negl}(\lambda)$ probability.

We stress that until we specify the underlying PCP, this claim is not equivalent to proving that Kilian's protocol is probabilistically special sound for any nontrivial relation. This is because the lower bound of $k/2N$ on $\mathsf{WIN}[\mathbf{V}_{\mathsf{PCP}}, x](\pi)$ depends on the adversary, who is free to set N to be an arbitrary polynomial. In particular, the adversary we construct for this experiment will set N according to the (inverse) success probability of the malicious prover.

Lemma 8.2. *Let* $\mathsf{PCP} = (\mathbf{P}_{\mathsf{PCP}}, \mathbf{V}_{\mathsf{PCP}})$ *be a PCP system with proof length* $\ell = \mathrm{poly}(|x|, \lambda)$ *and alphabet* Σ *satisfying* $|\Sigma|$*, and let* $R = \{0,1\}^{\mathsf{rc}}$ *denote the set of random coins for* $\mathbf{V}_{\mathsf{PCP}}$*. Let* V^{Kilian} *denote the verifier for* $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$*. Let* $k = 6\ell \ln(2|\Sigma|)$*.*

If VC *is a vector commitment satisfying post-quantum* position-binding*, then for any efficient quantum adversary* Adv *obtaining at most* $N(\lambda)$ *samples from* $\mathsf{Sampler}$*,*

$$\Pr\left[\begin{array}{c|r} \forall i \neq j, r_i \neq r_j & \mathsf{ck} \leftarrow \mathsf{VC.Gen}(1^\lambda, \Sigma, \ell) \\ \wedge\, \forall i, r_i \in S & \mathsf{cm} \leftarrow \mathsf{Adv}(\mathsf{ck}) \\ \wedge\, \forall i, V^{\mathsf{Kilian}}(x, (\mathsf{ck}, \mathsf{cm}, r_i, z_i)) = 1 & \{(r_i, z_i)\}_{i\in[k]} \leftarrow \mathsf{Adv}^{\mathsf{Sampler}(R)} \\ \mathsf{WIN}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \leq k/(2N) & \pi \leftarrow \mathsf{Ext}^{\mathsf{Kilian}}(\mathsf{ck}, \mathsf{cm}, \{(r_i, z_i)\}_{i\in[k]}) \end{array}\right] \leq \mathrm{negl}(\lambda) \ .$$

where S *is the set of randomnesses obtained by* Adv *from* $\mathsf{Sampler}$*.*

Proof. We call an input $(\mathsf{ck}, \mathsf{cm}, \{(r_i, z_i)\}_{i\in[k]})$ to $\mathsf{Ext}^{\mathsf{Kilian}}$ *admissible* for a set S if it satisfies the conditions

$$\forall i \neq j, r_i \neq r_j \wedge \forall i, r_i \in S \wedge \forall i, V^{\mathsf{Kilian}}(x, (\mathsf{ck}, \mathsf{cm}, r_i, z_i)) = 1 \ .$$

We call an input *consistent* if it is admissible and $\mathsf{Ext}^{\mathsf{Kilian}}$ does not output $\perp$ on this input. By the position-binding property of VC, the probability that Adv outputs an admissible input that is not consistent is at most $\mathrm{negl}(\lambda)$.

We now argue that the probability over S and $\{(r_i, z_i)\}_{i\in[k]}$ that $\{(r_i, z_i)\}_{i\in[k]}$ is consistent and $\mathsf{Ext}^{\mathsf{Kilian}}$ outputs $\pi \in B = \{\pi : \mathsf{WIN}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \leq k/(2N)\}$ is $\mathrm{negl}(\lambda)$. Consider a fixed string π such that $\mathsf{WIN}[\mathbf{V}_{\mathsf{PCP}}, x](\pi) \leq k/(2N)$. The probability that $\mathsf{Ext}^{\mathsf{Kilian}}$ outputs π is bounded by the probability that the answers given in $\{(z_i)_{i\in[k]}\}$ are consistent with π, which in turn is bounded by the probability that for all r_i, $\mathbf{V}^{\pi}_{\mathsf{PCP}}(x; r_i) = 1$.

Since the r_i are contained in a random set $S \subseteq R$ of size at most N, it suffices to bound the probability over S that at least k members of S have this property. By assumption, $\frac{1}{R}\sum_{r\in R} \mathbf{V}^{\pi}_{\mathsf{PCP}}(x; r) \leq k/(2N)$. Hence by a multiplicative Chernoff bound (Proposition 3.2),

$$\Pr\left[\sum_{r\in S} \mathbf{V}^{\pi}_{\mathsf{PCP}}(x; r) \geq k\right] \leq e^{-k/6} = (2|\Sigma|)^{-\ell} \ .$$

By a union bound over all $\pi \in B$ (in particular, $|B| \leq |\Sigma|^{\ell}$) we have that the probability that, on a consistent input, Ext outputs any $\pi \in B$ is at most $1/2^{\ell} = \mathrm{negl}(\lambda)$. □

Corollary 8.3. *Let* $\mathsf{PCP} = (\mathbf{P}_{\mathsf{PCP}}, \mathbf{V}_{\mathsf{PCP}})$ *be a PCP system, and let* ε *be a negligible function. If* VC *is a vector commitment satisfying post-quantum position-binding, then* $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ *is probabilistically special sound for the relation*

$$\mathfrak{R}_{\mathsf{PCP},\varepsilon}(\lambda) \coloneqq \{(x, \pi) : \Pr_r[\mathbf{V}^{\pi}_{\mathsf{PCP}}(x; r) = 1] > \varepsilon(\lambda)\} \ .$$

Proof. Fix some adversary Adv for PSSExp running in time $t(\lambda) = \mathrm{poly}(\lambda)$; note that the number of samples $N(\lambda)$ that the adversary can obtain from $\mathsf{Sampler}$ is at most $t(\lambda)$. Thus by Lemma 8.2 the probability that the extractor fails to output π that is not a witness for x in the relation $\mathfrak{R}_{\mathsf{PCP}}$ is at most $\mathrm{negl}(\lambda)$, since for all large enough λ, $\varepsilon(\lambda) < k(\lambda)/(2N(\lambda))$. □

8.3 Proof of Theorem 8.1

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

Claim 8.4. *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. □

By the above claim and Corollary 8.3, we have that for every negligible function ε, $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ is collapsing and probabilistically k-special sound for $\mathfrak{R}_{\mathsf{PCP},\varepsilon}(\lambda)$. Hence by Theorem 6.1, $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ is a post-quantum proof of knowledge for $\mathfrak{R}_{\mathsf{PCP},\varepsilon}(\lambda)$.

Setting $\varepsilon \coloneqq \varepsilon_{\mathsf{PCP}}$, we have that for all $x \notin \mathcal{L}(\mathfrak{R})$ and $\lambda \in \mathbb{N}$, $x \notin \mathcal{L}(\mathfrak{R}_{\mathsf{PCP},\varepsilon_{\mathsf{PCP}}}(\lambda))$, which implies that $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ is a post-quantum argument for $\mathfrak{R}$.

Moreover, if κ_{PCP} is negligible then we obtain an extractor for Kilian for $\mathfrak{R}$ by composing the extractor for Kilian for $\mathfrak{R}_{\mathsf{PCP},\kappa_{\mathsf{PCP}}}(\lambda)$ with the PCP extractor $\mathbf{E}$. The PCP of knowledge guarantee implies that whenever the Kilian extractor outputs a proof π, $(x, \mathbf{E}(x, \pi)) \in \mathfrak{R}$. Hence $\mathsf{Kilian}[\mathsf{PCP}, \mathsf{VC}]$ is a post-quantum argument of knowledge for $\mathfrak{R}$.

Acknowledgements

FM thanks Justin Holmgren for helpful discussions. 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. NS thanks Dominque Unruh for helpful discussions. NS is supported by DARPA under Agreement No. HR00112020023.

References

[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 *Public Key Cryptography*, volume 7778 of *Lecture Notes in Computer Science*, pages 55–72. Springer, 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.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[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.

[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.

[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.