Computationally binding quantum commitments

Dominique Unruh
University of Tartu

April 21, 2015

Abstract. We present a new definition of computationally binding commitment schemes in the quantum setting, which we call "collapse-binding". The definition applies to string commitments, composes in parallel, and works well with rewinding-based proofs. We give simple constructions of collapse-binding commitments in the random oracle model, giving evidence that they can be realized from hash functions like SHA-3. We evidence the usefulness of our definition by constructing three-round statistical zero-knowledge quantum arguments of knowledge for all NP languages.

Contents

1 Introduction

We study the definition and construction of computationally binding string commitment schemes in the quantum setting. A commitment scheme is a two-party protocol consisting of two phases, the commit and the open phase. The goal of the commitment is to allow the sender to transmit information related to a message m during the commit phase in

such a way that the recipient learns nothing about the message (hiding property). But at the same time, the sender cannot change his mind later about the message (binding property). Later, in the open phase, the sender reveals the message m and proves that this was indeed the message that he had in mind earlier. We will focus on non-interactive classical commitments, that is, the commit and open phase consists of a single classical message. However, the adversary who tries to break the binding or hiding property will be a quantum-polynomial-time algorithm. At the first glance, it seems that the definition of the binding property in this setting is straightforward; we just take the classical definition but consider quantum adversaries instead of classical ones:

Definition 1 (Classical-style binding – informal) *No quantum-polynomial-time algorithm A can output, except with negligible probability, a commitment c (i.e., the message sent during the commit phase) as well as two openings u, u' that open c to two different messages m, m'.*

(Formal definition in Section 2.) Unfortunately, this definition turns out to be inadequate in the quantum setting. Ambainis, Rosmanis, and Unruh [ARU14] show the existence of a commitment scheme (relative to a special oracle) such that: The commitment is classical-style binding. Yet there exists a quantum-polynomial-time adversary A that outputs a commitment c, then expects a message m as input, and then provides valid opening information for c and m. That is, the adversary can open the commitment c to any message of his choosing, even if he learns that message only after committing. This is in clear contradiction to the intuition of the binding property. How is this possible, as Definition 1 says that the adversary cannot produce two different openings for the same commitment? In the construction from [ARU14], the adversary has a quantum state $|\Psi\rangle$ that allows him to compute one opening for a message of his choosing, however, this computation will destroy the state $|\Psi\rangle$. Thus, the adversary cannot compute two openings simultaneously, hence the commitment is classically-binding. But he can open the commitment to an arbitrary message once, which shows that the commitment scheme is basically useless despite being classically-binding.[1]

1.1 Prior definitions

We now discuss various definitions that appeared in the literature and that circumvent the above limitation of the classical-binding property. (We do not discuss the hiding property here, because that one does not have any comparable problems. See Definition 8 below for the definition of hiding.) In each case, we discuss some limitations of the definitions to motivate the need for a new definition for computationally binding commitments. The reader only interested in our results can safely skip this section.

Sum-binding. The most obvious solution is to simply require that the adversary cannot open successfully to each of two messages: That is:

[1] Note that for classical adversaries, the classical-binding property gives useful guarantees: If an adversary can produce an opening for any message m using some classical algorithm, he can also produce two openings for different messages m, m' by running that algorithm twice.

Definition 2 (Sum-binding – informal) *Consider a bit commitment scheme. (I.e., one can only commit to $m = 0$ or $m = 1$.)*

Given an adversary A, let p_b be the probability that the recipient accepts in the following execution: A commits, then A is given b, and then A provides opening information for message b.

A commitment is sum-binding *iff for any quantum-polynomial-time adversary A, $p_0 + p_1 \leq 1 + \textit{negligible}$.*

Note that even with an ideal commitment, $p_0 + p_1 = 1$ is possible (the adversary just picks $b := 0$ in the commit phase with probability p_0, and $b := 1$ else). So $p_0 + p_1 \leq 1 + \textit{negligible}$ is the best we can expect if we allow for a negligible probability of an attack. The sum-binding definition has occurred implicitly and explicitly in different variants in [BCJL93, May97, DMS00, CDMS04, CSST11]. We use the name sum-binding here to distinguish it from the other definitions of binding discussed here since it does not have an established name.

Although it avoids the attack described above, the sum-binding definition has a number of disadvantages:

- It is specific to the bit commitment case. There is no straightforward generalization to the the string commitment case (i.e., where the message m does not have to be a single bit). See [CDMS04] for discussion why obvious approaches fail.[2]
- It is unclear how the definition behaves when we use the commitment several times. (I.e., it is not clear how it behaves under composition.) For example, given bits $m_1, \ldots, m_n$, what are the security guarantees if we commit to each of the m_i? (Be it in parallel, or sequentially.) Basically, we would expect that all commitments together form a binding commitment on the string $m = m_1 \ldots m_n$, but this is something we cannot even express using the sum-binding definition.
- It is not clear how useful sum-binding commitments are as subprotocols in larger protocols. That is, is the sum-binding property strong enough to allow to prove the security of complex protocols using commitments? While there are constructions of sum-binding in the literature (e.g., [DMS00]), we are not aware of research where (computational) sum-binding commitments are used as subprotocols.

CDMS-binding. Crépeau, Dumais, Mayers, and Salvail [CDMS04] suggest a generalization of the sum-binding property to string commitments. The basic idea is: Instead of bounding $p_0 + p_1 \leq 1 + \textit{negligible}$ where p_m is the probability that the adversary open his commitment as $m \in \{0,1\}$, we could bound $\sum_m p_m \leq 1 + \textit{negligible}$ where m ranges over all bitstrings. However, as discussed in [CDMS04], this would be too strong a requirement. (Basically, this is because the sum $\sum_m p_m$ has exponentially many summands, so even

[2] One obvious attempt would be: Let p_m be the probability that A opens the commitment as m when given m after the commit phase. Then for all m_0, m_1, we have $p_{m_0} + p_{m_1} \leq 1 + \textit{negligible}$.

However, this leaves the possibility that the adversary A achieves the following: In the commit phase, A outputs c, m_0, m_1 where m_0, m_1 are uniformly distributed. Then A gets a bit b. Then A opens c with message m_b. This should not be possible if c is binding, yet for this A, p_m is negligible for any fixed m. (Since $\Pr[m \in \{m_0, m_1\}]$ is negligible.)

negligible attack probabilities can add up to large probabilities.) Instead, they proposed the following definition:

Definition 3 (CDMS-binding – informal) *Let F be a family of functions. Fix a string commitment scheme. For $f \in F$, let $\tilde{p}_y^f$ be the probability that the recipient accepts in the following execution: A commits. A gets y. A tries to open the commitment to some m with $f(m) = y$.*

We call the commitment scheme F-CDMS-binding iff for all adversaries A and all $f \in F$, we have $\sum_y \tilde{p}_y^f \leq 1 + \text{negligible}$.

Now if all $f \in F$ have a polynomial-size range, the sum $\sum_y \tilde{p}_y^f$ will have polynomially many summands. The intuition behind this definition is that every function $f \in F$ represents some property of the committed message m (e.g., $f(m)$ is the parity of m). Then, if a commitment scheme is F-CDMS-binding, this intuitively means that the although the adversary might be able to change his mind about the message m, he cannot change his mind about $f(m)$. (E.g., if the parity function is in F, this means that the adversary will be committed to the parity of the message m.) [CDMS04] successfully used this definition (for a specific class F) to show that using quantum communication and a commitment, we can construct an oblivious transfer protocol. (Note however that their protocol is different and more complex than the original OT protocol from [BBCS91].)

Although the CDMS-binding definition generalizes the sum-binding definition to the case of string commitments, it comes with its own challenges:

- The definition is parametrized by a specific family F of functions that specifies in which way the commitment should be binding. This function family has to be chosen dependent on the particular use case. This makes the definition less universal and canonical.
- To the best of our knowledge, no construction of CDMS-binding commitments is known. Crépeau et al. [CDMS04] conjecture that the protocol from [CLS01] can be extended to a CDMS-binding one for functions F with small range, but no proof or construction is given.
- It is not known whether the definition is composable. If we commit to messages $m_1, \ldots, m_n$ individually using F-CDMS-binding commitments, does this constitute an F'-CDMS-binding commitment on $m := m_1 \| \ldots \| m_n$? If so, for which F'?
- While CDMS-binding commitments have successfully been used in a larger protocol (namely, the OT protocol from [CDMS04]), we believe that in many contexts, the definition is still not very easy to use. At least in classical cryptography, one often uses the fact that it is possible to extract the committed message by rewinding (basically, one runs the open phase, saves the opened message, and rewinds to before the opening phase). It is not clear how to do that with CDMS-binding commitments. For example, it is not clear how one could use CDMS-binding commitments in the construction of sigma-protocols that are quantum arguments of knowledge (as done in Section 7 below using our definition of binding commitments).

Perfectly-binding commitments. One possibility to solve all the problems mentioned so far is simply to use perfectly-binding commitments.

Definition 4 (Perfectly-binding – informal) *A commitment scheme is perfectly-binding if there exists no tuple (c, m, u, m', u') with $m \neq m'$ such that u is a valid opening for c with message m, and u' is a valid opening for c with message m'.*

However, if we restrict ourselves to perfectly-binding commitments, we get the following disadvantages:

- A perfectly-binding commitment cannot be statistically hiding [May97]. That is, the hiding property cannot hold against computationally unlimited adversaries. That means that we give up on information-theoretical security for one party just because we do not have a suitable definition for the computationally-binding property. For example, the constructions in [Unr12] are only computational zero-knowledge (not statistical zero-knowledge) because perfectly-binding commitments are used.
- Perfectly-binding commitments cannot be short. That is, the length of the commitment must be as long as the length of the committed message. So by using only perfectly-binding commitments, we may lose efficiency.

UC commitments. One further possibility is to use commitments that are UC-secure [Unr10]. Since the security of a protocol using a UC-secure commitment can be reduced to the security of the same protocol using an ideal (in particular perfectly-binding) commitment, UC-secure commitments are easy to use. Yet, this solution again comes with disadvantages:

- UC-commitments do not exist without the use of additional setup such as, e.g., a common reference strings (CRS). It is possible to chose the CRS in a pre-computation phase using a coin-toss protocol [DL09]. But that increases the round complexity of the resulting protocol (and, incidentally, loses the UC security and possibly even the concurrent composability of the resulting protocol).
- In the construction of UC-secure commitment schemes, trapdoors are used that allow the simulator to extract the committed message. This implies that constructions of UC-secure commitment are usually more complex, less efficient, and use stronger computational assumptions.
- At least when using a CRS, UC commitments cannot be short.

Damgård, Fehr, Lunemann, Salvail, and Schaffner [DFL$^+$09] use so-called dual-mode commitments, these are somewhat weaker than UC commitments. Yet, they also use extraction using a trapdoor in the CRS. Hence the disadvantages of UC commitments apply to dual-mode commitments as well.

1.2 Our contribution

We give a new definition for the computational-binding property for commitment schemes, called "collapse-binding" (Section 2). This definition is composable (several collapse-binding commitments are also collapse-binding together), works well with quantum

rewinding (see below), does not conflict with statistical hiding (as perfectly-binding commitments would), allows for short commitments (i.e., the commitment can be shorter than the committed message, in contrast to perfectly-binding commitments, and to extractable commitments in the CRS model). Basically, collapse-binding commitments seem to be in the quantum setting what computationally-binding commitments are in the classical setting.

We show that collision-resistant hash functions are not sufficient for getting collapse-binding or even just sum-binding commitments (Section 3), at least when using standard constructions, and relative to an oracle. We present a strengthening of collision-resistant hash functions, "collapsing hash functions" that can serve as a drop-in replacement for collision-resistant hash functions (Section 4). Using collapsing hash functions, we show several standard constructions of commitments to be collapse-binding (Section 5).

We conjecture that standard cryptographic hash functions such as SHA-3 [NIS14] are collapsing (and thus lead to collapse-binding commitments). We give evidence for this conjecture by proving that the random oracle is a collapsing hash function.

We show that the definition of collapse-binding commitments is usable by extending the construction of quantum proofs of knowledge from [Unr12] (Section 7). Their construction uses perfectly-binding commitments (actually, strict-binding, which is slightly stronger) to get proofs of knowledge. We show that when replacing the perfectly-binding commitments with collapse-binding ones, we get statistical zero-knowledge quantum arguments of knowledge. In particular, this shows that collapse-binding commitments work well together with rewinding.

1.3 Our techniques

Collapse-binding commitments. To explain the definition of collapse-binding commitments, first consider a perfectly-binding commitment. That is, when an adversary A outputs a commitment c, there is only one possible message m_c that A can open c to. Hence, if the adversary A outputs a superposition of messages that he can open c to, that superposition will necessarily be in the state $|m_c\rangle$. Hence, we can characterize perfectly-binding commitments by requiring: when an adversary outputs a superposition of messages that he can open the commitment c to, that superposition will necessarily be a single computational basis vector (i.e., no non-trivial superposition).

To express this more formally, consider the circuit in Figure 1 (a). Here the adversary A outputs a commitment c (classical message). Furthermore, he outputs three quantum registers S, U, M. S contains his state. M is supposed to contain a superposition of messages, U a superposition of corresponding opening informations. Then we

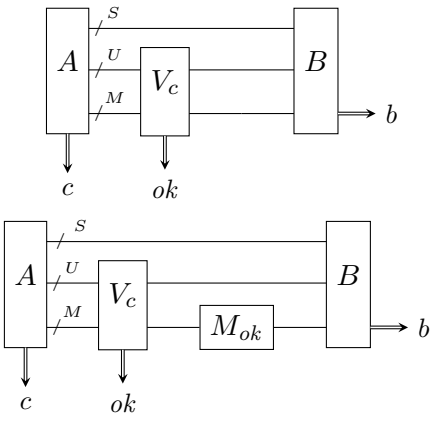


Figure 1: Games from the definition of collapse-binding commitments.

apply the measurement V_c. This measurement measures whether U, M contain matching opening information/message. More formally, V_c measures whether U, M is a superposition of states $|u, m\rangle$ such that u is valid opening information for message m and commitment c. Let $ok = 1$ if the measurement succeeds. Then we feed the registers S, U, M back to the second part B of the adversary. B outputs a classical bit b. As discussed before, a commitment is perfectly-binding iff for all adversaries A, the state of M after measuring $ok = 1$ is a computational basis vector.

The state of a register is a computational basis vector (or, synonymously: is in a collapsed state) iff measuring that register in the computational basis does not change that state. Consider the circuit in Figure 1 (b). Here we added a measurement M_{ok} on M after V_c. M_{ok} is a complete measurement in the computational basis, but is executed only if $ok = 1$. Since M_{ok} disturbs the state of M iff that state is not a computational basis vector, we can rephrase the definition of perfectly-binding commitments:

A commitment is perfectly-binding iff, for all computationally unlimited adversaries A, B, $\Pr[b = 1]$ is equal in Figures 1 (a) and 1 (b) where b is the output (i.e., guess) of B.[3]

Now we are ready to weaken this characterization to get a computational binding property. Basically, we require that the same holds for quantum-polynomial-time adversaries:

Definition 5 (Collapse-binding – informal) *A commitment is collapse-binding iff, for all quantum-polynomial-time adversaries A, B, $\Pr[b = 1]$ in Figure 1 (a) is negligibly close to $\Pr[b = 1]$ in Figure 1 (b).*

In other words, with a perfectly-binding commitment, the adversary cannot produce a superposition of different messages that are contained in the commitment. But with a collapse-binding commitment, the adversary is forced to produce a state *that looks like it is not a superposition* of different messages. For the purpose of computational security, this will often be as good.

We quickly explain why collapse-binding commitments work well with quantum rewinding. In the case of quantum rewinding (e.g., in the analysis of proofs of knowledge [Unr12]), one problem is that we might need to run an adversary until he opens a commitment c, then to measure the opened message, and then to go back to an earlier state by applying the inverse of the adversary. The problem is that measuring the opened message will disturb the state of the adversary, and thus make rewinding impossible. Except: if the opened message cannot be distinguished from being already in a collapsed state (as guaranteed by collapse-binding), then measuring the opened message does not disturb the state in a noticeable way and we can rewind. (See the discussion on arguments of knowledge below.)

Constructing collapse-binding commitments. Collapse-binding commitments are useful only if they exist. Perfectly-binding commitments are easily seen to be collapse-binding, but then we cannot have statistically hiding or short commitments. In the

[3]Our exposition above was not very rigorous, but it is easy to see that this is indeed an "if and only if".

classical setting, we get practical computationally-binding commitments from a collision-resistant hash function H. The most obvious construction is to send $c := H(m\|u)$ for uniformly random u of suitable length. We call this the "canonical commitment". The canonical commitment is easily seen to be classical-style binding if H is collision-resistant, and it is statistically hiding if H is a random oracle. To get rid of the random-oracle requirement, we can use a somewhat more complex constructions by Halevi and Micali [HM96] instead. Unfortunately, both the canonical commitment and the Halevi-Micali commitments are not collapse-binding if H is merely collision-resistant. In fact, relative to a specific oracle and using a specific collision-resistant hash function, there is a total break where the adversary can unveil the commitment to any message of his chosing. To show this, we tweak the technique from [ARU14] to construct a hash function H such that the adversary can sample an image c of H together with a quantum state $|\Psi\rangle$ such that: Given the state $|\Psi\rangle$, for any m, the adversary can find a random u with $H(m\|u) = c$. But this process destroys $|\Psi\rangle$, so the adversary cannot find two preimages of c; the hash function is collision-resistant. But the canonical commitment, based on this H, is trivially broken. Similar constructions break the Halevi-Micali commitments.

Since collision-resistance seems too weak a property in the quantum setting (at least for our purposes), we give a strengthening of collision-resistance: collapsing hash functions:

Definition 6 (Collapsing hash function – informal) *An adversary is valid if he outputs a classical value c, and a register M containing a superposition of messages m with $H(m) = c$. We call H collapsing iff no quantum-polynomial-time adversary can distinguish whether we measure M in the computational basis or not, before giving the register M back to the adversary. (This is formalized with games similar to those in Figure 1.)*

We can show that collapsing hash functions are collision-resistant, and they share a number of structural properties with collision-resistant functions. E.g., injective functions are collapsing, and the composition $H \circ H'$ of collapsing functions is collapsing.

Due to the similarity between the definition of collapsing hash functions and collapse-binding commitments, we can show that the canonical commitment and the Halevi-Micali commitments are collapse-binding if H is collapsing.

However, this leaves the question: do collapsing functions exist in the first place? We conjecture that common industrial hash function like SHA3 [NIS14] are actually collapsing (not only collision-resistant). In fact, we argue that the collapsing property should be a requirement for the design of future hash functions (in the sense that a hash function where the collapsing property is in doubt should not be selected for industry standards), since collision-resistance is not sufficient if we wish to achieve post-quantum secure cryptography. We support our conjecture that sufficiently unstructured functions are collapsing by proving that the random oracle is collapsing:

Random oracles are collapsing. We now sketch on a high level our proof that random oracles are collapsing, or, equivalently, that a random function is collapsing

with high probability. In our analysis, we assume that the adversary can query the random oracle on the superposition of different inputs; this is necessary for having a realistic modeling of hash functions [BDF+11]. As a first step, we identify a new property, "half-collision resistance":

Definition 7 (Half-collision resistance – informal) *A half-collision of H is a string x such that there exists an $x' \neq x$ with $H(x') = H(x)$. A hash function H is half-collision resistant if not adversary can: Output a half-collision with non-negligible probability. And never output a non-half-collision. (The adversary may output $\perp$ though.)*

That is, half-collision resistance says that the adversary cannot find non-injective inputs to H without sometimes accidentally outputting injective inputs. We show: if H is half-collision resistant, it is collapsing.

The proof idea is: if H is not collapsing, the adversary can produce a superposition M of messages m with $H(m) = c$ and notice whether M is being measured. The latter implies that M must be a superposition of at least two messages m with $H(m) = c$. Hence by measuring M, the adversary gets a half-collision. Much additional work is needed to make sure that the adversary does not accidentally measure the register M when it is not a nontrivial superposition.

(The half-collision resistance property might be useful independent of the proof that the random oracle is collapsing. When trying to construct collapsing hash functions based on other assumptions, half-collision resistance might be easier to verify since its definition consists of purely classical games.)

Next we construct a random function $H^* : X \to Y$ with $|Y| = \frac{2}{3}|X|$. That is, H^* is slightly compressing. The domain of H^* is partitioned into two sets X_1, X_2 with $|X_1| = 2|X_2|$. H^* is injective on X_2, and 2-to-1 on X_1. Besides those constraints, H^* is uniformly random. We can then show that H^* is half-collision resistant. (Basically, this means that the adversary cannot identify the subset X_1.) Furthermore, we can show that H^* is indistinguishable from a random function $H : X \to Y$. The latter fact is shown by step-wise rewriting of the definition of H^* until we reach H (crucially using the fact that random functions and random injections are indistinguishable [Zha13]). Since H^* is half-collision resistant, it is collapsing. And since H is indistinguishable from H^*, H is collapsing.

We now know that random functions $H : X \to Y$ are collapsing if $|Y| = \frac{2}{3}|X|$ (i.e., if they are slightly compressing). However, we want that H is collapsing for arbitrary X and Y, as long as Y has superpolynomial size. For $|X| \leq |Y|$, H is indistinguishable from a random injection, which in turn is collapsing. The interesting case is $|X| > |Y|$ (namely, when H is compressing). In this case, we show (following an idea from [Zha13]) that H can be written as $H = f_n \circ \cdots \circ f_1$ where all f_i are slightly compressing. (Some technical care is needed when $|Y|/|X|$ is not a power of $\frac{2}{3}$.) Since all f_i are collapsing, so is H. This shows that a random function H is collapsing, in other words, that the random oracle is collapsing (if its range has superpolynomial size).

Quantum arguments of knowledge. We illustrate the use of collapse-binding commitments by revisiting the construction of proofs of knowledge from Unruh [Unr12].

Unruh showed that a sigma-protocol (i.e., a particular kind of three round proof system) is a quantum proof of knowledge if it has two properties: *special soundness* (from two interactions with the same first and different second messages one can efficiently compute a witness) and *strict soundness* (the first and second message of a valid interaction determine the third). In the classical setting, only special soundness is needed. In the quantum setting, strict soundness is additionally required to allow for quantum rewinding: In the proof from [Unr12], we run the malicious prover to get his response (the third message). Then we measure the response. Then we rewind the prover (by applying the inverse of the unitary transformation representing the prover). Then we run the prover again to get a second answer. Special soundness then implies that from the two responses, we get a witness. However, we need to make sure that measuring the prover's response before rewinding does not disturb the state (too much). In [Unr12], this follows from strict soundness: strict soundness guarantees that the response is uniquely determined, and thus measuring the response does not disturb the state. To achieve strict soundness, [Unr12] lets the prover commit to all possible responses in the first message using perfectly-binding commitments.[4] The drawback of this solution is that the commitments cannot be statistically hiding, so we cannot get statistical zero-knowledge proofs using the method from [Unr12].

What happens if we replace the perfectly-binding commitments by collapse-binding commitments containing the response? In that case, the response will not necessarily be information-theoretically determined by the first two messages. However, the definition of collapse-binding commitments guarantees that measuring that response will be indistinguishable from not measuring it. Thus, if we measure the response, the state might be disturbed, but it will be computationally indistinguishable from not being disturbed. This is enough for the proof technique from [Unr12] to go through when using collapse-binding commitments, assuming the prover is computationally limited. The resulting protocol will not be a quantum proof of knowledge, but a quantum argument of knowledge (i.e., secure only against computationally limited provers). But in contrast to [Unr12], the proof system will be statistical zero-knowledge.

To summarize: from collapse-binding commitments (or from collapsing hash functions), we get three-round statistical zero-knowledge quantum arguments of knowledge for all languages in NP (with inverse polynomial knowledge error). To the best of our knowledge, not even three-round statistical zero-knowledge quantum *arguments* were known before.

1.4 Related work.

Commitments. Brassard, Crépeau, Jozsa, and Langlois [BCJL93] presented an information-theoretically hiding and binding commitment scheme using quantum communication. However, the protocol was flawed, Mayers [May97] showed that information-theoretically hiding and binding commitments are impossible. (This is no contradiction to our results, because our commitments are not information-theoretically binding.) Dumais, Mayers, and Salvail [DMS00] and Crépeau, Légaré, and Salvail [CLS01] constructed

[4] Actually, "strict-binding commitments" but this distinction is not relevant for this exposition.

statistically hiding commitments from quantum one-way permutations/functions, respectively. Their protocols use quantum communication, and are sum-binding. Crépeau, Dumais, Mayers, and Salvail [CDMS04] generalized the sum-binding definition to string commitments and constructed an OT protocol based on that definition. (However, it is not known whether the protocol composes even sequentially.) Damgård, Fehr, Lunemann, Salvail, and Schaffner [DFL+09] and Unruh [Unr10] showed a much simpler OT protocol to be secure, assuming much stronger commitment definitions in the CRS model, but achieving stronger security notions (sequential composability/UC). Ambainis, Rosmanis, and Unruh [ARU14] show that classical-style binding commitments are not necessarily even sum-binding.

Quantum random oracles. Random oracles were first explicitly considered in a quantum cryptographic context by Boneh, Dagdelen, Fischlin, Lehmann, Schaffner, and Zhandry [BDF+11] who stressed that the adversary should have superposition access to the random oracle. Zhandry [Zha13] showed that the random oracle is collision-resistant. In contrast, we show (based on his result) that the random oracle is collapsing (a stronger property).

Quantum rewinding and proof systems. Watrous [Wat09] showed how quantum rewinding can be used to prove the security of quantum zero-knowledge protocols. Unruh [Unr12] showed how a different flavor of quantum rewinding can be used for proving the security of quantum proofs of knowledge; we extend their technique to quantum arguments of knowledge. Unruh [Unr14a] constructs non-interactive computational zero-knowledge quantum arguments of knowledge in the random oracle model.

2 Definitions and basic properties

Preliminaries. For the necessary background in quantum computing, see, e.g., [NC10]. By $|i\rangle$ with $i \in I$ we denote the vectors of the computational basis of the Hilbert space with dimension $|I|$. We also use the symbol $|\cdot\rangle$ to refer to other (non-basis) vectors (e.g., $|\Psi\rangle$). And $\langle\Psi|$ is the conjugate transpose of $|\Psi\rangle$. $\|x\|$ refers to the Euclidean or ℓ^2-norm. We only consider finite dimensional Hilbert spaces. We denote $|+\rangle := \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$ and $|-\rangle := \frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle$. For a linear operator A on a Hilbert space, we denote by $A^\dagger$ its conjugate transpose. We denote by I the identity. We call an operator A on a Hilbert space a projector iff it is an orthogonal projector, i.e., a linear map with $P^2 = P$ and $P = P^\dagger$. By $\mathrm{TD}(\rho, \rho')$ we denote the trace distance between ρ and ρ', and by $F(\rho, \rho')$ the fidelity.

Given an algorithm A, let $x \leftarrow A(y)$ denote the result of running A with inputs y, and assigning the output to x. Let $x \xleftarrow{\$} M$ denote assigning a uniformly random element of M to x. We will use η to denote the security parameter, that is a positive integer that will be passed to all algorithms and adversaries and that indicates the required security level. By $a\|b$ we denote the concatenation of bitstrings a and b.

We call an algorithm quantum-polynomial-time if it is a quantum algorithm and its runtime is bounded by a polynomial in its input length with probability 1. We call an algorithm classical-polynomial-time if it performs only classical operations and its runtime is bounded by a polynomial in its input length with probability 1. We write 1^η for a bitstring (of 1's) of length η. (The latter is useful for making algorithms run in polynomial-time in the length of the security parameter, e.g., $A(1^\eta)$ will run polynomial-time in η.)

Commitments. A commitment scheme $(com, verify)$ consists of a quantum-polynomial-time algorithm com and a deterministic quantum-polynomial-time algorithm $verify$. $(c, u) \leftarrow com(1^\eta, m)$ returns a commitment c and the opening information u for the message m and security parameter η. c alone is supposed not to reveal anything about m (hiding). To open, we send (m, u) to the recipient who checks whether $verify(1^\eta, c, m, u) = 1$. Both com and $verify$ have classical input and output. com has a well-defined message space MSP_η that also depends on the security parameter η (e.g., $\{0,1\}^\eta$). Furthermore, for technical reasons, we assume that it is possible to find triples (c, m, u) with $verify(1^\eta, c, m, u) = 1$ with probability 1 in quantum-polynomial-time in η.[5]

We first state some standard properties of commitments.

Definition 8 *Let $(com, verify)$ be a commitment scheme. We define:*

- ***Perfect completeness:*** *$(com, verify)$ has* perfect completeness *iff for all $m \in \mathsf{MSP}_\eta$, $\Pr[verify(1^\eta, c, m, u) = 1 : (c, u) \leftarrow com(1^\eta, m)] = 1$.*
- ***Computational hiding:*** *$(com, verify)$ is* computationally hiding *iff for any quantum-polynomial-time A and any polynomial ℓ, there is a negligible μ such that for any η, any $m_0, m_1 \in \mathsf{MSP}_\eta$ with $|m_0|, |m_1| \leq \ell(\eta)$, and any $|\Psi\rangle$, $\left|P_0 - P_1\right| \leq \mu(\eta)$ where $P_i := \Pr[b = 1 : (c, u) \leftarrow com(1^\eta, m_i), b \leftarrow A(1^\eta, |\Psi\rangle, c)]$.*
- ***Statistical hiding:*** *Like computational hiding, except that we quantify over all A (not just quantum-polynomial-time A).*

Definition 9 (Classical-style binding) *A commitment scheme is* classical-style binding *iff for any quantum-polynomial-time algorithm A, the following is negligible in η:*

$$\Pr[verify(1^\eta, c, m, u) = 1 \wedge verify(1^\eta, c, m', u') = 1 \wedge m \neq m' : (c, m, u, m', u') \leftarrow A(1^\eta)]$$

Definition 10 (Collapse-binding) *For algorithms A, B, consider the following games:*

$$\begin{aligned}
&\mathsf{Game}_1: \quad (S, M, U, c) \leftarrow A(1^\eta),\ ok \leftarrow V_c(M, U),\ m \leftarrow M_{ok}(M),\ b \leftarrow B(1^\eta, S, M, U)\\
&\mathsf{Game}_2: \quad (S, M, U, c) \leftarrow A(1^\eta),\ ok \leftarrow V_c(M, U),\ b \leftarrow B(1^\eta, S, M, U)
\end{aligned}$$

Here S, M, U are quantum registers. V_c is a measurement whether M, U contains a valid opening, formally V_c is defined through the projector $\sum_{\substack{m,u \\ verify(1^\eta,c,m,u)=1}} |m\rangle\langle m| \otimes |u\rangle\langle u|$.

[5] This technical condition is necessary, e.g., for Definition 11 below. Without this condition, it is not clear that "valid" adversaries exist at all. Note that a commitment scheme with quantum-polynomial-time com and perfect completeness will always satisfies this technical condition: to find c, m, u, simply set $m := 0$ and compute $(m, u) \leftarrow com(1^\eta, m)$.

M_{ok} is a measurement of M in the computational basis if $ok = 1$, and does nothing if $ok = 0$ (i.e., it sets $m := \bot$ and does not touch the register M).

A commitment scheme is collapse-binding *iff for any quantum-polynomial-time algorithms A, B, the difference $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ is negligible.*

Instead of measuring using V_c whether the adversary outputs a correct opening information, we can quantify only over adversaries that always output correct opening information. This leads to the following equivalent definition of collapse-binding commitments. This definition is often easier to handle when proving that a given scheme is collapse-binding.

Definition 11 (Collapse-binding – variant) *For algorithms A, B, consider the following games:*

$$\begin{aligned}
&\mathsf{Game}_1: && (S, M, U, c) \leftarrow A(1^\eta),\ m \leftarrow M_{comp}(M),\ b \leftarrow B(1^\eta, S, M, U)\\
&\mathsf{Game}_2: && (S, M, U, c) \leftarrow A(1^\eta),\ b \leftarrow B(1^\eta, S, M, U)
\end{aligned}$$

Here S, M, U are quantum registers. $M_{comp}(M)$ is a measurement of M in the computational basis.

We call an adversary (A, B) valid *if $\Pr[\mathit{verify}(c, m, u) = 1] = 1$ when running $(S, M, U, c) \leftarrow A(1^\eta)$ and measuring M, U in the computational basis to obtain m, u.*

A commitment scheme is collapse-binding *iff for any quantum-polynomial-time valid adversary (A, B), the difference $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ is negligible.*

Lemma 12 *A commitment scheme $(\mathit{com}, \mathit{verify})$ is collapse-binding with respect to Definition 10 iff it is collapse-binding with respect to Definition 11.*

Proof. To avoid confusion, we call the games from Definition 10 $\mathsf{Game}_1, \mathsf{Game}_2$, while we call those from Definition 11 $\mathsf{Game}'_1, \mathsf{Game}'_2$. And the adversary in Definition 11 (that is used in $\mathsf{Game}'_1, \mathsf{Game}'_2$) we call (A', B').

First, assume that there is an adversary (A', B') breaking Definition 11, i.e., $\mu := |\Pr[b = 1 : \mathsf{Game}'_1] - \Pr[b = 1 : \mathsf{Game}'_2]|$ is non-negligible. Let $(A, B) := (A', B')$. By definition of validity, the measurement V_c from Definition 10 will succeed with probability 1 in Game_1 and Game_2. Hence that measurement has no effect, and thus $\Pr[b = 1 : \mathsf{Game}_1] = \Pr[b = 1 : \mathsf{Game}'_1]$ and $\Pr[b = 1 : \mathsf{Game}_2] = \Pr[b = 1 : \mathsf{Game}'_2]$. Thus $|\Pr[b = 1 : \mathsf{Game}'_1] - \Pr[b = 1 : \mathsf{Game}'_2]| = \mu$ is non-negligible. Thus (A, B) also breaks Definition 10.

Now, consider an adversary (A, B) breaking Definition 10, i.e., $\nu := |\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]|$ is non-negligible. Construct (A', B') as follows: $A'(1^\eta)$ runs $(S, M, U, c) \leftarrow A(1^\eta)$. Then it applies $ok \leftarrow V_c(M, U)$. If $ok = 1$, A' returns (S, M, U, c). Otherwise, A' picks (c, m, u) with $\mathit{verify}(1^\eta, c, m, u) = 1$,[6] initializes M, U with $|m\rangle|u\rangle$, and S with $|\bot\rangle$, and outputs c. (We assume that $|\bot\rangle$ is orthogonal to any state that A

[6]This is efficiently possible with probability 1 by assumption, see page 12.

would produce.) And B' does the following: If $ok = 0$, then B' outputs 0. If $ok = 1$, B' executes B.

A' is valid by construction: If $ok = 1$, $\mathit{verify}(1^\eta, c, m, u) = 1$ with probability 1 when measuring M, U as m, u, because M, U is in the image of V_c. And if $ok = 0$, $\mathit{verify}(1^\eta, c, m, u) = 1$ by choice of c, m, u.

We easily see that

$$\begin{aligned}
0 &= \Pr[b = 1 : \mathsf{Game}'_1 | ok = 0] = \Pr[b = 1 : \mathsf{Game}'_2 | ok = 0] = 0 \\
\alpha &:= \Pr[b = 1 : \mathsf{Game}_1 | ok = 0] = \Pr[b = 1 : \mathsf{Game}_2 | ok = 0] \\
\beta &:= \Pr[b = 1 : \mathsf{Game}_1 | ok = 1] = \Pr[b = 1 : \mathsf{Game}'_1 | ok = 1] \\
\gamma &:= \Pr[b = 1 : \mathsf{Game}_2 | ok = 1] = \Pr[b = 1 : \mathsf{Game}'_2 | ok = 1] \\
\delta &:= \Pr[ok = 1 : \mathsf{Game}_1] = \Pr[ok = 1 : \mathsf{Game}'_1] \\
&\quad = \Pr[ok = 1 : \mathsf{Game}_2] = \Pr[ok = 1 : \mathsf{Game}'_2]
\end{aligned}$$

and from this we calculate

$$\begin{aligned}
&\big|\Pr[b = 1 : \mathsf{Game}'_1] - \Pr[b = 1 : \mathsf{Game}'_2]\big| \\
&\quad = \Big|\big(0(1-\delta) + \beta\delta\big) - \big(0(1-\delta) + \gamma\delta\big)\Big| = \Big|\big(\alpha(1-\delta) + \beta\delta\big) - \big(\alpha(1-\delta) + \gamma\delta\big)\Big| \\
&\quad = \big|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\big| = \nu
\end{aligned}$$

which is non-negligible. Thus (A', B') breaks Definition 11. □

Definition 10 guarantees that the adversary cannot distinguish whether the register M is measured or not. However, it is not immediately obvious what happens when we measure M partially (e.g., we measure just one qubit). Could it be that such a partial measurement will be noticed? We expect that this is not the case, since a partial measurement lies half-way between no measurement and a complete measurement. The following lemma confirms that intuition: If a commitment scheme is collapse-binding, then Definition 10 also holds for partial measurements. (Assuming that the partial measurement is performed in the computational basis and can be implemented by a polynomial-time circuit.)

Lemma 13 (Collapse-binding w.r.t. partial measurements) *For a commitment scheme $(com, verify)$, and for algorithms A, B, consider the following games:*

$$\begin{aligned}
&\mathsf{Game}_1: \quad (S, M, U, c, f) \leftarrow A(1^\eta),\ ok \leftarrow V_c(M, U),\ x \leftarrow M^f_{ok}(M),\ b \leftarrow B(1^\eta, S, M, U) \\
&\mathsf{Game}_2: \quad (S, M, U, c, f) \leftarrow A(1^\eta),\ ok \leftarrow V_c(M, U), \qquad\qquad\qquad\quad b \leftarrow B(1^\eta, S, M, U)
\end{aligned}$$

Here f is a Boolean circuit (with multiple-bit output). V_c is as in Definition 10. M^f_{ok} is a measurement of M that returns $f(m)$ where m is the content of M if $ok = 1$, and does nothing if $ok = 0$ (i.e., it sets $m := \bot$ and does not touch the register M). More formally, if $ok = 1$, M_f is the measurement defined by the projectors $P_x := \sum_{m:f(m)=x} |m\rangle\langle m|$ for all x in the range of f, and if $ok = 0$, M_f is defined by the single projector $P_\bot := I$.

If $(com, verify)$ *is collapse-binding, then for any quantum-polynomial-time adversary* (A, B)*, the difference* $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ *is negligible.*

Proof. We start with Game_1. It is easy to see that V_c and M^f_{ok} commute, and that V_c is idempotent. Thus $\Pr[b = 1 : \mathsf{Game}_1] = \Pr[b = 1 : \mathsf{Game}_3]$ with:

$$\mathsf{Game}_3: \quad (S, M, U, c, f) \leftarrow A,\ ok' \leftarrow V_c,\ x \leftarrow M^f_{ok'},\ ok \leftarrow V_c,\ b \leftarrow B$$

(We omit the inputs of the various algorithms and measurements since they are unchanged throughout the proof.) If we consider the first three operations $(A, V_c, M^f_{ok'})$ as a single adversary, we can apply the collapse-binding property of *com*. Thus $|\Pr[b = 1 : \mathsf{Game}_3] - \Pr[b = 1 : \mathsf{Game}_4]| = \varepsilon_1$ for some negligible ε_1 with:

$$\mathsf{Game}_4: \quad (S, M, U, c, f) \leftarrow A,\ ok' \leftarrow V_c,\ x \leftarrow M^f_{ok'},\ ok \leftarrow V_c,\ m \leftarrow M_{ok},\ b \leftarrow B$$

We can see that $V_c, M^f_{ok'}, M_{ok}$ all commute. Furthermore V_c is idempotent, so we get $\Pr[b = 1 : \mathsf{Game}_4] = \Pr[b = 1 : \mathsf{Game}_5]$ with:

$$\mathsf{Game}_5: \quad (S, M, U, c, f) \leftarrow A,\ ok \leftarrow V_c,\ m \leftarrow M_{ok},\ x \leftarrow M^f_{ok},\ b \leftarrow B$$

(Note that we replace $M^f_{ok'}$ by M^f_{ok}.) The outcome of M^f_{ok} is determined by the outcome of M_{ok}, we have $\Pr[b = 1 : \mathsf{Game}_5] = \Pr[b = 1 : \mathsf{Game}_6]$ with:

$$\mathsf{Game}_6: \quad (S, M, U, c, f) \leftarrow A,\ ok \leftarrow V_c,\ m \leftarrow M_{ok},\ b \leftarrow B$$

Since $(com, verify)$ is collapse-binding, we get $|\Pr[b = 1 : \mathsf{Game}_6] - \Pr[b = 1 : \mathsf{Game}_2]| = \varepsilon_2$ for negligible ε_2.

Thus, summarizing, $|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]| \leq \varepsilon_1 + \varepsilon_2$ is negligible. □

Another question that naturally arises is whether collapse-binding commitments parallel compose. That is, if we commit to values $m_1, \ldots, m_n$ with n commitments, does this give a collapse-binding commitment on $m := (m_1, \ldots, m_n)$? Note that such a property is not obvious. For example, to the best of our knowledge, no prior definition of a quantum computational binding property in the literature is known to have this property. For collapse-binding commitments, however, the next lemma shows that the parallel composition of several commitments is still collapse-binding.

Lemma 14 (Parallel composition) *Let* $(com, verify)$ *be a collapse-binding commitment with message space* M*. Let* $n = n(\eta)$ *be polynomially-bounded and quantum-polynomial-time computable integer.*

Let $(com^n, verify^n)$ *be the* n*-fold parallel composition of* $(com, verify)$*. That is, its message space is* M^p*. And* $com^n(m_1, \ldots, m_n)$ *computes* $(c_i, u_i) \leftarrow com(m_i)$ *for* $i = 1, \ldots, n$*, and returns* (c, u) *with* $c := (c_1, \ldots, c_n)$ *and* $u := (u_1, \ldots, u_n)$*. And* $verify^n((c_1, \ldots, c_n), (m_1, \ldots, m_n), (u_1, \ldots, u_n)) = 1$ *iff* $\forall i.\, verify(c_i, m_i, u_i) = 1$*.*

Then $(com^n, verify^n)$ *is collapse-binding.*

Proof. By Lemma 12, to show that $(com^n, verify^n)$ is collapse-binding, we need to show that for any valid adversary A against $(com^n, verify^n)$, $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ is negligible, with $\mathsf{Game}_1, \mathsf{Game}_2$ as follows:

$$\begin{aligned}
\mathsf{Game}_1: &\quad (S, M, U, c) \leftarrow A(1^\eta),\ m \leftarrow M_{comp}(M),\ b \leftarrow B(1^\eta, S, M, U)\\
\mathsf{Game}_2: &\quad (S, M, U, c) \leftarrow A(1^\eta), \qquad\qquad\qquad\quad b \leftarrow B(1^\eta, S, M, U)
\end{aligned}$$

Using the definition of $(com^n, verify^n)$, this is equivalent to:

$$\begin{aligned}
\mathsf{Game}_1: &\quad (S, M_1, \ldots, M_n, U_1, \ldots, U_n, c_1, \ldots, c_n) \leftarrow A(1^\eta),\\
&\quad m_i \leftarrow M_{comp}(M_i) \text{ for } i = 1, \ldots, n,\\
&\quad b \leftarrow B(1^\eta, S, M_1, \ldots, M_n, U_1, \ldots, U_n)\\
\mathsf{Game}_2: &\quad (S, M_1, \ldots, M_n, U_1, \ldots, U_n, c_1, \ldots, c_n) \leftarrow A(1^\eta),\\
&\quad b \leftarrow B(1^\eta, S, M_1, \ldots, M_n, U_1, \ldots, U_n)
\end{aligned}$$

And the validity of A implies for all i that measuring M_i, U_i will always return m_i, u_i with $verify(c_i, m_i, u_i) = 1$.

We define hybrid games for $i = 0, \ldots, n$:

$$\begin{aligned}
\mathsf{Hyb}_j: &\quad (S, M_1, \ldots, M_n, U_1, \ldots, U_n, c_1, \ldots, c_n) \leftarrow A(1^\eta),\\
&\quad m_i \leftarrow M_{comp}(M_i) \text{ for } i = 1, \ldots, j,\\
&\quad b \leftarrow B(1^\eta, S, M_1, \ldots, M_n, U_1, \ldots, U_n)
\end{aligned}$$

Note that in Hyb_j, only $M_1, \ldots, M_j$ are measured. $M_{j+1}, \ldots, M_n$ are untouched. We immediately have

$$\Pr[b = 1 : \mathsf{Game}_1] = \Pr[b = 1 : \mathsf{Hyb}_n], \qquad \Pr[b = 1 : \mathsf{Game}_2] = \Pr[b = 1 : \mathsf{Hyb}_0]. \quad (1)$$

We define a new adversary (A', B') for $(com, verify)$ as follows: $A'(1^\eta)$ picks $j \xleftarrow{\$} \{1, \ldots, n\}$. Then he executes $(S, M_1, \ldots, M_n, U_1, \ldots, U_n, c_1, \ldots, c_n) \leftarrow A(1^\eta)$. He measures $m_i \leftarrow M_{comp}(M_i)$ for $i = 1, \ldots, j-1$, and then sets

$$S' := (j, S, M_1, \ldots, M_{j-1}, M_{j+1}, \ldots, M_n, U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_n)$$

and $M := M_j$ and $U := U_j$ and $c := c_j$ and returns (S', M, U, c). And $B'(1^\eta, S', M, U)$ splits S' again into $(j, S, M_1, \ldots, M_{j-1}, M_{j+1}, \ldots, M_n, U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_n)$ and lets $M_j := M$ and $U_j := U$ and runs $B(1^\eta, S, M_1, \ldots, M_n, U_1, \ldots, U_n)$.

As mentioned above, since A is valid for each i, measuring M_i, U_i returns m_i, u_i with $verify(1^\eta, c_i, m_i, u_i) = 1$. Hence measuring M, U as output by A' returns m, u with $verify(1^\eta, c, m, u) = 1$. Thus A' is valid for $(com, verify)$.

Thus $\left|\Pr[b = 1 : \mathsf{Game}'_1] - \Pr[b = 1 : \mathsf{Game}'_2]\right|$ is negligible where $\mathsf{Game}'_1, \mathsf{Game}'_2$ are as follows:

$$\begin{aligned}
\mathsf{Game}'_1: &\quad (S', M, U, c) \leftarrow A(1^\eta),\ m \leftarrow M_{comp}(M),\ b \leftarrow B(1^\eta, S', M, U)\\
\mathsf{Game}'_2: &\quad (S', M, U, c) \leftarrow A(1^\eta), \qquad\qquad\qquad\quad b \leftarrow B(1^\eta, S', M, U)
\end{aligned}$$

For any fixed choice of j, Game'_1 is the same as Hyb_j, and Game'_2 is the same as Hyb_{j-1}. Thus

$$\begin{aligned}\Pr[b=1:\mathsf{Game}'_1] &= \sum_{j=1}^{n} \tfrac{1}{n} \Pr[b=1:\mathsf{Hyb}_j],\\ \Pr[b=1:\mathsf{Game}'_1] &= \sum_{j=1}^{n} \tfrac{1}{n} \Pr[b=1:\mathsf{Hyb}_{j-1}].\end{aligned} \tag{2}$$

Hence

$$\begin{aligned}&\big|\Pr[b=1:\mathsf{Game}'_1] - \Pr[b=1:\mathsf{Game}'_2]\big|\\ &\overset{(2)}{=} \tfrac{1}{n}\big|\Pr[b=1:\mathsf{Hyb}_n] - \Pr[b=1:\mathsf{Hyb}_0]\big|\\ &\overset{(1)}{=} \tfrac{1}{n}\big|\Pr[b=1:\mathsf{Game}_1] - \Pr[b=1:\mathsf{Game}_2]\big| \qquad (3)\end{aligned}$$

We showed above that the lhs of (3) is negligible. Thus the rhs of (3) is negligible, too. Since n is polynomially-bounded in η, this implies that $\big|\Pr[b=1:\mathsf{Game}_1] - \Pr[b=1:\mathsf{Game}_2]\big|$ is negligible as well. As stated in the beginning of this proof, this implies that $(com^n, verify^n)$ is collapse-binding. □

3 Commitments from collision-resistant hash functions

In the following, we will often refer to hash functions. We will always assume that a hash function depends implicitly on the security parameter (in particular, the size of the range can depend on the security parameter). We also assume that the hash function is quantum-polynomial-time computable (in η and the input length).[7] Besides that, we do not assume any further properties such as collision-resistance unless explicitly mentioned.

Definition 15 (Canonical commitment scheme) *Given a hash function H and a parameter $\ell_u = \ell_u(\eta)$, the* canonical commitment scheme *for H is:*

- *Message space* $\mathsf{MSP}_\eta := \{0,1\}^*$.
- $com_{can}(m)$*: Pick* $u \xleftarrow{\$} \{0,1\}^{\ell_u}$. *Compute* $c := H(m\|u)$. *Return* (c,u).
- $verify_{can}(c,m,u)$*: Return* 1 *iff* $H(m\|u) = c$.

It is immediate to see that this scheme is classical-style binding if H is collision-resistant. However, in general it will not be hiding; for example, $H(m\|u)$ could leak the first bit of m. However, it is hiding if H is a random oracle:

Lemma 16 *Fix $\ell_u \geq 0$ and assume that $|Y| \leq 2^{\ell_u/8}$. For a random oracle $H : X \to Y$, the canonical commitment is statistically hiding.*

[7] When working in the random oracle model: Quantum-polynomial-time computable given access to the random oracle.

Proof. This lemma was proven in [Pas04, Lemma 9]. The statement of the lemma there additionally assumes that the message space of the canonical commitment is also $\{0,1\}^{\ell_u}$ (i.e., equal to the space of the randomness u). However, this is never used in the proof. Furthermore, the lemma there assumes that $|Y| = 2^{\ell_u/8}$, but the adaption to the case $|Y| \leq 2^{\ell_u/8}$ is straightforward. □

When using a hash function in the standard model, we can use the following commitment scheme instead:

Definition 17 (Bounded-length Halevi-Micali commitment [HM96]) *Fix integers $\ell = \ell(\eta)$, $n = n(\eta)$. Let $L := 4\ell+2n+4$. Let $H : \{0,1\}^L \to \{0,1\}^\ell$ be a hash function. Let $F = F(\eta)$ be a family of universal hash functions $f : \{0,1\}^L \to \{0,1\}^n$. We define the* bounded-length Halevi-Micali commitment *$(com_{HMb}, verify_{HMb})$ with $\mathsf{MSP}_\eta = \{0,1\}^n$ as:*

- *$com_{HMb}(m)$: Pick $f \in F$ and $u \in \{0,1\}^L$ uniformly at random, conditioned on $f(u) = m$. Compute $h := H(u)$. Let $c := (h, f)$. Return (c, u).*
- *$verify_{HMb}(c, m, u)$ with $c = (h, f)$: Check whether $f(u) = m$ and $h = H(u)$. If so, return 1.*

Definition 18 (Unbounded Halevi-Micali commitment [HM96]) *Fix an integer $\ell = \ell(\eta)$. Let $H : \{0,1\}^* \to \{0,1\}^\ell$ be a hash function. Let $L := 6\ell + 4$. Let F be a family of universal hash functions $f : \{0,1\}^L \to \{0,1\}^\ell$. We define the* unbounded Halevi-Micali commitment *$(com_{HMu}, verify_{HMu})$ as:*

- *$com_{HMu}(m)$: Pick $f \in F$ and $u \in \{0,1\}^L$ uniformly at random, conditioned on $f(u) = H(m)$. Compute $h := H(u)$. Let $c := (h, f)$. Return (c, u).*
- *$verify_{HMu}(c, m, u)$ with $c = (h, f)$: Check whether $f(u) = H(m)$ and $h = H(u)$. If so, return 1.*

Theorem 19 (Security of Halevi-Micali [HM96]) *If ℓ is superlogarithmic, then the Halevi-Micali commitment and the bounded-length Halevi-Micali commitment are statistically hiding. If H is collision-resistant, then the Halevi-Micali commitment and the bounded-length Halevi-Micali commitment are* classical-style *binding.*

Note that [HM96] did not prove the classical-style binding property against *quantum* adversaries. But the (very simple) proof of binding carries over unchanged to the quantum setting (if H is collision-resistant against quantum adversaries). The statistical hiding property holds against unlimited adversaries anyway, thus also against quantum adversaries.

The following theorem shows that collision-resistance does not seem to be enough to make the above constructions secure in the quantum setting, i.e., classical-style binding is all we get.

Theorem 20 *There is an oracle $\mathcal{O}$ relative to which there exists a collision-resistant*[8] *hash function H such that the canonical commitment scheme and both Halevi-Micali commitment schemes using H admit the following attack:*

There is a quantum-polynomial-time adversary $A^{\mathcal{O}}$ that outputs a commitment c, then expects a bit b, and then outputs with overwhelming probability a pair (m, u) such that verify$(c, m, u) = 1$ and the first bit of m is b.

Clearly, a commitment with that property should not be considered secure. This shows that collision-resistance is too weak a property for constructing commitments in the quantum setting, at least when using standard constructions.

Proof. [ARU14, Definition 6] defines a specific oracle $\mathcal{O}_{all}$ (more precisely, a probability distribution on oracles). We repeat only the parts of the construction that are relevant for our proof: Let $X := \{0,1\}^{\ell_1}$ and $Y := \{0,1\}^{\ell_2}$ for some arbitrary polynomially-bounded superlogarithmic ℓ_1, ℓ_2. For each $y \in Y$, let $S_y \subseteq X$ be a uniformly random subset of a certain size k. Let $\mathcal{O}_V$ be an oracle that tests membership in S_y, more precisely $\mathcal{O}_V(y, x) = 1$ iff $x \in S_y$. ($\mathcal{O}_V$ may be queried in superposition.) Finally, $\mathcal{O}_{all}$ is defined to be an oracle consisting of $\mathcal{O}_V$ and several other oracles (some of them implementing unitary transformations).

We use the following important facts about $\mathcal{O}_{all}$:

Fact 1 (Hardness of two values) *Let A be an algorithm making a polynomial number of oracle queries. Then $\Pr[x \neq x' \ \wedge \ x, x' \in S_y : (y, x) \leftarrow A^{\mathcal{O}_{all}}(1^\eta)]$ is negligible.*

This fact is a reformulation of [ARU14, Corollary 7 (i)].

Fact 2 (Searching one value) *There is a pair (E_1, E_2) of quantum-polynomial-time oracle algorithms such that:*

- *$E_1^{\mathcal{O}_{all}}(1^\eta)$ outputs $y \in Y$ and a quantum state $|\Psi(y)\rangle$.*
- *Given a Boolean circuit P with $|\{x \in S_y : P(x) = 1\}| \geq |S_y|/3$, $E_2^{\mathcal{O}_{all}}(1^\eta, y, |\Psi(y)\rangle, P)$ outputs $x \in S_y$ with $P(x) = 1$ with overwhelming probability.*

This is a special case of [ARU14, Theorem 5].[9]

Informally, Fact 2 tells us that if we choose $y \in Y$ ourselves, we get a quantum trapdoor $|\Psi(y)\rangle$ that allows us to search *one* value $x \in S_y$ satisfying a predicate of our choice, as long as this predicate is satisfied $\frac{1}{3}$ of the time. (But note: we cannot get two such x in the same S_y, as this would violate Fact 1.)

Let $h_2 : \{0,1\}^* \to \{0,1\}^\ell$ (for some arbitrary polynomially-bounded superlogarithmic ℓ) be uniformly random. We can then define the oracle $\mathcal{O}$ to be the oracle containing

[8] H is collision-resistant iff for any quantum-polynomial-time A, $\Pr[x \neq x' \wedge H(x) = H(x') : (x, x') \leftarrow A(1^\eta)]$ is negligible.

[9] We have fixed $\delta_{\min} := 1/3$ and n to be the security parameter, and we have removed the argument $|\Sigma\Psi\rangle$ from E_1 because $|\Sigma\Psi\rangle$ can be produced by E_1 using the oracle $\mathcal{O}_\Psi$ contained in $\mathcal{O}_{all}$.

$\mathcal{O}_{all}$ and h_2. (I.e., $\mathcal{O}$ gives access to $\mathcal{O}_{all}$ and an additional random oracle.) Note that since h_2 and $\mathcal{O}_{all}$ are independent, Fact 1 still applies when A is given access to $\mathcal{O}$.

We now construct a hash function $H : \{0,1\}^* \to \{0,1\}^\ell$. For $x \in X$, $y \in Y$ with $\mathcal{O}_V(y,x) = 1$, let $h_1(x\|y) := 0\|y$ and let $h_1(z) := 1\|z$ everywhere else. Let $H := h_2 \circ h_1$.

Claim 1 (Collision-resistance of H) *H is collision-resistant (relative to $\mathcal{O}$).*

To show this, we show that h_1 and h_2 are collision-resistant relative to $\mathcal{O}$. This then shows that $H = h_2 \circ h_1$ is collision-resistant relative to $\mathcal{O}$. Any collision of h_1 must be of the form $h_1(x\|y) = h_1(x'\|y')$ with $x\|y \neq x'\|y'$ and $\mathcal{O}_V(y',x') = \mathcal{O}_V(y,x) = 1$ since h_1 is injective everywhere else. By definition of h_1, this implies that $0\|y = 0\|y'$, thus $y = y'$ and $x \neq x'$. And then $\mathcal{O}_V(y',x') = \mathcal{O}_V(y,x) = 1$ implies by definition of $\mathcal{O}_V$ that $x, x' \in S_y$. By Fact 1, a polynomial-time adversary with oracle access to $\mathcal{O}$ finds such x, x', y only with negligible probability. This shows that h_1 is collision-resistant relative to $\mathcal{O}$.

By [Zha13, Theorem 3.1], h_2 is collision-resistant (given oracle access to h_2).[10] Since $\mathcal{O}_{all}$ is chosen independently of h_2, it can be simulated with no extra queries to h_2. I.e., an adversary breaking collision resistance of h_2 using $\mathcal{O} = (\mathcal{O}_{all}, h_2)$ can be transformed into one breaking collision resistance of h_2 using h_2. Hence h_2 is also collision-resistant given oracle access to $\mathcal{O}$.

Thus h_1, h_2 are collision-resistant relative to $\mathcal{O}$, and thus $H = h_2 \circ h_1$ is collision-resistant relative to $\mathcal{O}$.

Attack on the canonical commitment scheme. Let ℓ_m be some arbitrary message length, and ℓ_u the length of the opening information (see Definition 15). For this attack, we assume that the length parameters ℓ_1, ℓ_2 in the construction of $\mathcal{O}_{all}$ have been chosen such that $\ell_m + \ell_u = \ell_1 + \ell_2$. (This is always possible, since ℓ_1, ℓ_2 are only required to be superlogarithmic.) The adversary A does the following:

- Let E_1, E_2 be the algorithms from Fact 2.
- $(y, |\Psi(y)\rangle) \leftarrow E_1^{\mathcal{O}_{all}}(1^\eta)$. Let $c := h_2(0\|y)$ and send c as the commitment.
- Upon input b, choose P such that $P(x) := 1$ iff the first bit of x is b. Run $x \leftarrow E_2^{\mathcal{O}_{all}}(1^\eta, y, |\Psi(y)\rangle, P)$. Split $x\|y$ as $m\|u := x\|y$ with $|m| = \ell_m, |u| = \ell_u$ and send (m,u). (Note: the lengths of m, u do not necessarily match the lengths of x, y, but their combined length does since $\ell_1 + \ell_2 = \ell_m + \ell_u$.)

Since $S_y \subseteq Y$ is a random set of (superpolynomial) cardinality k, we have that the fraction of S_y having leading bit b (i.e., satisfying P) is at least $\frac{1}{3}$ with overwhelming probability. Thus x as returned by E_2 satisfies, by Fact 2, with overwhelming probability $x \in S_y$ and $P(x) = 1$. From $P(x) = 1$ it follows that the first bit of m is b as

[10] Strictly speaking, [Zha13, Theorem 3.1] only applies to random oracles with finite but arbitrary large domain, not to h_2 which has domain $\{0,1\}^*$. However, if an adversary finds a collision in h_2 with non-negligible probability μ, then there must be a length ℓ^* such that the adversary finds a collision of length at most ℓ^* with probability at least $\mu/2$. Thus an adversary breaking collision-resistance of h_2 can be transformed into an adversary breaking collision-resistance of a random oracle with finite domain. [Zha13, Theorem 3.1] then applies.

required. And $x \in S_y$ implies $\mathcal{O}_V(y,x) = 1$ which implies $h_1(x\|y) = 0\|y$. Hence $H(m\|u) = h_2(h_1(x\|y)) = h_2(0\|y) = c$. Thus $\mathit{verify}_{can}(c,m,u) = 1$. This shows that the attack on the canonical commitment is successful with overwhelming probability.

Attack on the bounded-length Halevi-Micali commitment. Let n be the message length, and ℓ, L as in Definition 17. For this attack, we assume that the length parameters ℓ_1, ℓ_2 have been chosen such that $\ell_1 + \ell_2 = L$. (This is always possible, since ℓ_1, ℓ_2 are only required to be superlogarithmic.) The adversary A does the following:

- Let E_1, E_2 be the algorithms from Fact 2.
- $(y, |\Psi(y)\rangle) \leftarrow E_1^{\mathcal{O}_{all}}(1^\eta)$. Pick $f \in F$ (the family of universal hash functions). Let $h := h_2(0\|y)$ and let $c := (h, f)$ and send c as the commitment.
- Upon input b, choose P such that $P(x) := 1$ iff the first bit of $f(x)$ is b. Run $x \leftarrow E_2^{\mathcal{O}_{all}}(1^\eta, y, |\Psi(y)\rangle, P)$. Let $u := x$ and $m := f(u)$ and send (m, u).

Similarly as for the attack on the canonical commitment, we get that A gets with overwhelming probability an $x \in S_y$ with $P(x) = 1$ which then implies $\mathit{verify}_{HMb}(c,m,u) = 1$.

Attack on the unbounded Halevi-Micali commitment. We describe the attack on the unbounded Halevi-Micali commitment. Let ℓ_m be a superpolynomial message length, and L the length of the opening information (see Definition 18). For this attack, we assume that the length parameters ℓ_1, ℓ_2 have been chosen such that $\ell_1 + \ell_2 = \ell_m$. (This is always possible, since ℓ_1, ℓ_2 are only required to be superlogarithmic.) The adversary A does the following:

- Let E_1, E_2 be the algorithms from Fact 2.
- $(y, |\Psi(y)\rangle) \leftarrow E_1^{\mathcal{O}_{all}}(1^\eta)$. Pick $f \in F$ and $u \in \{0,1\}^L$ such that $f(u) = h_2(0\|y)$. Compute $h := H(u)$. Let $c := (h, f)$ and send c as the commitment.
- Upon input b, let $P(x) := 1$ iff the first bit of x is b. Run $x \leftarrow E_2^{\mathcal{O}_{all}}(1^\eta, y, |\Psi(y)\rangle, P)$. Let $m := x\|y$. Send (m, u).

Similarly as for the attack on the canonical commitment, we get that A gets with overwhelming probability an $x \in S_y$ with $P(x) = 1$ which then implies $\mathit{verify}_{HMu}(c,m,u) = 1$.
□

4 Collapsing hash functions

As seen in the previous section, for many protocols collision-resistance is not a sufficiently strong property in the quantum setting. In the following, we propose a strengthening of the collision-resistance property that seems more useful in the quantum setting, namely "collapsing" hash functions. We believe that collapsing hash functions are a natural assumption for real-life hash functions such as SHA-3 etc. This belief is supported by the fact that the random oracle is collapsing (see Section 6).

The definition of collapsing hash functions is similar to that of collapsing commitments (Definition 11).

Definition 21 (Collapsing) *For a function H and algorithms A, B, consider the following games:*

$$\begin{aligned}
\mathsf{Game}_1: &\quad (S,M,c) \leftarrow A(1^\eta),\ m \leftarrow M_{comp}(M),\ b \leftarrow B(1^\eta,S,M)\\
\mathsf{Game}_2: &\quad (S,M,c) \leftarrow A(1^\eta),\ b \leftarrow B(1^\eta,S,M)
\end{aligned}$$

Here S, M are quantum registers. $M_{comp}(M)$ is a measurement of M in the computational basis.

We call an adversary (A,B) valid *if $\Pr[H(m)=c]=1$ when we run $(S,M,c)\leftarrow A(1^\eta)$ and measure M in the computational basis as m.*

A hash function H is collapsing *iff for any quantum-polynomial-time valid adversary (A,B), the difference $adv := \left|\Pr[b=1:\mathsf{Game}_1]-\Pr[b=1:\mathsf{Game}_2]\right|$ is negligible. (We call adv the* advantage*.)*

Notice that the definition of collapsing hash functions is inherently quantum, even though the object we consider (the hash function H) is classical. We know of no classical analogue to collapsing hash functions. However, a collapsing hash function will necessarily be collision-resistant, see Lemma 23 below.

We proceed to give a number of useful properties of collapsing hash functions.

Lemma 22 *An injective function H is collapsing with advantage 0.*

Proof. Consider an adversary (A,B) against Definition 21. Since (A,B) is valid, by definition we have that $m \leftarrow M_{comp}(M)$ in Game_1 returns m with $H(m)=c$. Since H is injective, this means there is only one such m. Thus M is in state $|m\rangle$ before applying $m \leftarrow M_{comp}(M)$, and the measurement $M_{comp}(M)$ does not change the state of M. Thus $\Pr[b=1:\mathsf{Game}_1]=\Pr[b=1:\mathsf{Game}_2]$. □

Lemma 23 *A collapsing hash function is collision resistant.*

Proof. Assume the hash function H is not collision resistant. Then there is a quantum adversary C that outputs a collision (m,m') with $H(x)=H(x')$ with non-negligible probability μ.

We construct a quantum-polynomial-time adversary (A,B) for Definition 21.

Let A be the following quantum algorithm: It runs C to get a collision (m,m'). If (m,m') is a collision, it stores m,m' in the register S, and initializes M with $|\Psi_{m,m'}\rangle := \frac{1}{\sqrt{2}}|m\rangle+\frac{1}{\sqrt{2}}|m'\rangle$. It sets $c:=H(m)=H(m')$ and returns (S,M,c). If (m,m') is not a collision, A stores $\perp$ in the register S, initializes M with $|0\rangle$, sets $c:=H(0)$, and returns (S,M,c).

The algorithm B retrieves m,m' from S. If S contains $\perp$ instead, B returns $b:=0$. Otherwise B measures whether M contains $|\Psi_{m,m'}\rangle$, i.e., B measures M with the projector $|\Psi_{m,m'}\rangle\langle\Psi_{m,m'}|$. If this measurement succeeds, B returns $b:=1$, else B returns $b:=0$.

By construction, (A,B) is valid.

In Game_2, with probability $1-\mu$, B finds S to contain $\perp$ and returns $b=0$. If S contains a collision m, m', then by construction of A, M contains $|\Psi_{m,m'}\rangle$, so B outputs $b=1$ with probability 1 in this case. Hence $\Pr[b=1:\mathsf{Game}_2]=\mu$.

In Game_1, with probability $1-\mu$, B finds S to contain $\perp$ and returns $b=0$. If S contains a collision m, m', then by construction the state of M before $M_{comp}(M)$ is $|\Psi_{m,m'}\rangle$, hence after that measurement it is $|m\rangle$ or $|m'\rangle$ (each with probability $\frac{1}{2}$). In each case, the measurement performed by B (projector $|\Psi_{m,m'}\rangle\langle\Psi_{m,m'}|$) succeeds with probability $\frac{1}{2}$. Thus, if S contains a collision, B returns $b=1$ with probability $\frac{1}{2}$. Hence $\Pr[b=1:\mathsf{Game}_1]=\mu/2$.

Hence $\left|\Pr[b=1:\mathsf{Game}_1]-\Pr[b=1:\mathsf{Game}_2]\right|=\frac{\mu}{2}$ is non-negligible, in contradiction to the assumption that H is collapsing. □

Definition 21 guarantees that the adversary cannot distinguish whether the register M is measured or not. Like in the case of commitments (cf. the discussion before Lemma 13) we can ask what happens when a partial measurement is performed. Analogous to Lemma 13 we get that a partial measurement cannot be noticed, either:

Lemma 24 (Collapsing w.r.t. partial measurements) *For a function H and algorithms A, B, consider the following games:*

$$\begin{aligned}
&\mathsf{Game}_1': \quad (S,M,c,f)\leftarrow A(1^\eta),\ x\leftarrow M^f(M),\ b\leftarrow B(1^\eta,S,M)\\
&\mathsf{Game}_2: \quad (S,M,c,f)\leftarrow A(1^\eta),\qquad\qquad\quad b\leftarrow B(1^\eta,S,M)
\end{aligned}$$

Here f is a Boolean circuit (with multiple-bit output).[11] *And S, M are quantum registers. $M^f(M)$ measures $f(m)$ where m is the content of M in the computational basis. Formally, $M^f(M)$ is the measurement defined by the projectors $P_x := \sum_{m:f(m)=x}|m\rangle\langle m|$ for all x in the range of f.*

If H is collapsing, then for any quantum-polynomial-time valid adversary (A,B), the difference $\left|\Pr[b=1:\mathsf{Game}_1']-\Pr[b=1:\mathsf{Game}_2]\right|$ is negligible.

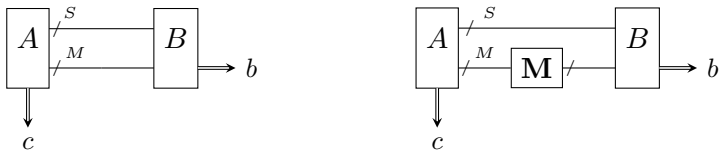
Proof. The proof is analogous to that of Lemma 13. (The proof actually becomes a bit simpler, because all occurrences of and arguments relating to V_c may be omitted.) □

Lemma 25 *If a valid adversary (A,B) breaks the collapsing property of $g\circ f$ with advantage ε, then there are valid adversaries (A',B') and (A'',B'') with advantages $\varepsilon',\varepsilon''$ against g, f, respectively, such that $\varepsilon\le\varepsilon'+\varepsilon''$.*

(A',B') and (A'',B'') each perform only two additional evaluations of f in comparison to (A,B). (And one additional measurement in the computational basis. But no additional evaluations of g.)

[11] In the random oracle model, we also allow f to contain gates for evaluating the random oracle.

Proof. Consider the following circuits:

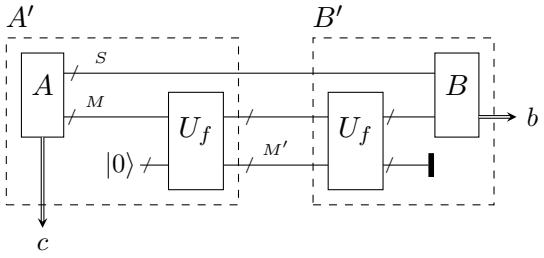


(4)

Here **M** represents a measurement in the computational basis. (Discarding the outcome.) By definition of ε, we have

$$\varepsilon = \Big|\Pr[b = 1 : \text{lhs of (4)}] - \Pr[b = 1 : \text{rhs of (4)}]\Big|.$$

Let $U_f : |x\rangle|y\rangle \mapsto |x\rangle|y \oplus f(x)\rangle$. Since U_f is self-inverse, introducing two consecutive applications of U_f into the lhs of (4) does not change the outcome probability. That is, with



(5)

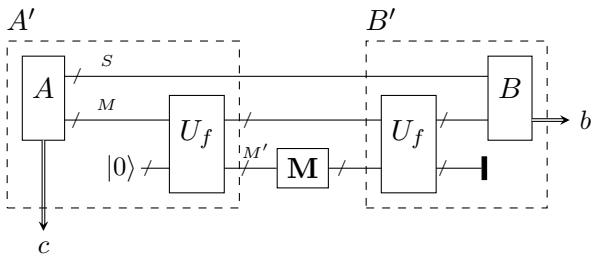
we have

$$\Pr[b = 1 : \text{lhs of (4)}] = \Pr[b = 1 : (5)].$$

The dashed boxes in (5) define a new adversary (A', B') against g. The top two wires leaving A' contain the state of A', while the bottom wire M' contains the superposition of hashed values. Since A is valid for $g \circ f$, M contains a superposition of values m with $g \circ f(m) = c$. (By this we mean formally that the projector $\sum_{m,u:g\circ f(m)=c} |m\rangle\langle m|$ applied to M passes with probability 1.) Hence M' contains a superposition of values $m' = f(m)$ with $g(m') = c$. Thus A' is valid for g. Let ε' be the advantage of A' against g. That is, we have

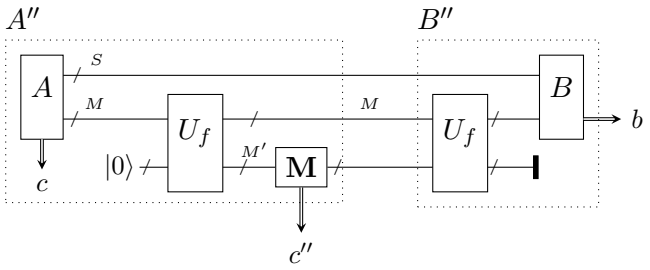
$$\varepsilon' = \Big|\Pr[b = 1 : (5)] - \Pr[b = 1 : (6)]\Big|$$

with the following circuit (6):



(6)

We now change the circuit slightly: Instead of discarding the outcome of the measurement $\mathbf{M}$, we assign it to the classical variable c''.



(7)

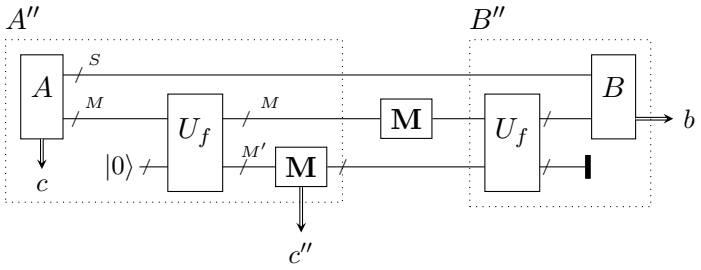
Obviously, not discarding c'' does not change the distribution of b, hence

$$\Pr[b = 1 : (6)] = \Pr[b = 1 : (7)].$$

The dotted lines in (7) define an adversary (A'', B'') against f. The wires S and M' together form the state of (A'', B''), and the middle wire M is supposed to contain the hashed values. If M contains the value m, then M' contains $f(m)$ and c'' will be $f(m)$. Thus, if we measure a particular value c'', then M contains a superposition of values m with $f(m) = c''$. Thus, (A'', B'') is valid for f. Let ε'' be the advantage of (A'', B'') against f. Then we have

$$\varepsilon'' = \big|\Pr[b = 1 : (7)] - \Pr[b = 1 : (8)]\big|$$

with the following circuit (8):



(8)

The subcircuit consisting of the two U_f and the two measurements $\mathbf{M}$ is easily seen to be equivalent to a measurement $\mathbf{M}$ on the M wire (since we do not use the outcome c''). Thus

$$\Pr[b = 1 : (8)] = \Pr[b = 1 : \text{rhs of } (4)].$$

Collecting all inequalities, we get:

$$\varepsilon = \big|\Pr[b = 1 : \text{lhs of } (4)] - \Pr[b = 1 : \text{rhs of } (4)]\big| \leq \varepsilon' + \varepsilon''. \qquad \square$$

Corollary 26 *If f and g are collapsing, so is $g \circ f$.*

Proof. Immediate from Lemma 25. $\square$

5 Commitments from collapsing hash functions

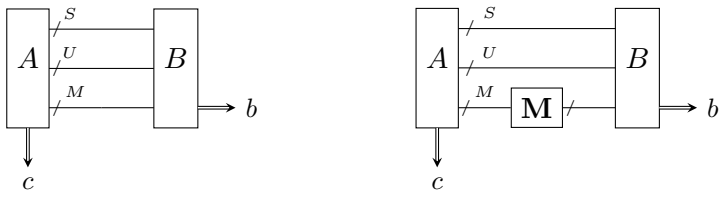
In Section 3 we saw that collision-resistant hash functions are not sufficient for several standard constructions of commitment schemes. We will now show that those same constructions are secure in the quantum setting when using collapsing hash functions instead.

The following lemma (and its Corollary 28 below) allow us to extend the message space of a collapsing commitment by hashing the message with a collapsing hash function. Besides being useful in its own right, we need it in the analysis of the unbounded Halevi-Micali commitment. The proof of the lemma is similar to that of Corollary 26.

Lemma 27 *Let f be a hash function. Let $(com, verify)$ be a commitment scheme. Let $com_f(1^\eta, m) := com(f(m))$ and $verify_f(1^\eta, c, m, u) = verify(1^\eta, c, f(m), u)$. If a valid adversary (A, B) breaks the collapse binding property of $(com_f, verify_f)$ with advantage ε (with respect to Definition 11), then there are valid adversaries (A', B') and (A'', B'') with advantages $\varepsilon', \varepsilon''$ against the collapse binding property of $(com, verify)$ and the collapsing property of f, respectively, such that $\varepsilon \leq \varepsilon' + \varepsilon''$.*

(A', B') and (A'', B'') each perform only two additional evaluations of f in comparison to (A, B). (And one additional measurement in the computational basis. But no additional evaluations of com or verify.)

Proof. Consider the following circuits:

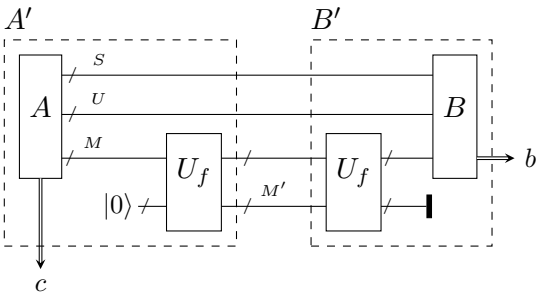


(9)

Here **M** represents a measurement in the computational basis. (Discarding the outcome.) By definition of ε, we have

$$\varepsilon = \Big|\Pr[b = 1 : \text{lhs of (9)}] - \Pr[b = 1 : \text{rhs of (9)}]\Big|.$$

Let $U_f : |x\rangle|y\rangle \mapsto |x\rangle|y \oplus f(x)\rangle$. Since U_f is self-inverse, introducing two consecutive applications of U_f into the lhs of (9) does not change the outcome probability. That is, with



(10)

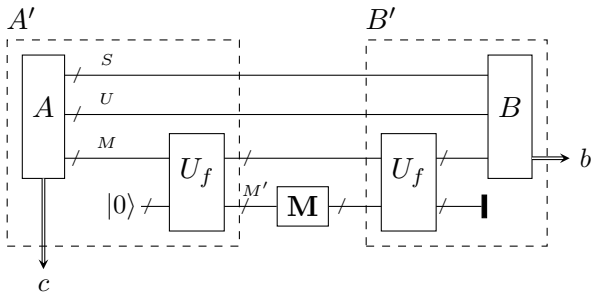
we have

$$\Pr[b = 1 : \text{lhs of (9)}] = \Pr[b = 1 : (10)].$$

The dashed boxes in (10) define a new adversary (A', B') against $(com_f, verify_f)$. The wires S, M leaving A' contain the state of A', the wire M' contains the committed message, and the wire U contains the opening information. Since A is valid for $(com_f, verify_f)$, M, U contains a superposition of values m, u with $verify(1^\eta, c, f(m), u) = 1$. (By this we mean formally that the projector $\sum_{m,u:verify(1^\eta,c,f(m),u)=1} |m\rangle\langle m| \otimes |u\rangle\langle u|$ applied to M, U passes with probability 1.) Hence M', U contains a superposition of values $m' = f(m)$ with $verify(1^\eta, c, m', u) = 1$. Thus A' is valid for $(com, verify)$. Let ε' be the advantage of A' against g. That is, we have

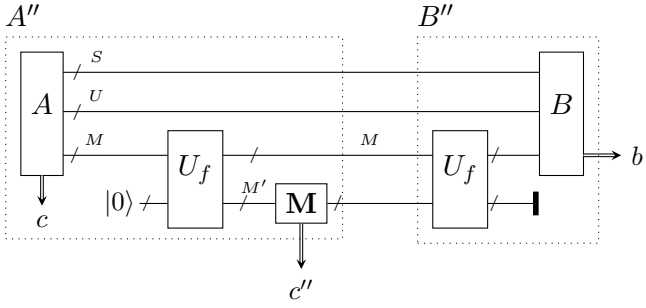
$$\varepsilon' = \big|\Pr[b = 1 : (10)] - \Pr[b = 1 : (11)]\big|$$

with the following circuit (11):



(11)

We now change the circuit slightly: Instead of discarding the outcome of the measurement $\mathbf{M}$, we assign it to the classical variable c''.



(12)

Obviously, not discarding c'' does not change the distribution of b, hence

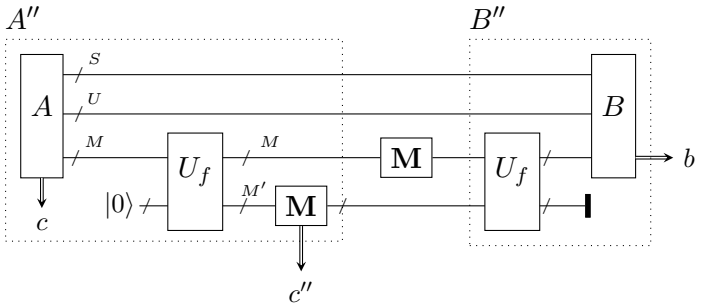
$$\Pr[b = 1 : (11)] = \Pr[b = 1 : (12)].$$

The dotted lines in (12) define an adversary (A'', B'') against f. The wires S, U, and M' together form the state of (A'', B''), and the middle wire M is supposed to contain the hashed values. If M contains the value m, then M' contains $f(m)$ and c'' will be $f(m)$.

Thus, if we measure a particular value c'', then M contains a superposition of values m with $f(m) = c''$. Thus, (A'', B'') is valid for f. Let ε'' be the advantage of (A'', B'') against f. Then we have

$$\varepsilon'' = \big|\Pr[b = 1 : (12)] - \Pr[b = 1 : (13)]\big|$$

with the following circuit (13):



(13)

The subcircuit consisting of the two U_f and the two measurements $\mathbf{M}$ is easily seen to be equivalent to a measurement $\mathbf{M}$ on the M wire (since we do not use the outcome c''). Thus

$$\Pr[b = 1 : (13)] = \Pr[b = 1 : \text{rhs of } (9)].$$

Collecting all inequalities, we get:

$$\varepsilon = \big|\Pr[b = 1 : \text{lhs of } (9)] - \Pr[b = 1 : \text{rhs of } (9)]\big| \leq \varepsilon' + \varepsilon''. \qquad \square$$

Corollary 28 *Let f be a collapsing function. Let $(com, verify)$ be a collapse binding commitment scheme. Let $com_f(1^\eta, m) := com(1^\eta, f(m))$ and $verify_f(1^\eta, c, m, u) = verify(1^\eta, c, f(m), u)$. Then $(com_f, verify_f)$ is a collapse-binding commitment scheme.*

Proof. Immediate from Lemma 27. $\square$

Lemma 29 *If H is collapsing, then the canonical commitment scheme $(com_{can}, verify_{can})$, and the bounded-length Halevi-Micali commitment $(com_{HMb}, verify_{HMb})$, and the unbounded Halevi-Micali commitment $(com_{HMu}, verify_{HMu})$ are collapse-binding. (For any choice of the parameters ℓ_u, ℓ, n.)*

We give the proof ideafirst. To show that the canonical commitment com_{can} is collapse-binding, we use the characterization of collapse-binding from Definition 11. We need to show that the adversary cannot distinguish between a measurement on register M and no measurement on register M, assuming the adversary outputs M, U containing a superposition of m, u with $verify_{can}(c, m, u) = 1$. The condition $verify_{can}(c, m, u) = 1$ is equivalent to $H(m\|u) = c$. Hence the adversary outputs in M, U a superposition of preimages of c under H. Since H is collapsing, this implies that the adversary cannot distinguish between a measurement on M, U and no measurement on M, U. This also implies (using some additional work) that the adversary cannot distinguish between a measurement on M and no measurement on M. Hence com_{can} is collapse-binding. The Halevi-Micali commitments are handled similarly.

Proof. We investigate the three commitment schemes one by one:

Canonical commitment. We show that $(com_{can}, verify_{can})$ is collapse-binding. Fix a valid quantum-polynomial-time adversary (A, B) against the canonical commitment scheme (with respect to Definition 11). By concatenating the registers M, U into a single register M', we get an adversary (A', B') against the hash function H. Written in terms of A', B' (and unfolding the definition of $verify_{can}$), the games from Definition 11 become:

$$\begin{aligned}\mathsf{Game}_1: &\quad (S, M', c) \leftarrow A'(1^\eta),\ m \leftarrow M^f(M'),\ b \leftarrow B'(1^\eta, S, M')\\ \mathsf{Game}_2: &\quad (S, M', c) \leftarrow A'(1^\eta), \qquad\qquad\qquad\quad b \leftarrow B'(1^\eta, S, M')\end{aligned}$$

where $f(m\|u) := m$ and M^f is defined as in Lemma 24.

Since (A, B) is valid, we have $verify_{can}(1^\eta, c, m, u) = 1$ when m, u is the result of measuring M, U. When we define $m' := m\|u$, we have $verify(1^\eta, c, m, u) = 1$ iff $H(m') = c$ by definition of *verify*. Thus in an execution with (A', B'), we have $H(m') = c$. Thus (A', B') is valid.

Since (A', B') is valid and quantum-polynomial-time, by Lemma 24, $|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]|$ is negligible. (In Lemma 24, f is chosen by the adversary, but we can transform A' to output f himself.) Since Game_1 and Game_2 are equivalent to the games from Definition 11, it follows that $(com_{can}, verify_{can})$ is collapsing.

Bounded-length Halevi-Micali commitment. We show that $(com_{HMb}, verify_{HMb})$ is collapse-binding. Fix a valid quantum-polynomial-time adversary (A, B) against the commitment scheme (with respect to Definition 11). By unfolding the definition of $(com_{HMb}, verify_{HMb})$, see Definition 17, the games from Definition 11 become:

$$\begin{aligned}\mathsf{Game}_1: &\quad (S, M, U, h, f) \leftarrow A(1^\eta),\ m \leftarrow M_{comp}(M),\ b \leftarrow B(1^\eta, S, M, U)\\ \mathsf{Game}_2: &\quad (S, M, U, h, f) \leftarrow A(1^\eta), \qquad\qquad\qquad\quad b \leftarrow B(1^\eta, S, M, U)\end{aligned}$$

and validity of A implies that M, U are such that when measuring them, we get m, u with $f(u) = m$ and $h = H(u)$. We need to show that $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ is negligible.

Since $f(u) = m$, measuring M in the computational basis is equivalent to applying the measurement M^f on U. Here M^f is as in Lemma 24. Thus we have $\Pr[b = 1 : \mathsf{Game}_1] = \Pr[b = 1 : \mathsf{Game}'_1]$ with

$$\mathsf{Game}'_1: \quad (S, M, U, h, f) \leftarrow A(1^\eta),\ m \leftarrow M^f(U),\ b \leftarrow B(1^\eta, S, M, U).$$

By Lemma 24, we get that $\left|\Pr[b = 1 : \mathsf{Game}'_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ is negligible. (Here we instantiate S, M, c, f in Lemma 24 with $S := (S, M)$, $M := U$, $c := h$.)

Thus $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right|$ is also negligible, hence $(com_{HMb}, verify_{HMb})$ is collapse-binding.

Unbounded Halevi-Micali commitment. We show that $(com_{HMu}, verify_{HMu})$ is collapse-binding. Let $(com, verify) := (com_{HMb}, verify_{HMb})$ and $f := H$. Then $(com_f, verify_f)$ as defined in Corollary 28 is the same as $(com_{HMu}, verify_{HMu})$. We showed above that $(com_{HMb}, verify_{HMb})$ is collapse-binding. And $f = H$ is collapsing by assumption. Thus by Corollary 28, $(com_f, verify_f)$ is collapse binding. Hence $(com_{HMu}, verify_{HMu})$ is collapse-binding. □

6 Random oracles are collapsing

In Section 5 we saw that collapsing hash functions imply collapse-binding commitments. In this section, we explore the existence of collapsing hash functions. Specifically, we show that the random oracle is collapsing. This implies that there are simple collapse-binding commitments in the random oracle model. Furthermore, it supports the assumption that real-life hash functions such as SHA-3 etc. are collapse-binding. Under this assumption, the constructions from Section 5 are collapse-binding using those hash functions (that is, under this assumption, we do not need the random oracle).

For the remainder of this section, X and Y are sets, and $H : X \to Y$ is a random oracle. And Y is finite. And $X \subseteq \{0,1\}^*$ (finite or infinite). And $q \geq 1$ always refers to an upper bound on the number of oracle queries performed by the adversary.

We start by defining a seemingly unrelated property (half-collision resistance) that will turn out to imply the collapsing property. We will need half-collision resistance in our proof that the random oracle is collapsing. However, the concept of half-collision resistance might be of use for constructions in the standard model, too: since half-collision resistance is defined by a classical game, it might be easier to construct hash functions that are half-collision resistant.

Definition 30 *A* half-collision *of a hash function $f : X \to Y$ is a value x such that $\exists x' \neq x. f(x) = f(x')$.*

An adversary A has advantage ε against half-collision resistance *iff*

- *with probability* 1*, the output of A is a half-collision or $\perp$, and*
- *with probability at ε, A outputs a half-collision.*

Lemma 31 *If (A, B) is valid and has advantage μ against the collapsing property of a hash function f, then there is an adversary D with advantage $\geq \mu^2/4$ against the half-collision resistance of f. The time-complexity of D is linear in that of (A, B). (If f is given as an oracle, D makes $4q + 4$ queries to f when (A, B) makes q queries.)*

Proof sketch: By definition, a valid adversary A will always output in register M a superposition of messages m with $H(m) = c$ (all with the same c). So we have two cases: M contains a superposition of a single message m, or M contains a superposition of several messages that have the same image c, i.e., a superposition of half-collisions. Thus, in the second case, we can find a half-collisions by measuring M. But, an adversary against half-collision resistance must never output a non-half-collision (no false positives).

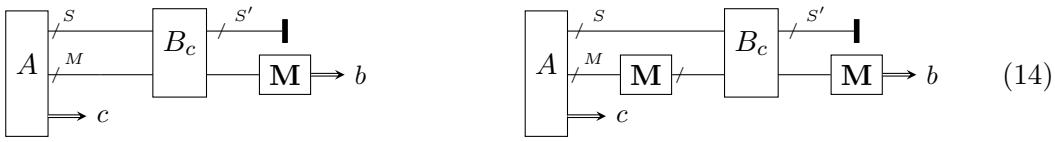
Thus, we need a possibility to test whether M contains only a single message. (In this case, we abort.)

Note that when M contains only a single message, then the adversary B cannot distinguish between a measurement on M and no measurement on M. To exploit this, we run an execution where M is measured and an execution where M is not measured in superposition (roughly speaking), and we make it depend on a control qubit in state $|+\rangle$ which execution is used. Then, in the case where M contains only a single message, the control qubit stays unentangled with the rest of the circuit. By measuring whether the qubit is still in state $|+\rangle$, the half-collision resistance adversary can detect whether M contains one or several messages. (It may err and incorrectly assume that M contains only one message, but an error in that direction is permitted.) Thus we have constructed an adversary against half-collision resistance.

Proof. We first construct a slight modification B_c of the adversary algorithm B. B_c is parametrized over an image c of f, and $B_c(S, M)$ first measures its input register M with the projector $P_c := \sum_{m:f(m)=c} |m\rangle\langle m|$. That is, B_c measures whether f applied to the content of M would return c. If so, B_c executes B and returns the output b of B. If not, B_c returns $b = 0$. B_c needs one more query to f than B.

Furthermore, we assume that B_c is implemented as a unitary circuit. That is, we assume that the input register S (which contains the state of the adversary in the games from Definition 21), contains sufficiently many ancillae for implementing B_c unitarily, and that B_c has an output register S' in addition to the register containing b.

In the games from Definition 21, the register M is in the image of P_c anyway, so in those games, B_c behaves the same as B. Thus, $\Pr[b = 1]$ differs by μ in the following circuits:

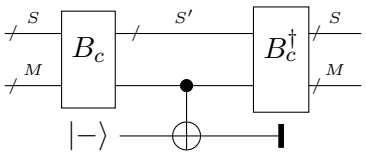


(14)

M denotes a (single- or multi-bit) measurement in the computational basis. (The first occurrence of **M** has only an outgoing quantum wire, so it discards the outcome. The second occurrence has only an outgoing classical wire b so it discards the post-measurement state.)

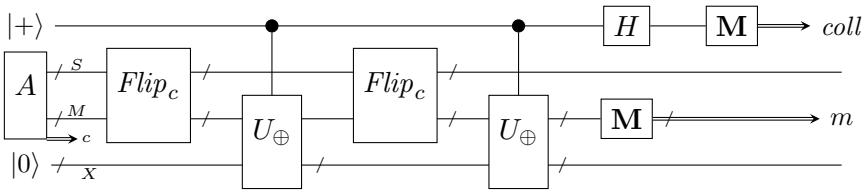
Let P_c^B be the projector onto the space of all input states where B_c will output $b = 1$ with probability 1. In particular, since B_c returns $b = 0$ for $|s\rangle|m\rangle$ with $f(m) \neq c$, we have that $\operatorname{im} P_c \subseteq \operatorname{im}(1 - P_c^B)$, so P_c and P_c^B commute. (Strictly speaking, we should refer to $I_S \otimes P_c$ here, since P_C operates on M only. But here and in the remainder of the proof, we implicitly omit tensor products with the identity since for every operator, it is clear on which registers it operates.) Let $\mathit{Flip}_c := (1 - 2P_c^B)$. It is easy to verify that

$Flip_c$ is implemented by the following quantum circuit:



We are now ready to define the adversary D against half-collision resistance:

- Execute the following quantum circuit:

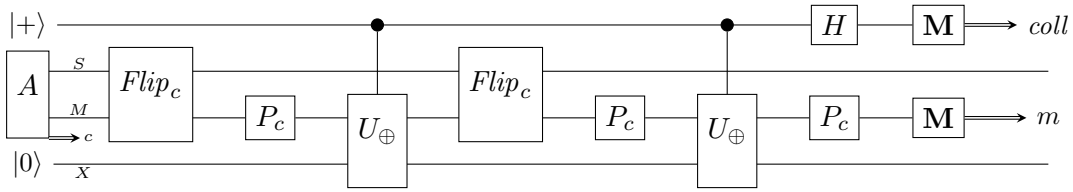


(15)

  Here $\mathbf{M}$ is a measurement in the computational basis, and $U_\oplus|x\rangle|y\rangle := |x\rangle|x \oplus y\rangle$ is a unitary operating on registers M, X (where X has the same dimension as M).
- If $coll = 0$, return $\perp$.
- If $coll = 1$, return m.

Claim 2 *With probability* 1*, D outputs $\perp$ or a half-collision of f.*

To show this claim, first recall that P_c and P_c^B commute. Thus also P_c and $Flip_c = 1 - 2P_c^B$ commute. Furthermore, since A is a valid adversary, the output of A is in $\operatorname{im} P_c$. And furthermore $U_\oplus$ and P_c are easily seen to commute. Thus we can perform the following transformations on the circuit (15) without changing the output probabilities: Add a projector P_c after A (on the M wire) and commute it until before after the second $U_\oplus$. Add a projector P_c after A and commute it until before the second $U_\oplus$. And add a projector P_c after A and commute it until before the first $U_\oplus$. We get the following result:



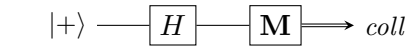
(16)

(Note that this circuit contains projectors, which may in general mean that the final quantum state does not have norm 1 and that thus the output probabilities for $coll, m$ do not add up to 1. In the present case, however, this is not the case because we argued that the output probabilities are the same as in (15).)

For a fixed value of c (as output by A in the above circuit), we distinguish three cases:

- Case 1 "$c \notin \operatorname{im} f$": This never happens because A is valid (and thus the register M contains a superposition of values m with $f(m) = c$).

- Case 2 "c has exactly one preimage under f": Let $m_0 := f^{-1}(c)$. Then P_c projects onto $|m_0\rangle$. Thus after an application of P_c on the M register, $U_\oplus$ has the same effect as the unitary $U_{m_0} : |y\rangle \to |y \oplus m_0\rangle$ on X. So we can replace both controlled $U_\oplus$ by controlled U_{m_0} in circuit (16). But the controlled U_{m_0} operated only on wires 1 and 4, while Flip_c and P_c operate on wires 2 and 3. Thus the first U_{m_0} can be commuted past Flip_c and P_c and then cancels out with the second controlled U_{m_0}. These transformations have not changed the probability distribution of *coll*. The first wire of the circuit has become the following:

  Thus $\mathit{coll} = 0$ with probability 1. Thus D outputs $\perp$ with probability 1.

- Case 3 "c has at least two preimages under f": m is the result of measuring the third wire in the computational basis. Just before that measurement, $P_c = \sum_{m:f(m)=c} |m\rangle\langle m|$ was applied to that wire. Thus $f(m) = c$ with probability 1. Since c has at least two preimages, this implies that m is a half-collision. Thus D outputs a half-collision or $\perp$ in this case (depending on the value of *coll*).

This proves Claim 2.

Claim 3 *With probability at least $\mu^2/4$, D outputs some $m \neq \perp$.*

To prove this claim, we first define some variables. For fixed c, let $|\Psi'_c\rangle$ be the output state (in registers S, M) of A when A outputs c. Let $|\Psi_c\rangle := |\Psi'_c\rangle|0\rangle$ (living in registers S, M, X). Recall that P_c^B operates on registers S and M, and that $U_\oplus : |x\rangle|y\rangle \to |x\rangle|x \oplus y\rangle$ operates on M and X.

Let

$$\alpha_c := \big\|P_c^B|\Psi_c\rangle\big\| \qquad \text{and} \qquad \beta_c := \big\|P_c^B U_\oplus|\Psi_c\rangle\big\|.$$

Note that by construction, α_c^2 is the probability that the left circuit in (14) returns $b = 1$, conditioned on the output c, and β_c^2 is the probability that the right circuit in (14) outputs $b = 1$, conditioned on c. (For β_c^2, notice that applying $U_\oplus$ has the same effect as measuring M.) Thus

$$\mu = \Big|\Pr[b = 1 : \text{left circuit}] - \Pr[b = 1 : \text{right circuit}]\Big| = \Big|\sum_c \Pr[c]\alpha_c^2 - \sum_c \Pr[c]\beta_c^2\Big| \quad (17)$$

where $\Pr[c]$ denotes the probability that A yields the classical output c.

Let $|\Phi_c\rangle$ denote the final state in the execution of the circuit (15) just before measuring *coll* and m. We have

$$\begin{aligned}|\Phi_c\rangle &= (H \otimes I_{SMX})\Big(\tfrac{1}{\sqrt{2}}|0\rangle \otimes \underbrace{\mathit{Flip}_c\mathit{Flip}_c}_{=I}|\Psi_c\rangle + \tfrac{1}{\sqrt{2}}|1\rangle \otimes U_\oplus \mathit{Flip}_c U_\oplus \mathit{Flip}_c|\Psi_c\rangle\Big) \\ &= \tfrac{1}{2}|0\rangle|\Psi_c\rangle + \tfrac{1}{2}|1\rangle|\Psi_c\rangle + \tfrac{1}{2}|0\rangle \otimes U_\oplus \mathit{Flip}_c U_\oplus \mathit{Flip}_c|\Psi_c\rangle - \tfrac{1}{2}|1\rangle \otimes U_\oplus \mathit{Flip}_c U_\oplus \mathit{Flip}_c|\Psi_c\rangle.\end{aligned}$$

Let $\operatorname{Re} x$ denote the real part of x. Since *coll* is the result of measuring the first wire in circuit (15), the probability $\Pr[coll = 1 : c]$ of having $coll = 1$ conditioned on a particular value of c is:

$$\begin{aligned}
&\Pr[coll = 1 : c] \\
&= \left\| \tfrac{1}{2}|\Psi_c\rangle - \tfrac{1}{2} U_\oplus Flip_c U_\oplus Flip_c |\Psi_c\rangle \right\|^2 \\
&= \Big\| \tfrac{1}{2}|\Psi_c\rangle - \tfrac{1}{2} U_\oplus U_\oplus |\Psi_c\rangle + U_\oplus U_\oplus P_c^B |\Psi_c\rangle \qquad \text{(since } Flip_c = I - 2P_c^B\text{)} \\
&\qquad\qquad + U_\oplus P_c^B U_\oplus |\Psi_c\rangle - 2 U_\oplus P_c^B U_\oplus P_c^B |\Psi_c\rangle \Big\|^2 \\
&= \left\| P_c^B |\Psi_c\rangle + U_\oplus P_c^B U_\oplus |\Psi_c\rangle - 2 U_\oplus P_c^B U_\oplus P_c^B |\Psi_c\rangle \right\|^2 \qquad \text{(since } U_\oplus U_\oplus = I\text{)} \\
&= \underbrace{\langle \Psi_c | P_c^B \cdot P_c^B | \Psi_c \rangle}_{=\alpha_c^2} + \underbrace{2 \operatorname{Re} \langle \Psi_c | P_c^B \cdot U_\oplus P_c^B U_\oplus | \Psi_c \rangle}_{=2 \operatorname{Re}\langle \Psi_c | U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle} - 4 \operatorname{Re} \langle \Psi_c | P_c^B \cdot U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle \\
&\quad + \underbrace{\langle \Psi_c | U_\oplus P_c^B U_\oplus \cdot U_\oplus P_c^B U_\oplus | \Psi_c \rangle}_{=\beta_c^2} - \underbrace{4 \operatorname{Re} \langle \Psi_c | U_\oplus P_c^B U_\oplus \cdot U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle}_{=4 \operatorname{Re}\langle \Psi_c | U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle} \\
&\quad + \underbrace{4 \langle \Psi_c | P_c^B U_\oplus P_c^B U_\oplus \cdot U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle}_{=4 \operatorname{Re}\langle \Psi_c | P_c^B U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle} \\
&= \alpha_c^2 + \beta_c^2 - 2 \operatorname{Re} \langle \Psi_c | U_\oplus P_c^B U_\oplus P_c^B | \Psi_c \rangle \\
&\geq \alpha_c^2 + \beta_c^2 - 2 \big| \underbrace{\langle \Psi_c | U_\oplus P_c^B}_{\text{norm is } \beta_c} \cdot U_\oplus \cdot \underbrace{P_c^B | \Psi_c \rangle}_{\text{norm is } \alpha_c^2} \big| \\
&\geq \alpha_c^2 + \beta_c^2 - 2\beta_c\alpha_c = (\alpha_c - \beta_c)^2. \qquad (18)
\end{aligned}$$

We can now bound the probability $\Pr[coll = 1]$ that in circuit (15), we measure $coll = 1$:

$$\begin{aligned}
\Pr[coll = 1] &= \sum_c \Pr[c] \Pr[coll = 1 : c] \overset{(18)}{\geq} \sum_c \Pr[c](\alpha_c - \beta_c)^2 \overset{(*)}{\geq} \tfrac{1}{4} \sum_c \Pr[c](\alpha_c^2 - \beta_c^2)^2 \\
&\overset{(**)}{\geq} \tfrac{1}{4} \Big(\sum_c \Pr[c](\alpha_c^2 - \beta_c^2) \Big)^2 \overset{(17)}{=} \tfrac{1}{4}\mu^2.
\end{aligned}$$

Here $(*)$ uses the fact that for $\alpha, \beta \in [0,1]$, $|\alpha - \beta| \geq \frac{1}{2}|\alpha^2 - \beta^2|$.[12] And $(**)$ uses Jensen's inequality. Since D outputs $\perp$ only if $coll = 0$, it follows that D outputs $m \neq \perp$ with probability at least $\frac{1}{4}\mu^2$. Claim 3 follows.

From Claims 2 and 3, it follows that D has advantage $\geq \mu^2/4$ against the half-collision resistance of f. By inspection of the circuit of D, we see that D invokes one instance of A and four instances of B_c. And B_c invokes B and performs at most one more evaluation of f. Thus D performs at most $4q + 4$ queries to f, when f is given as an oracle and (A, B) performs q queries. □

[12] Proof: $|\alpha^2 - \beta^2| \leq \max_{x \in [0,1]} \frac{\partial x^2}{\partial x} \cdot |\alpha - \beta| = \max_{x \in [0,1]} 2x \cdot |\alpha - \beta| = 2|\alpha - \beta|$.

Corollary 32 (Distinguishing random functions and injections [Zha13]) *Assume that $|X| \leq |Y|$. Let $H : X \to Y$ be a uniformly random function. Let $\hat{H} : X \to Y$ be a uniformly random injection. Then for any q-query adversary A,*

$$\left|\Pr[A^H = 1] - \Pr[A^{\hat{H}} = 1]\right| \in O(q^3/|Y|).$$

Proof. [Zha13, Section 3.1] shows this lemma for the case $|X| = |Y|$. For the general case, let $H' : Y \to Y$ be a random function and $\hat{H}' : Y \to Y$ be a random permutation. Then $H' \circ \hat{H}$ has the same distribution as H, and $\hat{H}' \circ \hat{H}$ has the same distribution as $\hat{H}$. Since the corollary holds for $|X| = |Y|$, we have that H' and $\hat{H}'$ can be distinguished with probability at most $O(q^3/|Y|)$ by a q-query adversary, and thus $H' \circ \hat{H}$ and $\hat{H}' \circ \hat{H}$ can be distinguished with probability at most $O(q^3/|Y|)$. Thus H and $\hat{H}$ can be distinguished with probability at most $O(q^3/|Y|)$. □

Lemma 33 *Assume $|X| \leq |Y|$. Then H is collapsing with advantage $O(q^3/|Y|)$.*

Proof. Let $\hat{H} : X \to Y$ be a random injective function. Let $\mathsf{Game}_1, \mathsf{Game}_2$ refer to the games from Definition 21, and $\widehat{\mathsf{Game}_1}, \widehat{\mathsf{Game}_2}$ refer to those games with $\hat{H}$ instead of H. Since $\hat{H}$ is injective, by Lemma 22, $\hat{H}$ is collapsing with advantage 0, i.e., $\Pr[b = 1 : \widehat{\mathsf{Game}_1}] = \Pr[b = 1 : \widehat{\mathsf{Game}_2}]$.

By Corollary 32, an adversary making q queries can distinguish H and $\hat{H}$ only with probability $O(q^3/|Y|)$. Thus $\left|\Pr[b = 1 : \mathsf{Game}_i] - \Pr[b = 1 : \widehat{\mathsf{Game}_i}]\right| \in O(q^3/|Y|)$ for $i = 1, 2$. Altogether $\left|\Pr[b = 1 : \mathsf{Game}_1] - \Pr[b = 1 : \mathsf{Game}_2]\right| \in O(q^3/|Y|)$. □

For the next lemma, we fix some notation first: $[N] := \{1, \ldots, N\}$. For functions $f : [M] \to [N]$ and $g : [M'] \to [N]$, let $f + g : [M + M'] \to [N]$ be defined via $(f+g)(x) := f(x)$ for $x = 1, \ldots, M$ and $(f+g)(x) = g(x-M)$ for $x = M+1, \ldots, M+M'$. For functions $f : [M] \to [N]$ and $g : [M'] \to [N']$, let $f|g : [M + M'] \to [N + N']$ be defined via $(f|g)(x) := f(x)$ for $x = 1, \ldots, M$ and $(f|g)(x) := g(x - M) + N$ for $x = M + 1, \ldots, M + M'$.

Lemma 34 *Assume that $M \geq N$. Let $\hat{f}, \hat{g} : [N] \to [N]$ and $\hat{h} : [M] \to [M]$ and $\hat{\varphi} : [N + M] \to [N + M]$ be uniformly distributed permutations (all independent), and let $H : [2N + M] \to [N + M]$ be a uniformly distributed function.*

Then for any q-query adversary A,

$$\left|\Pr[A^H = 1] - \Pr[A^{\hat{\varphi} \circ ((\hat{f}+\hat{g})|\hat{h})} = 1]\right| \in O(q^3/N).$$

Proof. Let the following functions be uniformly distributed:

- $f, g : [N] \to [N]$.
- $h : [M] \to [M]$.
- $\varphi : [N + M] \to [N + M]$.
- $v : [N] \to [N + M]$ and $w : [M] \to [N + M]$.
- $a : [N] \to [2N]$ and $b : [2N] \to [2N]$ and $c : [2N] \to [N + M]$.

- $H : [2N + M] \to [N + M]$.

Let $\hat{f}, \hat{g}, \hat{h}, \hat{\varphi}, \hat{v}, \hat{w}, \hat{a}, \hat{b}, \hat{c}$ be uniformly distributed *injective* functions (with the same domains and ranges as the functions above). All functions are chosen independently.

For two functions f, g, let $f \approx g$ mean that that $\left|\Pr[A^f = 1] - \Pr[A^g = 1]\right| \in O(q^3/N)$ for any q-query adversary. And let $f \equiv g$ mean that f and g have the same distribution. We will show the following facts:

$$\varphi \approx \hat{\varphi} \qquad v \approx \hat{v} \qquad c \approx \hat{c} \tag{19}$$

$$w \circ \hat{h} \equiv w \qquad \hat{c} \circ \hat{a} \equiv \hat{v} \qquad \hat{c} \circ \hat{b} \equiv \hat{c} \tag{20}$$

$$\forall \alpha \text{ with range } [N], \forall \beta \text{ with range } [M] : \varphi \circ (\alpha|\beta) \equiv (v \circ \alpha) + (w \circ \beta) \tag{21}$$

$$\hat{a} \circ (\hat{f} + \hat{g}) \approx \hat{b} \tag{22}$$

$$c + w \equiv H \tag{23}$$

From Corollary 32 we get (19). (Corollary 32 gives the bound $O(q^3/(N + M))$ for the distinguishing probability. Since $M \geq N$ this also implies the desired bound $O(q^3/N)$.)

The equations in (20) are immediate by definition of $w, \hat{h}, w, \hat{c}, \hat{a}, \hat{v}, \hat{b}$.

In (21), first notice that $\varphi \equiv v + w$ and thus $\varphi \circ (\alpha|\beta) \equiv (v + w) \circ (\alpha|\beta)$. Let the range of α be $[N']$ and that of β be $[M']$. By case distinction over $x \in \{1, \ldots, N'\}$ and $x \in \{N' + 1, \ldots, N' + M'\}$ we check $(v + w) \circ (\alpha|\beta)(x) = (v \circ \alpha) + (w \circ \beta)(x)$. Then (21) follows.

To show (22), let $f_1, g_1 : [N] \to [2N]$ be two uniformly random functions conditioned on having identical range. Let $f_2, g_2 : [N] \to [2N]$ be two uniformly random functions conditioned on having disjoint range. [Zha13, Theorem 4.1] states that a q-query adversary distinguishes f_1, g_1 from f_2, g_2 with probability at most $O(q^3/N)$. (This is the "Set Equality Problem".) As a consequence, a q-query adversary distinguishes $f_1 + g_1$ and $f_2 + g_2$ with probability at most $O(q^3/N)$. One can verify that $f_1 + g_1 \equiv \hat{a} \circ (\hat{f} + \hat{g})$ and $f_2 + g_2 \equiv \hat{b}$. Thus an adversary distinguishing $\hat{a} \circ (\hat{f} + \hat{g})$ and $\hat{b}$ also distinguishes $f_1 + g_1$ and $f_2 + g_2$ with the same probability. (22) follows.

And (23) is immediate.

We then have:

$$\begin{aligned}
\hat{\varphi} \circ ((\hat{f} + \hat{g})|\hat{h}) &\overset{(19)}{\approx} \varphi \circ ((\hat{f} + \hat{g})|\hat{h}) \overset{(21)}{\equiv} (v \circ (\hat{f} + \hat{g})) + (w \circ \hat{h}) \overset{(20)}{\equiv} (v \circ (\hat{f} + \hat{g})) + w \\
&\overset{(19)}{\approx} (\hat{v} \circ (\hat{f} + \hat{g})) + w \overset{(20)}{\equiv} (\hat{c} \circ \hat{a} \circ (\hat{f} + \hat{g})) + w \overset{(22)}{\approx} (\hat{c} \circ \hat{b}) + w \\
&\overset{(20)}{\equiv} \hat{c} + w \overset{(19)}{\approx} c + w \overset{(23)}{\equiv} H.
\end{aligned}$$

(In some of these steps, we implicitly perform a reduction to (19) or (22). E.g., in $\hat{c} + w \approx c + w$, we use queries to $\hat{c}$ to simulate queries to $\hat{c} + w$. In some of these reductions, it is not obvious that this simulation does not double the number of queries from q to $2q$. However, $O(q^3/N) = O((2q)^3/N)$, so (19) or (22) apply with the same asymptotic bound to adversaries making $2q$ queries.)

Note that $\approx$ is transitive (the distinguishing advantage of the adversary may double, but will still be in $O(q^3/N)$). Thus $\hat{\varphi} \circ ((\hat{f} + \hat{g})|\hat{h}) \approx H$ which proves the lemma. □

Lemma 35 *Assume that* $|Y| = \lceil \frac{2}{3}|X| \rceil$. *Then* H *is collapsing with advantage* $O(\sqrt{q^3/|X|})$.

Proof sketch: For simplicity, we consider the case $|Y| = 2N$, $|X| = 3N$. Then, by Lemma 34 with $M := N$, H is indistinguishable from $H^* := \hat{\varphi} \circ \big((\hat{f} + \hat{g})|\hat{h}\big)$. Furthermore, for a random permutation π, H and $H \circ \pi$ are identically distributed, and $H \circ \pi$ is indistinguishable from $H^* \circ \pi$. Thus it is sufficient to show that $H^* \circ \pi$ is collapsing. In turn, by Lemma 31, it is sufficient to show that $H^* \circ \pi$ is half-collision resistant. To show that, observe that the half-collisions of H^* are the inputs $1, \ldots, 2N$, but not $2N+1, \ldots, 3N$. Thus the half-collisions of $H^* \circ \pi$ are $P := \pi^{-1}(\{1, \ldots, 2N\})$. So, the half-collision resistance adversary has to find elements of P, without false positives, while given oracle access to $H^* \circ \pi$. But $H^* \circ \pi$ is indistinguishable from $H \circ \pi$, so the adversary would also be able to find elements in P given $H \circ \pi$. Since $H \circ \pi$ is a random function, independent of P, the adversary cannot do that without getting false positives. Hence $H^* \circ \pi$ is half-collision resistant and thus collapsing. Hence H is collapsing.

Proof. Let $N := |X| - |Y|$ and $M := 2|Y| - |X|$. Then $M - N = 3|Y| - 2|X| \geq 3 \cdot \frac{2}{3}|X| + 2|X| = 0$, hence $M \geq N$.

Since $|X| = 2N + M$ and $|Y| = N + M$, we can assume that $X = [2N + M]$ and $Y = [N + M]$. Thus, by Lemma 34, H is indistinguishable from $H^* := \hat{\varphi} \circ ((\hat{f} + \hat{g})|\hat{h})$ where $\hat{\varphi}, \hat{f}, \hat{g}, \hat{h}$ are random permutations as in Lemma 34. More precisely, a $(4q+4)$-query adversary distinguishes H and H^* with probability at most $\delta \in O((4q+4)^3/N) = O(q^3/N)$.

Let $\pi : X \to X$ be a uniformly random permutation. Then also $H \circ \pi$ and $H^* \circ \pi$ are distinguished with probability at most δ. And since $H \circ \pi$ and H have the same distribution, H and $H^* \circ \pi$ are distinguished with probability at most δ.

Assume a q-query adversary (A, B) against the collapsing property of H with some advantage ε. We will prove that $\varepsilon \in O(\sqrt{q^3/N}) = O(\sqrt{q^3/|X|})$.

Since H and $H^* \circ \pi$ cannot be distinguished with probability greater than δ, it follows that (A, B) has advantage $\geq \varepsilon - 2\delta$ against the collapsing property of $H^* \circ \pi$. (Because each of the probabilities in Definition 21 can change by at most δ.)

Thus, by Lemma 31, there is an adversary D with advantage $\geq (\varepsilon - 2\delta)^2/4$ against the half-collision resistance of $H^* \circ \pi$. D makes at most $4q + 4$ queries. That is:

$$\Pr[D^{H^* \circ \pi} \text{ outputs half-collision}] \geq \tfrac{(\varepsilon - 2\delta)^2}{4}, \quad \Pr[D^{H^* \circ \pi} \text{ outputs non-half-collision}] = 0$$

By definition of $+$ and $|$ from Lemma 34, and since $\hat{\varphi}, \hat{f}, \hat{g}, \hat{h}$ are all permutations, the half-collisions of $H^* = \hat{\varphi} \circ ((\hat{f} + \hat{g})|\hat{h})$ are $1, \ldots, 2N$, while $2N+1, \ldots, 2N+M$ are the non-half-collisions. Thus the half-collisions of $H^* \circ \pi$ are $P := \pi^{-1}(\{1, \ldots, 2N\})$. Thus we have

$$\Pr[D^{H^* \circ \pi} \in P] \geq \tfrac{(\varepsilon - 2\delta)^2}{4}, \qquad \Pr[D^{H^* \circ \pi} \in X \setminus P] = 0.$$

Since H^* is δ-indistinguishable from H by $(4q+4)$-query adversaries (in particular by D), it follows that

$$p_{hit} := \Pr[D^{H \circ \pi} \in P] \geq \tfrac{(\varepsilon - 2\delta)^2}{4} - \delta, \qquad p_{miss} := \Pr[D^{H \circ \pi} \in X \setminus P] \leq \delta. \tag{24}$$

Since π is a permutation and H is a random function, $(H \circ \pi, \pi)$ and (H, π) are identically distributed. Thus

$$p_{hit} = \Pr[D^H \in P], \qquad p_{miss} = \Pr[D^H \in X \setminus P].$$

(Recall that $P = \pi^{-1}(\{1, \ldots, 2N\})$.) Since P is a uniformly random subset of X of size $2N$, and independent of H, we have:

$$p_{hit} = \frac{|P|}{|X|} \Pr[D^H \in X], \qquad p_{miss} = \frac{|X \setminus P|}{|X|} \Pr[D^H \in X]. \tag{25}$$

Then

$$\tfrac{(\varepsilon-2\delta)^2}{4} - \delta \overset{(24)}{\leq} p_{hit} \overset{(25)}{=} \frac{|P|}{|X|} \frac{|X|}{|X \setminus P|} p_{miss} = \frac{2N}{M} p_{miss} \overset{(24)}{\leq} \frac{2N}{M} \delta \overset{M \geq N}{\leq} 2\delta.$$

Solving this inequality for ε, we get $\varepsilon \leq \sqrt{12\delta} + 2\delta \in O(\sqrt{q^3/N}) = O(\sqrt{q^3/|X|})$.

Since ε was the advantage of an arbitrary q-query adversary (A, B) against the collapsing property of H, the lemma follows. □

Theorem 36 *Let Y be finite, and $X \subseteq \{0,1\}^*$ (finite or infinite). Then $H : X \to Y$ is collapsing with advantage $O(\sqrt{q^3/|Y|})$.*

Proof sketch: H is indistinguishable from a composition $f_n \circ \cdots \circ f_1$ of random functions $f_n : X_n \to Y_n$ with $|X_{n+1}| = |Y_n| = \frac{2}{3}|X_n|$. By Lemma 35, each f_n is collapsing. Thus, by Corollary 26, $f_n \circ \cdots \circ f_1$ is collapsing and hence H is collapsing.

Proof. We first consider the case that $X = [M]$ for some M.

Since we are interested only in the asymptotic complexity, we can further assume $|Y| \geq 11$. And since the case $|X| \leq |Y|$ is covered by Lemma 33, we can assume $|X| > |Y|$.

Let $t(x) := \lceil \frac{2}{3} x \rceil$. Let $n \geq 0$ be the smallest integer such that $t^n(M) \leq |Y|$. (Such integer always exists: if $t^i(M) > |Y| \geq 11$, then $t^{i+1}(M) < t^i(M)$, hence $t^i(M)$ decreases until it becomes smaller-equal $|Y|$.)

Further, since $t^{i-1}(M) > |Y| \geq 11$ for $i = 1, \ldots, n$ we have that

$$t^i(M) = \lceil \tfrac{2}{3} t^{i-1}(M) \rceil \leq \tfrac{2}{3} t^{i-1}(M) + 1 \leq \tfrac{2}{3} t^{i-1}(M) + \tfrac{1}{12} t^{i-1}(M) = \tfrac{3}{4} t^{i-1}(M)$$

for $i = 1, \ldots, n$. Thus

$$t^{n-i}(M) \geq (\tfrac{4}{3})^i t^n(M) \geq (\tfrac{4}{3})^i \tfrac{2}{3} t^{n-1}(M) \geq (\tfrac{4}{3})^i \tfrac{2}{3} |Y| \quad \text{for} \quad i = 0, \ldots, n. \tag{26}$$

For $i = 1, \ldots, n$, let $f_i : [t^{i-1}(M)] \to [t^i(M)]$ be a uniformly random function.

Let $h : [t^n(M)] \to Y$ be a uniformly random function. (h and all f_i are independent.)

We define $H^* : X \to Y$ as $H^* := h \circ f_n \circ \cdots \circ f_1$.

We will first show that H^* is collapsing. By Lemma 35, there is a constant $C > 0$ such that each f_i is collapsing with advantage $C\sqrt{q^3/t^{i-1}(M)}$.

And by Lemma 33, h is collapsing with advantage $C'q^3/|Y|$ for some constant $C' > 0$. (Note that the domain of h is smaller than its range by definition of n.)

Let ε_i denote a tight upper bound such that $f_n \circ \cdots \circ f_{n-i+1}$ is collapsing with advantage ε_i for $(q+2)$-query adversaries. Then $\varepsilon_0 = 0$ since the identity is collapsing with advantage 0 by Lemma 22.

By Lemma 25 (with $f := f_{n-i+1}$ and $g := f_n \circ \cdots \circ f_{n-i+2}$), we then get for $i \geq 1$:

$$\underbrace{\varepsilon_i}_{\substack{\text{advantage against} \\ g\circ f = f_n\circ\cdots\circ f_{n-i+1} \\ \text{in } q+2 \text{ queries}}} \leq \underbrace{\varepsilon_{i-1}}_{\substack{\text{advantage against} \\ g = f_n\circ\cdots\circ f_{n-i+2} \\ \text{in } q+2 \text{ queries}}} + \underbrace{C\sqrt{(q+4)^3/t^{n-i}(M)}}_{\substack{\text{advantage against } f=f_{n-i+1} \\ \text{in } q+4 \text{ queries}}}. \tag{27}$$

Note that when applying Lemma 25, both the adversary for the left and the right summand in (27) makes two more queries to $f = f_{n-i+1}$ than the $(q+2)$-query adversary having the advantage ε_i against $g \circ f$. However, since all f_i are chosen independently, the adversary in the left summand (who attacks $g = f_n \circ \cdots \circ f_{n-i+2}$) can chose $f = f_{n-i+1}$ himself. Thus the advantage in the left summand is with respect to adversaries that make only $q+2$ queries to g.

From (27) we get (using $\varepsilon_0 = 0$):

$$\begin{aligned}\varepsilon_n &\leq \sum_{i=1}^{n} C\sqrt{(q+4)^3/t^{n-i}(M)} \overset{(26)}{\leq} \sum_{i=1}^{n} C\sqrt{(q+4)^3/(\tfrac{4}{3})^i \tfrac{2}{3}|Y|} \\ &\leq C\sqrt{\tfrac{3}{2}(q+4)^3/|Y|} \cdot \sum_{i=1}^{\infty} (\tfrac{3}{4})^{i/2} \in O(\sqrt{q^3/|Y|}).\end{aligned}$$

By definition, ε_n upper bounds the advantage of a $(q+2)$-query adversary against $f_n \circ \cdots \circ f_1$. And the advantage of a q-query adversary against h is $\leq C'q^3/|Y|$ (see above). Thus by Lemma 25 (with $g := h$ and $f := f_n \circ \cdots \circ f_1$) the advantage of a q-query adversary against $H^* = h \circ f_n \circ \cdots \circ f_1$ is at most

$$\delta := C'q^3/|Y| + \varepsilon_n \in O(\sqrt{q^3/|Y|}).$$

We have shown that H^* is collapsing with advantage $O(\sqrt{q^3/|Y|})$. To show that H is collapsing, we will first show that H and H^* are indistinguishable.

Let $H_i : [t^i(M)] \to Y$ denote a uniformly random function. [Zha12, Corollary VII.5 (Small range distributions)] shows that for independent uniformly random functions $a : A \to B$, $b : B \to C$, and $c : A \to C$, a q-query adversary can distinguish $b \circ a$ from c with probability at most $27q^3/|B|$. Hence for $i = 1, \ldots, n$, $H_i \circ f_i$ and H_{i-1} can be distinguished with probability at most $27q^3/t^i(M)$. Hence $H_i^* := H_i \circ f_i \circ \cdots \circ f_1$ and $H_{i-1}^* = H_{i-1} \circ f_{i-1} \circ \cdots \circ f_1$ can be distinguished with probability at most $27q^3/t^i(M)$. Thus H_0^* and H_n^* can be distinguished with probability at most

$$\gamma := \sum_{i=1}^{n} 27q^3/t^i(M) \overset{(26)}{\leq} \sum_{i=1}^{n} 27q^3 \tfrac{3}{2}(\tfrac{3}{4})^{n-i}/|Y| \leq \tfrac{81}{2}q^3/|Y| \cdot \sum_{i=0}^{\infty} (\tfrac{3}{4})^i \in O(q^3/|Y|).$$

Note that H_n is identically distributed to h, hence H_n^* is identically distributed to H^*. And $H_0^* = H_0 : [M] \to Y$ is identically distributed to H. Hence H^* and H can be distinguished with probability at most γ.

Now we can prove that H is collapsing. Since H^* is collapsing with advantage $\leq \delta$, we have that for any q-query adversary,

$$\big|\Pr[b = 1 : \mathsf{Game}_1 \text{ using } H^*] - \Pr[b = 1 : \mathsf{Game}_2 \text{ using } H^*]\big| \leq \delta. \tag{28}$$

Here $\mathsf{Game}_1, \mathsf{Game}_2$ are the games from Definition 21. Since H and H^* are indistinguishable (with advantage $\leq \gamma$), replacing H^* by H changes the two probabilities in (28) by at most γ. Hence we get

$$\big|\Pr[b = 1 : \mathsf{Game}_1 \text{ using } H] - \Pr[b = 1 : \mathsf{Game}_2 \text{ using } H]\big| \leq \delta + 2\gamma$$

Thus H is collapsing with advantage $\leq \delta + 2\gamma \in O(\sqrt{q^3/|Y|})$. The theorem follows for the case of $|X| = [M]$.

The case for finite X follows immediately by setting $M := |X|$. (This is just a relabeling of the elements of X.)

Finally, we prove the theorem for infinite $X \subseteq \{0,1\}^*$: Fix a q-query adversary (A, B). Let μ denote the advantage of (A, B) against the collapsing property of $H : X \to Y$. Let $X_i := \{x \in X : |x| \leq i\}$. Let $H_i : X_i \to Y$ be a uniformly random function. Let μ_i denote the advantage of (A, B) against the collapsing property of H_i. Then $\mu_i \to \mu$ for $i \to \infty$. Furthermore, we have already shown that against a finite $H_i : X_i \to Y$, the advantage is at most $\delta + 2\gamma$. Hence $\mu_i \leq \delta + 2\gamma$. Note that $\delta + 2\gamma$ does not depend on the size of X_i, only on the size of Y. Thus $\mu_i \leq \delta + 2\gamma$ implies $\mu \leq \delta + 2\gamma$. Thus H is collapsing with advantage $\leq \delta + 2\gamma \in O(\sqrt{q^3/|Y|})$ for infinite X. □

7 Zero-knowledge arguments of knowledge

In this section, we study the security of sigma-protocols. A sigma-protocol is a specific kind three-round proof system in which the verifier's message consists only of random bits. Sigma-protocols play an important role in classical constructions of zero-knowledge proof systems for two reasons: For a number of simple but important languages, sigma-protocols exist. And given sigma-protocols for simple languages, there are efficient constructions for more complex languages. (There are constructions for conjunctions and disjunctions of sigma-protocols, as well as more complex threshold constructions [CDS94].)

In the classical setting, it is relatively simple to give conditions under which sigma-protocols are zero-knowledge proofs of knowledge. In the quantum setting, however, analyzing the security of sigma-protocols turns out to be much harder. Watrous [Wat09] presented a rewinding technique for proving the zero-knowledge property of sigma-protocols (see also Theorem 39 below). Unruh [Unr12] showed that sigma-protocols are quantum proofs of knowledge under a specific additional condition called "strict soundness". This condition requires that the third message ("response") in a valid interaction is uniquely determined by the first two. However, strict soundness is a strong

additional assumption. [Unr12] showed how to achieve strict soundness by committing to the response already in the first message. However, the commitment scheme used for this needed to be perfectly-binding (actually, it needed to satisfy a somewhat stronger property, called "strict binding"). In particular, this implies that the commitment scheme cannot be information-theoretically hiding (hence the resulting protocol cannot be statistical zero-knowledge), and we cannot have short commitments (a perfectly-binding commitment will always be at least as long as the message inside).

Furthermore, Ambainis, Rosmanis, and Unruh [ARU14] showed that the condition of strict soundness is necessary, at least relative to an oracle. They also showed that even if we assume that strict soundness holds, but only against computationally limited adversaries,[13] the resulting sigma-protocol will, in general, not be a quantum argument of knowledge.[14] Even more, it might not even be a quantum argument. That is, a computationally limited adversary can successfully prove a wrong statement.

In this section we show how we can use collapse-binding commitments as a drop-in replacement for the perfectly-binding commitments in the construction from [Unr12]. One particular consequence is that given collapse-binding hash functions we can construct three-round statistical zero-knowledge quantum arguments of knowledge from sigma-protocols (without using a common-reference string). This assumes the sigma-protocol is statistical honest-verifier zero-knowledge and has special soundness. And that the challenge space (the set from which the verifier picks his random message) is polynomially-bounded. These properties, however, are also needed in the classical setting.

7.1 Interactive proof systems

An interactive proof system (P, V) for some relation R consists of two interactive quantum machines P and V that get classical inputs $(x, w) \in R$ and x, respectively. Afterwards, V outputs a bit. For formal definitions see [Unr12]. (In general, P and V can exchange quantum messages, but our concrete constructions below will be classical.)

We consider two important properties of interactive proof systems: First, we want them to be arguments of knowledge. Informally, they should convince the verifier that the prover knows a witness w for the statement x (i.e., $(x, w) \in R$). Second, we want them to be zero-knowledge. Informally, the proof should not leaks anything about the witness besides its existence.

Quantum arguments of knowledge. The following definition of quantum arguments of knowledge follows the definition from [Unr15], with one difference: we have formulated security against uniform malicious provers. That is, while in [Unr15] the statement x and the auxiliary input $|\Psi\rangle$ are all-quantified, in our setting they are chosen by an quantum-polynomial-time algorithm Z. The reason we consider only uniform malicious provers here is: A non-uniform adversary can break any non-interactive commitment

[13] I.e., it is hard to find two different valid interactions where the first two messages are equal but the response is different.

[14] Argument and argument of knowledge are the variants of proof and proof of knowledge that consider a computationally limited malicious prover.

(with classical messages) that is not already perfectly-binding. (Namely, the auxiliary input can simply contain one commitment and two different openings.) Thus, since we consider only non-interactive commitments in this paper, we need a uniform definition of quantum arguments of knowledge. For a motivation of the remaining definitional choices, see [Unr15].

Definition 37 (Quantum Arguments of Knowledge) *We call an interactive proof system* (P,V) *for a relation* R (uniformly) quantum-computationally extractable with knowledge error κ *if there exists a constant* $d>0$*, a polynomially-bounded function* $p>0$*, and a quantum-polynomial-time oracle algorithm* K *such that for any unitary quantum-polynomial-time algorithm* P^**, for any polynomial* ℓ*, and for any quantum-polynomial-time algorithm* Z *(input generator), there exists a negligible* μ *such that for any security parameter* $\eta\in\mathbb{N}$*, we have that*

$$\begin{aligned}\Pr[\langle\mathsf{P}^*(1^\eta,x,Z),\mathsf{V}(1^\eta,x)\rangle=1:(x,Z)\leftarrow\mathsf{Z}(1^\eta)]\geq\kappa(\eta)\implies\\ \Pr[(x,w)\in R:(x,Z)\leftarrow\mathsf{Z}(1^\eta),w\leftarrow\mathsf{K}^{\mathsf{P}^*(1^\eta,x,Z)}(1^\eta,x)]\\ \geq\tfrac{1}{p(\eta)}\Big(\Pr\big[\langle\mathsf{P}^*(1^\eta,x,Z),\mathsf{V}(1^\eta,x)\rangle=1:(x,Z)\leftarrow\mathsf{Z}(1^\eta)\big]-\kappa(\eta)\Big)^d-\mu(\eta).\end{aligned}$$

Here $\langle\mathsf{P}^*(1^\eta,x,Z),\mathsf{V}(1^\eta,x)\rangle$ *is the output of* V *after an interaction between* P^* *and* V *on the respective inputs* x *and* Z*.* Z *is a quantum register,* x *is classical, both initialized using the algorithm* Z*. And* $\mathsf{K}^{\mathsf{P}^*(1^\eta,x,Z)}$ *refers to an execution of* K *with black-box access to* $\mathsf{P}^*(1^\eta,x,Z)$*. That is,* K *can apply the unitary* U_x *describing the prover* P^* *and its inverse* $U_x^\dagger$*. (See [Unr12] for a more detailed description of that black-box execution model.)*

Quantum zero-knowledge. Roughly speaking, (P,V) is *quantum-computationally zero-knowledge* iff for any quantum-polynomial-time malicious verifier V^*, there exists a quantum-polynomial-time simulator S such that for any $(x,w)\in R$, the output state of S is quantum computationally indistinguishable from the from the output state of V^* in an interaction with $\mathsf{P}(1^\eta,x,w)$.

Similarly, *quantum statistical zero-knowledge* is defined in the same way, except that V^* is not required to be quantum-polynomial-time.

We will not use the definition of quantum zero-knowledge directly, only the imported Theorem 39 from [Unr15] will refer to it. We therefore omit the formal definition and refer to [Unr15].

7.2 Sigma-protocols

We now introduce sigma-protocols (following [Unr14a] with modifications as mentioned in the footnotes). The notions are like the standard classical definitions, all that was done to adopt them to the quantum setting was to make the adversary quantum-polynomial-time.

A *sigma-protocol* for a relation R is a three-message proof system. It is described by its challenge space N_z (where $|N_z|\geq 2$), a classical-polynomial-time prover (P_1,P_2) and a deterministic classical-polynomial-time verifier V. The first message from the prover is

$a \leftarrow P_1(1^\eta, x, w)$ and is called the *commitment*, the uniformly random reply from the verifier is $z \xleftarrow{\$} N_z$ (called *challenge*), and the prover answers with $r \leftarrow P_2(1^\eta, x, w, z)$ (the *response*). We assume P_1, P_2 to share state. Finally $V(1^\eta, x, a, z, r)$ outputs whether the verifier accepts.

Definition 38 (Properties of sigma-protocols) *Let $\Sigma = (N_z, P_1, P_2, V)$ be a sigma-protocol. We define:*

- ***Completeness:*** *For any quantum-polynomial-time algorithm A,*

$$\Pr[(x,w) \in R \wedge ok = 0 : (x,w) \leftarrow A(1^\eta), \\ a \leftarrow P_1(1^\eta, x, w), z \xleftarrow{\$} N_z, r \leftarrow P_2(1^\eta, x, w, z), ok \leftarrow V(1^\eta, x, a, z, r)]$$

  is negligible.

- ***Computational special soundness:*** *There is a quantum-polynomial-time algorithm E_Σ (the extractor)*[15] *such that for any quantum-polynomial-time A, we have that*

$$\Pr[(x,w) \notin R \wedge z \neq z' \wedge ok = ok' = 1 : (x, a, z, r, z', r') \leftarrow A(1^\eta), \\ ok \leftarrow V(1^\eta, x, a, z, r),\ ok' \leftarrow V(1^\eta, x, a, z', r'),\ w \leftarrow E_\Sigma(1^\eta, x, a, z, r, z', r')]$$

  is negligible.

- ***Honest-verifier zero-knowledge (HVZK):*** *There is a quantum-polynomial-time algorithm S_Σ (the simulator)*[16] *such that for any quantum-polynomial-time algorithm A and any polynomial ℓ, the following is negligible for all $(x,w) \in R$ with $|x|, |w| \leq \ell(\eta)$ and all states $|\Psi\rangle$:*

$$\big|\Pr[b = 1 : a \leftarrow P_1(1^\eta, x, w),\ z \xleftarrow{\$} N_z,\ r \leftarrow P_2(1^\eta, x, w, z),\ b \leftarrow A(1^\eta, |\Psi\rangle, a, z, r)] \\ - \Pr[b = 1 : (a, z, r) \leftarrow S_\Sigma(1^\eta, x),\ b \leftarrow A(1^\eta, |\Psi\rangle, a, z, r)]\big|$$

- ***Statistical honest-verifier zero-knowledge (SHVZK):*** *Like HVZK, except that we quantify over computationally unlimited A (not only quantum-polynomial-time A).*

Note that the above are the standard conditions expected from sigma-protocols in the classical setting. In contrast, for a sigma-protocol to be a *quantum* proof of knowledge, a much more restrictive condition is required, strict soundness [Unr12, ARU14]. We show below how to circumvent this necessity by adding collapse-binding commitments to the sigma-protocol (at least when we only need a quantum *argument* of knowledge).

Remark 1. Any sigma-protocol (N_z, P_1, P_2, V) can be seen as an interactive proof (P, V) in a natural way: P sends the output a of P_1 to V. V picks $z \xleftarrow{\$} N_z$ and sends it to P. P sends the resulting output r of P_2 to V. V checks the triple (a, z, r) using V.

The following theorem is shown in [Unr15]:

[15] [Unr14a] requires a classical E_Σ here. By allowing E_Σ to be quantum here, we weaken the notion of computational special soundness slightly, and thus strengthen our results below.

[16] [Unr14a] requires a classical S_Σ here. By allowing E_Σ to be quantum here, we weaken the notion of HVZK/SHVZK slightly, and thus strengthen our results below.

Theorem 39 (HVZK implies zero-knowledge [Unr15]) *Let $\Sigma = (N_z, P_1, P_2, V)$ be a sigma-protocol. We consider Σ as an interactive proof (P, V), see Remark 1.*

If $|N_z|$ is polynomially-bounded and is SHVZK, then Σ is quantum statistical zero-knowledge.

If $|N_z|$ is polynomially-bounded and Σ is HVZK, then Σ is quantum computational zero-knowledge.

Due to this theorem, it will be sufficient to verify that the sigma-protocols we construct are HVZK/SHVZK. We will hence not need to use the definition of quantum zero-knowledge explicitly in the following.

7.3 Constructing zero-knowledge arguments of knowledge

In [Unr12], the following idea was used to construct quantum proofs of knowledge: We assume a sigma-protocol with special soundness and with polynomial-size $|N_z|$. We convert it into a sigma-protocol with strict soundness as follows: When the prover sends his commitment $a \leftarrow P_1(x, w)$, he additionally sends $com(r_z)$ for all $z \in N_z$ where r_z is the response to the challenge z. When the prover receives the challenge z, he opens $com(r_z)$ instead of sending r_z. If the commitment has the "strict binding" property, the resulting sigma-protocol has strict soundness (without losing the special soundness or HVZK property).[17] Strict binding is a strengthening of perfect binding, it means that not only the message in the commitment is information-theoretically determined, but also the opening information.

Given a sigma-protocol with strict and special soundness, we can show that it is a proof of knowledge. Basically, [Unr12] runs the protocol twice (using the inverse of the unitary malicious prover to rewind) to get two responses r, r' for different challenges $z \neq z'$. The difficulty here is that measuring r can disturb the state of the malicious prover, leading to a corrupt value r'. The trick here is that due to the strict soundness, the value r is essentially uniquely determined, and therefore the measurement does not introduce too much disturbance.[18]

Unfortunately, that technique needs commitments with the strict binding property. First, it is easy to see that strict binding commitments must be longer than the messages they contain. Short strict binding commitments are not possible. Furthermore, the only known construction of strict binding commitments [Unr12] uses quantum 1-1 one-way functions. No candidates for those are known.

We show below that the same technique of committing to the responses works with collapse-binding commitments. The crucial point in the analysis from [Unr12] was that measuring the committed response does not change the state. The collapse-binding property guarantees something slightly weaker: when measuring the committed response, the state may change, but this cannot be noticed by a computationally limited adversary.

[17] This part was done only implicitly in [Unr12], in the analysis of the Hamiltonian cycle proof system.

[18] There is some disturbance due to the fact that it is not determined whether r is a valid response or an invalid one.

So with collapse-binding commitments, an analog reasoning as in [Unr12] can be used, except that we get security only against quantum-polynomial-time adversaries. I.e., we get a quantum argument of knowledge.

We will now describe this in more detail.

First, we formalize the sigma-protocol in which we commit to the responses:

Definition 40 (Sigma-protocol with committed responses) *Let (N_z, P_1, P_2, V) be a sigma-protocol with polynomially-bounded $|N_z|$. Let $(com, verify)$ be a commitment scheme (with the responses of (N_z, P_1, P_2, V) as message space). We construct a sigma-protocol (N_z, P_1', P_2', V') as follows:*

- *$P_1'(1^\eta, x, w)$ runs: $a \leftarrow P_1(1^\eta, x, w)$. For each $z \in N_z$: $r_z \leftarrow P_2(1^\eta, x, w, z)$ [19] and $(c_z, u_z) \leftarrow com(1^\eta, r_z)$. Let $a' := (a, (c_z)_{z \in N_z})$ and return a'.*
- *$P_2'(1^\eta, x, w, z)$ returns $r' := (r_z, u_z)$.*
- *$V'(1^\eta, x, a', z, r')$ with $a' = (a, (c_z)_{z \in N_z})$ and $r' = (r, u)$: Check whether $verify(1^\eta, c_z, r, u) = 1$ and $V(1^\eta, a, z, r) = 1$. Iff so, return 1.*

We show that the above construction is a quantum argument of knowledge:

Theorem 41 (Quantum argument of knowledge) *If (N_z, P_1, P_2, V) is a sigma-protocol with computational special soundness for a relation R, and $(com, verify)$ is collapse-binding, then (N_z, P_1', P_2', V') from Definition 40 is computationally quantum extractable for R with knowledge error $1/\sqrt{|N_z|}$.*

The proof of this theorem will rely on the following lemma from [Unr12]. (That lemma is the core lemma of the rewinding technique from [Unr12].)

Lemma 42 (Extraction via quantum rewinding [Unr12]) *Let C be a set with $|C| = c$. Let $(P_i)_{i \in C}$ be projectors. Let $|\Phi\rangle$ be a unit vector. Let $V := \sum_{i \in C} \frac{1}{c} \|P_i|\Phi\rangle\|^2$ and $E := \sum_{i,j \in C, i \neq j} \frac{1}{c^2} \|P_i P_j |\Phi\rangle\|^2$. Then, if $V \geq \frac{1}{\sqrt{c}}$, $E \geq V(V^2 - \frac{1}{c})$.*

Proof of Theorem 41. Recall that any sigma-protocol can be seen as an interactive proof system by Remark 1. Let (P, V) denote the interactive proof system resulting from the sigma-protocol (N_z, P_1', P_2', V'). (In particular, the verifier V sends a random $z \in N_z$, and in the end checks whether $verify(1^\eta, c_z, r, u) = 1$ and $V(1^\eta, a, z, r) = 1$.)

Let P^* denote a malicious prover, i.e., a unitary quantum-polynomial-time algorithm. Since P^* attacks a sigma-protocol, it sends two messages. We can thus assume that P^* is of the following form:

- It operates on quantum registers Z, C, R, U. Here Z contains the internal state of P^* (initialized by algorithm Z). C is the register that will contain the first message $a' = (a, (c_z)_z)$ sent by P^*. R, U contains the second message $r' = (r, u)$ sent by P^*. And C, R, U are initialized with $|0\rangle$.
- The unitary U_x describes the unitary operation of P^* on Z, C during the first invocation of P^*. U_x is parametrized by the classical input x of P^*. The message $a' = (a, (c_z)_z)$ is obtained by measuring C in the computational basis.

[19] We can run P_2 several times using the final state of P_1 because P_1 is classical.

- The unitary U_z describes the unitary operation of P^* on Z, R, U during the second invocation of P^*. U_z is parametrized by the challenge z that P^* receives. The message $r' = (r, u)$ is obtained by measuring R and U in the computational basis.

We fix some additional notation for this proof:

- V_z: The projector on R, U onto the span of all $|r, u\rangle$ with $\mathit{verify}(1^\eta, c_z, r, u) = 1$. (That is, V_z measures whether measuring R, U would yield a valid opening of c_z.)
- W_z: The projector on R onto the span of all $|r\rangle$ with $V(1^\eta, a, z, r) = 1$. (That is, W_z measures whether measuring R yields a valid response r for challenge z.)
- $P_z := U_z^\dagger W_z V_z U_z$. Since V_z and W_z are projectors and diagonal in the computational basis, they commute and their product is a projector. And since U_z is a unitary, P_z is a projector (acting on registers Z, R, U).
- $x \leftarrow \mathbf{M}(X)$ denotes that x is assigned the result of measuring the register X in the computational basis.
- $ok \leftarrow P(X)$ means that ok is assigned 1 iff measuring the register X with projector P succeeds. (With P being, e.g., one of V_z, W_z, P_z.)
- We write $U(X)$ or $U(X)$ to mean that the unitary U is applied to the register X. (With U being, e.g., one of U_x, U_z).

With that notation, we can rewrite the success probability of the malicious prover as follows:

$$\begin{aligned}
\Pr_V &:= \Pr[\mathsf{P}^*(1^\eta, x, Z), \mathsf{V}(1^\eta, x)\rangle = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta)] \\
&= \Pr[ok_c = ok_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z \xleftarrow{\$} N_z, \\
&\qquad U_z(ZRU),\ r \leftarrow \mathbf{M}(R),\ u \leftarrow \mathbf{M}(U),\ ok_c = \mathit{verify}(1^\eta, c_z, r, u),\ ok_v = V(1^\eta, a, z, r)] \\
&= \Pr[ok = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z \xleftarrow{\$} N_z,\ ok \leftarrow P_z(ZRU)].
\end{aligned}$$

We now construct the extractor $\mathsf{K}^{\mathsf{P}^*(1^\eta, x, Z)}(1^\eta, x)$ required by Definition 37. It operates on quantum registers S, C, R, U as follows:

$$\begin{aligned}
&(x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z, z' \xleftarrow{\$} N_z,\ U_z(ZRU),\ ok_c \leftarrow V_z(RU), \\
&r \leftarrow \mathbf{M}(R),\ U_z^\dagger(ZRU),\ U_{z'}(ZRU),\ r' \leftarrow \mathbf{M}(R),\ w \leftarrow E_\Sigma(1^\eta, x, a, z, r, z', r'),\ \text{return } w.
\end{aligned}$$

Here E_Σ is the extractor of the sigma-protocol (N_z, P_1, P_2, V). This extractor exists because the sigma-protocol has computational special soundness (see Definition 38). Note that K only uses black-box access to P (via the unitaries $U_x, U_z, U_{z'}$ and their inverses).

We will now bound the success probability of the extractor

$$\begin{aligned}
\Pr_E &:= \Pr[(x, w) \in R : w \leftarrow \mathsf{K}^{\mathsf{P}^*(1^\eta, x, Z)}(1^\eta, x)] \\
&= \Pr[(x, w) \in R : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z, z' \xleftarrow{\$} N_z, \\
&\qquad U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ r \leftarrow \mathbf{M}(R),\ U_z^\dagger(ZRU),\ U_{z'}(ZRU), \\
&\qquad r' \leftarrow \mathbf{M}(R),\ w \leftarrow E_\Sigma(1^\eta, x, a, z, r, z', r')] \\
&= \Pr[(x, w) \in R : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z, z' \xleftarrow{\$} N_z,
\end{aligned}$$

$$U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ r \leftarrow \mathbf{M}(R),\ ok_v \leftarrow V(1^\eta, x, a, z, r),\ U_z^\dagger(ZRU),$$
$$U_{z'}(ZRU),\ r' \leftarrow \mathbf{M}(R),\ ok'_v \leftarrow V(1^\eta, x, a, z', r'),\ w \leftarrow E_\Sigma(1^\eta, x, a, z, r, z', r')].$$

Due to the computational special soundness of (N_z, P_1, P_2, V), in the previous game, with overwhelming probability, $z \neq z'$ and $ok_v = 1$ and $ok_{v'} = 1$ implies $(x, w) \in R$. Thus there exists a negligible μ_1 such that

$$\begin{aligned}\Pr_E \geq \Pr[&z \neq z' \wedge ok_v = ok'_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z, z' \xleftarrow{\$} N_z,\\ &U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ r \leftarrow \mathbf{M}(R),\ ok_v \leftarrow V(1^\eta, x, a, z, r),\ U_z^\dagger(ZRU),\\ &U_{z'}(ZRU),\ r' \leftarrow \mathbf{M}(R),\ ok'_v \leftarrow V(1^\eta, x, a, z', r')] - \mu_1 =: \Pr'_E - \mu_1.\end{aligned}$$

Instead of computing $ok_v \leftarrow V(1^\eta, x, a, z, r)$ using the just measured r, we can instead measure whether the register R contains a value r that would make $V(1^\eta, x, a, z, r) = 1$ true. I.e., we can replace $ok_v \leftarrow V(1^\eta, x, a, z, r)$ by a measurement using the projector W_z. Since at that point, R was just measured in the computational basis, the measurement using W_z does not disturb the state of the system. Similarly, we can replace $ok'_v \leftarrow V(1^\eta, x, a, z', r')$ by a measurement using $W_{z'}$. We get:

$$\begin{aligned}\Pr'_E &= \Pr[z \neq z' \wedge ok_v = ok'_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\\ &\qquad z, z' \xleftarrow{\$} N_z,\ U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ r \leftarrow \mathbf{M}(R),\ ok_v \leftarrow W_z(R),\ U_z^\dagger(ZRU),\\ &\qquad U_{z'}(ZRU),\ r' \leftarrow \mathbf{M}(R),\ ok'_v \leftarrow W_{z'}(R)]\\ &= \Pr[z \neq z' \wedge ok_v = ok'_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\\ &\qquad z, z' \xleftarrow{\$} N_z,\ U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ r \leftarrow \mathbf{M}_{ok_c}(R),\ ok_v \leftarrow W_z(R),\ U_z^\dagger(ZRU),\\ &\qquad U_{z'}(ZRU),\ r' \leftarrow \mathbf{M}(R),\ ok'_v \leftarrow W_{z'}(R)].\end{aligned}$$

In the last probability, $r \leftarrow \mathbf{M}_{ok_c}(R)$ refers to a measurement on R that is only executed if $ok_c = 1$. (And $r := \perp$ otherwise.) The last two probabilities are equal because $\mathbf{M}(R)$ and $\mathbf{M}_{ok_c}(R)$ only differ if $ok_c = 0$, in which case "$z \neq z' \wedge ok_v = ok'_v = 1$" is false anyway.

Since V_z measures whether R, U contains $|r, u\rangle$ with $\mathit{verify}(1^\eta, c_z, r, u) = 1$, and since $(\mathit{com}, \mathit{verify})$ is collapse-binding, and since the outcome r is never used, we have that no quantum-polynomial-time adversary can distinguish between "$ok_c \leftarrow V_z(RU),\ r \leftarrow \mathbf{M}(R)$" and "$ok_c \leftarrow V_z(RU)$", except with negligible probability. (Cf. Definition 10.) Thus there is a negligible μ_2 such that

$$\begin{aligned}\Pr'_E \geq \Pr[&z \neq z' \wedge ok_v = ok'_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\\ &z, z' \xleftarrow{\$} N_z,\ U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ ok_v \leftarrow W_z(R),\ U_z^\dagger(ZRU),\\ &U_{z'}(ZRU),\ r' \leftarrow \mathbf{M}(R),\ ok'_v \leftarrow W_{z'}(R)] - \mu_2 =: \Pr''_E - \mu_2.\end{aligned}$$

Since $\mathbf{M}(R)$ and $W_{z'}(R)$ and $V_{z'}(RU)$ commute, and since adding additional/removing operations after all values z, z', ok_v, ok'_v are fixed does not change the distribution of those values, we have that "$r' \leftarrow \mathbf{M}(R), ok'_v \leftarrow W_{z'}(R)$" and "$ok'_c \leftarrow V_{z'}(RU), ok'_v \leftarrow$

$W_z(R), U^\dagger_{z'}(ZRU)$" lead to the same distribution of z, z', ok_v, ok'_v. This justifies $(*)$ in the following calculation:

$$\begin{aligned}
\Pr''_E &\overset{(*)}{=} \Pr[z \neq z' \wedge ok_v = ok'_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\\
&\qquad z, z' \overset{\$}{\leftarrow} N_z,\ U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ ok_v \leftarrow W_z(R),\ U^\dagger_z(ZRU),\\
&\qquad U_{z'}(ZRU),\ ok'_c \leftarrow V_{z'}(RU),\ ok'_v \leftarrow W_{z'}(R),\ U^\dagger_{z'}(ZRU)]\\
&\geq \Pr[z \neq z' \wedge ok_c = ok_v = 1 \wedge ok'_c = ok'_v = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\\
&\qquad (a, (c_z)_z) \leftarrow \mathbf{M}(C),\ z, z' \overset{\$}{\leftarrow} N_z,\ U_z(ZRU),\ ok_c \leftarrow V_z(RU),\ ok_v \leftarrow W_z(R),\\
&\qquad U^\dagger_z(ZRU),\ U_{z'}(ZRU),\ ok'_c \leftarrow V_{z'}(RU),\ ok'_v \leftarrow W_{z'}(R),\ U^\dagger_{z'}(ZRU)]\\
&= \Pr[z \neq z' \wedge ok = 1 \wedge ok = 1 : (x, Z) \leftarrow \mathsf{Z}(1^\eta),\ U_x(ZC),\ (a, (c_z)_z) \leftarrow \mathbf{M}(C),\\
&\qquad z, z' \overset{\$}{\leftarrow} N_z,\ ok \leftarrow P_z(ZRU),\ ok \leftarrow P_{z'}(ZRU)].
\end{aligned}$$

Let $\alpha_{a'} := \Pr[a' = (a, (c_z)_z)]$ in the previous game, and let $|\psi_{a'}\rangle$ denote the post-measurement-state of registers Z, R, U after the measurement $(a, (c_z)_z) \leftarrow \mathbf{M}(C)$. Then

$$\Pr''_E = \sum_{a'} \alpha_{a'} \underbrace{\sum_{\substack{z,z'\\ z\neq z'}} \frac{1}{|N_z|^2} \Big\| P_{z'} P_z |\psi_{a'}\rangle \Big\|^2}_{=:E_{a'}}.$$

Furthermore, note that

$$\Pr_V = \sum_{a'} \alpha_{a'} \underbrace{\sum_{z} \frac{1}{|N_z|} \Big\| P_z |\psi_{a'}\rangle \Big\|^2}_{=:V_{a'}}.$$

Lemma 42 implies that if $V_{a'} \geq 1/\sqrt{|N_z|}$, then $E_{a'} \geq V_{a'}(V_{a'}^2 - 1/|N_z|)$. Or stated differently: $E_{a'} \geq \varphi(V_{a'})$ where $\varphi(x) := 0$ for $x < 1/\sqrt{|N_z|}$ and $\varphi(x) := x(x^2 - 1/|N_z|)$ for $x \geq 1/\sqrt{|N_z|}$. Since φ is convex on $[0, 1]$, by Jensen's inequality we get $\Pr''_E \geq \varphi(\Pr_V)$. In other words $\Pr''_E \geq \Pr_V(\Pr_V^2 - 1/|N_z|)$ whenever $\Pr_V \geq 1/\sqrt{|N_z|}$. Furthermore, the inequalities derived above give $\Pr_E \geq \Pr''_E - \mu$ for $\mu := \mu_1 + \mu_2$. And μ is negligible. It follows that:

$$\Pr_V \geq \frac{1}{\sqrt{N_z}} \qquad \Longrightarrow \qquad \Pr_E \geq \Pr_V\Big(\Pr_V^2 - \frac{1}{|N_z|}\Big) - \mu \geq \Big(\Pr_V - \frac{1}{\sqrt{|N_z|}}\Big)^3 - \mu.$$

Thus (P, V) is quantum-computationally extractable for R with knowledge error $\kappa := 1/\sqrt{|N_z|}$. □

Finally, we show that our protocol is still HVZK/SHVZK. From this we conclude below (Corollary 44) that our protocol is quantum zero-knowledge.

Lemma 43 *If $|N_z|$ is polynomially-bounded, and (N_z, P_1, P_2, V) is HVZK and $(com, verify)$ is computationally hiding, and com is a polynomial-time algorithm, then (N_z, P'_1, P'_2, V') is HVZK.*

If $|N_z|$ is polynomially-bounded, and (N_z, P_1, P_2, V) is SHVZK and $(com, verify)$ is statistically hiding, and com is a polynomial-time algorithm, then (N_z, P_1', P_2', V') is SHVZK.

Proof. We first prove the computational case of the lemma. Assume that $|N_z|$ is polynomially-bounded, and (N_z, P_1, P_2, V) is HVZK and $(com, verify)$ is computationally hiding.

We need to show that (N_z, P_1', P_2', V') is HVZK. By definition of HVZK, and by construction of (N_z, P_1', P_2', V'), that means that for any quantum-polynomial-time A,

$$\begin{aligned}\Pr_1 := \Pr[b=1 : {} & a \leftarrow P_1(1^\eta, x, w),\ \text{for each } z\colon r_z \leftarrow P_2(1^\eta, x, w, z),\\ & \text{for each } z\colon (c_z, u_z) \leftarrow com(1^\eta, r_z),\ z \xleftarrow{\$} N_z,\\ & b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)]\\ \approx \Pr[b=1 : {} & (a, (c_z)_z, z, r_z, u_z) \leftarrow S_\Sigma'(1^\eta, x),\\ & b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)] =: \Pr_{sim} \end{aligned} \tag{29}$$

Here S_Σ' is a quantum-polynomial-time simulator that we will construct below. And $\approx$ means that the difference between the probabilities is negligible.

We then calculate:

$$\begin{aligned}\Pr_1 = \Pr[b=1 : {} & a \leftarrow P_1(1^\eta, x, w),\ z \xleftarrow{\$} N_z,\ r_z \leftarrow P_2(1^\eta, x, w, z),\ (r_z, u_z) \leftarrow com(1^\eta, r_z),\\ & \text{for each } z' \neq z\colon r_{z'} \leftarrow P_2(1^\eta, x, w, z),\\ & \text{for each } z' \neq z\colon (c_{z'}, u_{z'}) \leftarrow com(1^\eta, r_{z'}),\ b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)]\\ \overset{(*)}{\approx} \Pr[b=1 : {} & a \leftarrow P_1(1^\eta, x, w),\ z \xleftarrow{\$} N_z,\ r_z \leftarrow P_2(1^\eta, x, w, z),\ (r_z, u_z) \leftarrow com(1^\eta, r_z),\\ & \text{for each } z' \neq z\colon r_{z'} \leftarrow P_2(1^\eta, x, w, z),\\ & \text{for each } z' \neq z\colon (c_{z'}, u_{z'}) \leftarrow com(1^\eta, 0),\ b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)]\\ = \Pr[b=1 : {} & a \leftarrow P_1(1^\eta, x, w),\ z \xleftarrow{\$} N_z,\ r_z \leftarrow P_2(1^\eta, x, w, z),\ (r_z, u_z) \leftarrow com(1^\eta, r_z),\\ & \text{for each } z' \neq z\colon (c_{z'}, u_{z'}) \leftarrow com(1^\eta, 0),\ b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)]\\ \overset{(**)}{\approx} \Pr[b=1 : {} & (a, z, r_z) \leftarrow S_\Sigma(1^\eta, x),\ (r_z, u_z) \leftarrow com(1^\eta, r_z),\\ & \text{for each } z' \neq z\colon (c_{z'}, u_{z'}) \leftarrow com(1^\eta, 0),\ b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)] =: \Pr_2 \end{aligned}$$

Here $(*)$ uses that $(com, verify)$ is computationally hiding and A is quantum-polynomial-time. $com(1^\eta, 0)$ refers to a commitment to some fixed message 0 in the message space of com. And $(**)$ follows from the HVZK property of (N_z, P_1, P_2, V) for suitable quantum-polynomial-time S_Σ.

Let $S_\Sigma'(1^\eta, x)$ perform the following steps:

$$\begin{aligned}& (a, z, r_z) \leftarrow S_\Sigma(1^\eta, x),\ (r_z, u_z) \leftarrow com(1^\eta, r_z),\\ & \text{for each } z' \neq z\colon (c_{z'}, u_{z'}) \leftarrow com(1^\eta, 0),\ \text{return } (a, (c_z), z, r_z, u_z).\end{aligned}$$

Then S_Σ' is quantum-polynomial-time, and

$$\Pr_2 = \Pr[b=1 : (a, (c_z)_z, z, r_z, u_z) \leftarrow S_\Sigma'(1^\eta, x),\ b' \leftarrow A(1^\eta, |\Psi\rangle, a, (c_z)_z, z, r_z, u_z)] = \Pr_{sim}\,.$$

Hence $\Pr_1 \approx \Pr_{sim}$, so (29) holds, and it follows that (N_z, P_1', P_2', V') is HVZK. This shows the lemma in the computational case.

The statistical case of the lemma is shown fully analogously, except that we do not assume A to be quantum-polynomial-time (and thus have to use the statistical hiding property of $(com, verify)$ and the SHVZK property of (N_z, P_1, P_2, V)). □

Corollary 44 (Zero-knowledge) *If $|N_z|$ is polynomially-bounded, and (N_z, P_1, P_2, V) is HVZK and $(com, verify)$ is computationally hiding, and com is a polynomial-time algorithm, then (N_z, P_1', P_2', V') is computational zero-knowledge.*

If $|N_z|$ is polynomially-bounded, and (N_z, P_1, P_2, V) is SHVZK and $(com, verify)$ is statistically hiding, and com is a polynomial-time algorithm, then (N_z, P_1', P_2', V') is statistical zero-knowledge.

Proof. Immediate from Lemma 43 and Theorem 39. □

8 Interactive quantum commitments

The definition of the collapse-binding property (Definition 10) was formulated specifically for non-interactive commitments where only classical messages are exchanged and where the verification in the opening phase is deterministic.

For completeness, we show here how the definition can be generalized to interactive commitments that may send quantum states and have a quantum verification algorithm. Note that we still consider the case where the message that we commit to is classical. Also, for technical reasons, we consider only commitments where the opening phase consists of a single quantum message.

We stress that, in contrast to Definition 10, we have not investigated this definition further. For example, we do not know whether commitments according to Definition 45 below are useful for constructing zero-knowledge arguments. We mainly state this definition for reference and leave it to future research to see how well the definition behaves.

We model an interactive commitment using two interactive algorithms SND (sender) and RCP (recipient) for the commit phase, and a quantum algorithm VER for the opening phase. $(S, U, R) \leftarrow \langle \mathsf{SND}(1^\eta, m), \mathsf{RCP}(1^\eta) \rangle$ denotes an execution of the interaction between SND and RCP where SND is committing to the message m (and with security parameter η). Here the quantum registers S, U, R contain the state of SND, the opening information (which consists of a single message), and the state of RCP, respectively. The algorithm VER takes the security parameter and quantum registers M, U, R as input and outputs a single bit, indicating whether the opening phase succeeded.

Definition 45 (Collapse-binding – generalized) *Let Z be an auxiliary quantum register. Let P^η_{VER} be a projector on quantum registers M, U, R, Z (parametric in the*

security parameter η), such that for any η and any quantum state $|\Psi\rangle$ on M,U,R, $\Pr[\mathsf{VER}(1^\eta,M,U,R)=1 : MUR \leftarrow |\Psi\rangle] = \left|P^\eta_{\mathsf{VER}}(|\Psi\rangle\otimes|0\rangle)\right|^2$.[20]

For an interactive algorithm A and a non-interactive algorithm B, consider the following games:

$$\begin{aligned}
\mathsf{Game}_1: \quad & (S,M,U,R)\leftarrow\langle A(1^\eta),\mathsf{RCP}(1^\eta)\rangle, Z\leftarrow|0\rangle, ok\leftarrow P^\eta_{\mathsf{VER}}(M,U,R,Z),\\
& \qquad\qquad m\leftarrow M_{ok}(M), b\leftarrow B(1^\eta,S,M,U)\\
\mathsf{Game}_2: \quad & (S,M,U,R)\leftarrow\langle A(1^\eta),\mathsf{RCP}(1^\eta)\rangle, Z\leftarrow|0\rangle, ok\leftarrow P^\eta_{\mathsf{VER}}(M,U,R,Z),\\
& \qquad\qquad b\leftarrow B(1^\eta,S,M,U)
\end{aligned}$$

Here $(S,M,U,R)\leftarrow\langle A(1^\eta),\mathsf{RCP}(1^\eta)\rangle$ denotes an interaction between A and the honest recipient RCP*. The quantum registers S,M,U are output by A, the register R contains the final state of* RCP*. $Z\leftarrow|0\rangle$ means the quantum register Z is initialized with $|0\rangle$. $ok\leftarrow P_{\mathsf{VER}}(M,U,R,Z)$ means that ok is the output of measuring the joint register M,U,R,Z with projector P_{VER}. M_{ok} is as in Definition 10.*

We say $(\mathsf{SND},\mathsf{RCP},\mathsf{VER})$ *is* collapse-binding *relative to P_{VER} iff for any quantum-polynomial-time interactive algorithm A and any quantum-polynomial-time algorithm B, the difference $\left|\Pr[b=1:\mathsf{Game}_1]-\Pr[b=1:\mathsf{Game}_2]\right|$ is negligible in η.*

We stress that the choice of the purification P^η_{VER} of VER is not irrelevant. In general, different purifications P^η_{VER} may have different post-measurement state even if they realize the same algorithm VER [Unr14b]. Thus Definition 45 could be satisfied with one P^η_{VER} and not satisfied with another. We conjecture that in most cases where the definition is used, one will simply need the existence of some quantum-polynomial-time P^η_{VER}.

Notice that for a non-interactive commitment scheme with classical messages, Definition 45 coincides with Definition 10: In the non-interactive case, RCP simply stores the classical message c it receives, hence $(S,M,U,R)\leftarrow\langle A,\mathsf{RCP}\rangle$ becomes $(S,M,U,c)\leftarrow A(1^\eta)$. And the projector P_{VER} from Definition 45 can be chosen as $\sum_c V_c\otimes|c\rangle\langle c|$ where V_c is the projector from Definition 10.

9 Open problems

We list some questions for future research:

- We have constructed quantum arguments of knowledge from sigma-protocols by using collapse-binding commitments. However, our construction requires the challenge space N_z of the sigma-protocol to be of polynomially-bounded size. As a consequence, the resulting argument of knowledge will have a noticeable knowledge error; for a negligible knowledge error we need to use sequential repetition, resulting in a proof system with non-constant round complexity. Are there general

[20] In other words, P_{VER} performs the same measurement as VER does using an ancilla system Z. Given a quantum-polynomial-time VER, we can always construct a polynomial-size quantum circuit implementing P: Let U be the purification of VER using ancillae Z. Let P_1 be the projector that measures whether VER outputs 1. Then $P_{\mathsf{VER}}:=U^\dagger P_1 U$ is one possible choice for P_{VER}.

constructions of arguments of knowledge from sigma-protocols that do not require the challenge space to be polynomially-bounded?
- Can we use collapse-binding commitments to construct a quantum OT protocol? For example, using the construction from [BBCS91] or a variation thereof?
- How are the various definitions of computationally binding commitments related? That is, which implications and separations exist between sum-binding, CDMS-binding, collapse-binding, and UC-secure commitments?

Acknowledgements. We thank Ansis Rosmanis for discussions on insecure commitments based on collision-resistant hash functions. This research by the European Social Fund's Doctoral Studies and Internationalisation Programme DoRa, by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS, by European Social Fund through the Estonian Doctoral School in Information and Communication Technology, and by the Estonian ICT program 2011-2015 (3.2.1201.13-0022).

References

[ARU14] Andris Ambainis, Ansis Rosmanis, and Dominique Unruh. Quantum attacks on classical proof systems (the hardness of quantum rewinding). In *FOCS 2014*, pages 474–483. IEEE, 2014. Preprint on IACR ePrint 2014/296.

[BBCS91] Charles H. Bennett, Gilles Brassard, Claude Crépeau, and Marie-Hélène Skubiszewska. Practical quantum oblivious transfer. In *Crypto '91*, volume 576 of *LNCS*, pages 351–366. Springer, 1991.

[BCJL93] G. Brassard, C. Crépeau, R. Jozsa, and D. Langlois. A quantum bit commitment scheme provably unbreakable by both parties. In *FOCS '93*, pages 362–371, Los Alamitos, CA, USA, 1993. IEEE.

[BDF+11] Dan Boneh, Özgür Dagdelen, Marc Fischlin, Anja Lehmann, Christian Schaffner, and Mark Zhandry. Random oracles in a quantum world. In *Asiacrypt 2011*, pages 41–69, Berlin, Heidelberg, 2011. Springer-Verlag.

[CDMS04] Claude Crépeau, Paul Dumais, Dominic Mayers, and Louis Salvail. Computational collapse of quantum state with application to oblivious transfer. In *TCC 2004*, volume 2951 of *LNCS*, pages 374–393. Springer, 2004.

[CDS94] Ronald Cramer, Ivan Damgård, and Berry Schoenmakers. Proofs of partial knowledge and simplified design of witness hiding protocols. In Yvo Desmedt, editor, *Crypto 94*, volume 839 of *Lecture Notes in Computer Science*, pages 174–187. Springer, 1994.

[CLS01] Claude Crépeau, Frédéric Légaré, and Louis Salvail. How to convert the flavor of a quantum bit commitment. In *Eurocrypt 2001*, volume 2045 of *LNCS*, pages 60–77. Springer, 2001.

[CSST11] Claude Crépeau, Louis Salvail, Jean-Raymond Simard, and Alain Tapp. Two provers in isolation. In Dong Hong Lee and Xiaoyun Wang, editors, *Asiacrypt 2011*, volume 7072 of *LNCS*, pages 407–430. Springer, 2011.

[DFL+09] Ivan Damgård, Serge Fehr, Carolin Lunemann, Louis Salvail, and Christian Schaffner. Improving the security of quantum protocols via commit-and-open. In *Crypto 2009*, volume 5677 of *LNCS*, pages 408–427. Springer, 2009.

[DL09] Ivan Damgård and Carolin Lunemann. Quantum-secure coin-flipping and applications. In *Asiacrypt 2009*, volume 5912, pages 52–69. Springer, 2009.

[DMS00] Paul Dumais, Dominic Mayers, and Louis Salvail. Perfectly concealing quantum bit commitment from any quantum one-way permutation. In *Eurocrypt 2000*, volume 1807 of *LNCS*, pages 300–315. Springer, 2000.

[HM96] Shai Halevi and Silvio Micali. Practical and provably-secure commitment schemes from collision-free hashing. In Neal Koblitz, editor, *Crypto '96*, volume 1109 of *LNCS*, pages 201–215. Springer, 1996.

[May97] Dominic Mayers. Unconditionally Secure Quantum Bit Commitment is Impossible. *Physical Review Letters*, 78(17):3414–3417, 1997. Online available at `http://arxiv.org/abs/quant-ph/9605044`.

[NC10] M. Nielsen and I. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, 10th anniversary edition, 2010.

[NIS14] NIST. SHA-3 standard: Permutation-based hash and extendable-output functions. Draft FIPS 202, 2014. Available at `http://csrc.nist.gov/publications/drafts/fips-202/fips_202_draft.pdf`.

[Pas04] Rafael Pass. Alternative variants of zero-knowledge proofs. Licentiate thesis, KTH Numerical Analysis and Computer Science, Stockholm, 2004. Online available at `http://www.cs.cornell.edu/~rafael/papers/raf-lic.pdf`.

[Unr10] Dominique Unruh. Universally composable quantum multi-party computation. In *Eurocrypt 2010*, volume 6110 of *LNCS*, pages 486–505. Springer, May 2010.

[Unr12] Dominique Unruh. Quantum proofs of knowledge. In *Eurocrypt 2012*, volume 7237 of *LNCS*, pages 135–152. Springer, April 2012. Full version is IACR ePrint 2010/212.

[Unr14a] Dominique Unruh. Non-interactive zero-knowledge proofs in the quantum random oracle model. IACR ePrint 2014/587, 2014.

[Unr14b] Dominique Unruh. Uniqueness of representing POVM using projective measurement. Physics Stack Exchange answer, 2014. `http://physics.stackexchange.com/q/144037` (version: 2014-10-31).

[Unr15] Dominique Unruh. Quantum proofs of knowledge. IACR ePrint 2010/212/20150211:174234, February 2015. Updated full version of [Unr12].

[Wat09] John Watrous. Zero-knowledge against quantum attacks. *SIAM J. Comput.*, 39(1):25–58, 2009. Online available at `https://cs.uwaterloo.ca/~watrous/Papers/ZeroKnowledgeAgainstQuantum.pdf`.

[Zha12] Mark Zhandry. How to construct quantum random functions. In *FOCS 2013*, pages 679–687, Los Alamitos, CA, USA, 2012. IEEE Computer Society. Online version is IACR ePrint 2012/182.

[Zha13] Mark Zhandry. A note on the quantum collision and set equality problems. arXiv:1312.1027v3 [cs.CC], December 2013.