Quantum Information and Computation, Vol. 16, No. 11&12 (2016) 1048–1070
© Rinton Press

AN ADAPTIVE ATTACK ON WIESNER'S QUANTUM MONEY

DANIEL NAGAJ
Institute of Physics, Slovak Academy of Sciences
Dúbravská cesta 9, 84215 Bratislava, Slovakia

OR SATTATH
Computer Science Division, UC Berkeley

AHARON BRODUTCH
Institute for Quantum Computing and Department of Physics and Astronomy, University of Waterloo

DOMINIQUE UNRUH
University of Tartu, Estonia

Received May 28, 2015
Revised May 24, 2016

Unlike classical money, which is hard to forge for practical reasons (e.g. producing paper with a certain property), quantum money is attractive because its security might be based on the no-cloning theorem. The first quantum money scheme was introduced by Wiesner circa 1970. Although more sophisticated quantum money schemes were proposed, Wiesner's scheme remained appealing because it is both conceptually clean as well as relatively easy to implement.

We show efficient adaptive attacks on Wiesner's quantum money scheme [1] (and its variant by Bennett et al. [2]), when valid money is accepted and passed on, while invalid money is destroyed. We propose two attacks, the first is inspired by the Elitzur-Vaidman bomb testing problem [3, 4], while the second is based on the idea of *protective measurements* [5]. It allows us to break Wiesner's scheme with 4 possible states per qubit, and generalizations which use more than 4 states per qubit. The attack shows that Wiesner's scheme can only be safe if the bank replaces valid notes after validation.

Keywords: quantum money, adaptive attack, quantum Zeno effect, protective measurement

Communicated by: R Cleve & R de Wolf

1 Introduction

One of the main requirements for any medium of money is that it should not be easily copied. For this very reason, it is appealing to construct *quantum money*: its security would follow from the laws of quantum mechanics, or more specifically, the no-cloning theorem [6]. Indeed, quantum money was one of the earliest quantum information protocols, introduced by Stephen Wiesner circa 1970, although it took some time to be published [1].

Wiesner's quantum money scheme uses only single-qubit memory and single-qubit measurements, as follows: A bank creates a note of size n with a public serial number s, and for each serial number a random (classical) private key $k^{(s)} \in \{0, 1, +, -\}^n$. The corresponding banknote contains a *quantum money state* $|\$_s\rangle = |k_1^{(s)}\rangle \otimes |k_2^{(s)}\rangle \otimes \ldots \otimes |k_n^{(s)}\rangle$, where $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$. The serial number together with the quantum money state, i.e. the pair $(s, |\$_s\rangle)$, form *legitimate* quantum money.

In order to validate her money (and pay with it), Alice sends the (potentially forged) banknote $(s, |\psi\rangle)$ to the bank. The bank measures each of the qubits of $|\psi\rangle$ in its respective basis; the i^{th} qubit is measured in the basis $\{|0\rangle, |1\rangle\}$ if $k_i^{(s)} \in \{0, 1\}$, and in the $\{|+\rangle, |-\rangle\}$ basis otherwise. The money is declared valid if and only if all the measurement outcomes agree with the measurements on the legal state $|\$_s\rangle$.

One downside of Wiesner's original scheme is that the bank must keep a database containing the secret key for every serial number. In a follow-up paper, Bennett, Brassard, Breidbart and Wiesner used a fixed pseudo-random function for choosing the secret keys for all the serial numbers, which implies that the memory required by the bank does not grow as a function of the number of legitimate serial numbers [2]. All our results apply both to Wiesner's and to the Bennett et al. scheme.

There are two specifications which are important for our work: we have to determine whether after a successful validity test, the money state is returned to Alice (or passed to Bob who does business with Alice), or replaced with a new quantum money state and a new serial number. Next, after a failed validity test, is the post-measurement state (a bad bill) returned to Alice for inspection? These distinctions are crucial for the scheme's security. We define *strict testing* to be the variant of Wiesner's scheme in which the valid money state is returned to the owner after a successful test, and the post measurement state of a failed test is destroyed.

Wiesner proposed his quantum money scheme more than 40 years ago, and it was believed to be secure although a complete security proof was never published. However, Lutomirski [7] and Aaronson [8] independently observed that the scheme is insecure if the bank returns valid notes as well as the post measurement state after detecting an invalid note. Farhi et al. showed that "single copy tomography" [9] can be performed in a much more general setting. When we own a single copy of an unknown state $|\psi\rangle$ and have access to a projective measurement (validation test) $\{|\psi\rangle\langle\psi|, \mathbb{I} - |\psi\rangle\langle\psi|\}$ provided as a black box, we can efficiently estimate the reduced local density matrices of the state $|\psi\rangle$. This approach, based on Jordan's lemma, pointed out the security threat posed by having access to a validation procedure (in particular, the post-measurement state). Lutomirski's conclusion was that as long as the bank does not return the post-measurement state of an invalid note, the scheme should remain safe.

On the other hand, Molina, Watrous and Vidick proved that Wiesner's scheme is secure against *simple counterfeiting attacks* [10]. In this model, an attacker is given a single copy of an authenticated state, and attempts to create two banknotes with the same serial number which, independently, pass the bank's validity test. During the counterfeiting process, the attacker does not have access to the validity test. They showed that the success probability of the optimal attack on Wiesner's quantum money scheme is $\left(\frac{3}{4}\right)^n$.

Pastawski et al. [11, Theorem 5] proved security under a more general setting. In their setting a user is given one bank note, and creates polynomially many faked banknotes. The bank validates all the banknotes. They show that the probability that at least two banknotes would pass the validation is exponentially small. However, in a general attack (as in our case), the attacker could make some use of the quantum validation procedure, even if the bank strictly discourages failed tests and does not return bad banknotes.

Main results. We show that in a *strict testing* variant of Wiesner's scheme (that is, if only valid money is returned to the owner), given a single valid quantum money state $(s, |\$_s\rangle)$, a counterfeiter can efficiently create as many copies of $|\$_s\rangle$ as he wishes (hence, the scheme is insecure). He can rely on the quantum Zeno effect for protection – if he disturbs the quantum money state only slightly, the bill is likely to be projected back to the original state after a test. Interestingly, this allows a counterfeiter to distinguish the four different qubit states with an arbitrarily small probability of being caught.

The BT (bomb-testing) attack: The simplest attack is based on the Zeno assisted Elitzur-Vaidman bomb test [4]. The *bomb testing* (BT) attack lets us tell whether the first qubit of our quantum money is in the state $|+\rangle$ or not. By repeating this test for each of the money qubits (and for each of the four possible states), we can identify the quantum money state. We use an ancillary *probe* qubit, initialized to the $|0\rangle$ state. We repeat the following steps $N = \frac{\pi}{2\delta}$ times, as depicted in Fig. 3 on p. 1055:

1. Rotate the probe by a small angle δ.

2. Apply a C-NOT from the probe qubit to the money qubit.

3. Send the quantum money to the bank for validation (and get it back when verified).

If the money qubit is in the $|+\rangle$ state, it stays invariant under the NOT operation, and therefore also by the C-NOT operation controlled by a probe. Hence, at the end of the procedure, the probe qubit will be in the state $|1\rangle$. If the quantum money state is in either the $|0\rangle$ or $|1\rangle$ state, the probe will be in the $|0\rangle$ state, using the same analysis of the Elitzur-Vaidman bomb tester. In these two cases the maximal rotation induced on the money state at any one time is at most δ and the bank's probability of detecting a counterfeiter is at most δ^2; hence, the overall probability of detection by the bank is $O(\delta)$. The last case, $|\$_i\rangle = |-\rangle$, is somewhat different: after the first iteration, the probe has angle $-\delta$. At the end of the second iteration, the probe returns to state $|0\rangle$, etc. Therefore, at the end of the procedure, the probe is in the state $|0\rangle$ with certainty (as long as N is even), while the money state is left invariant.

One might hope that a simple generalization of Wiesner's strict testing scheme using r states as a basis instead of 4 states (and a d dimensional qudit instead of a qubit) will be able to hold off our attack. However, we show that this is not sufficient. More precisely, let the generalized money scheme use n random states from the set $\{|\beta_1\rangle, \dots, |\beta_r\rangle\}$, where $|\beta_i\rangle \in \mathcal{C}^d$, and let

$$\theta_{min} = \min_{1 \le i \neq j \le r} \arccos |\langle \beta_i | \beta_j \rangle|.$$

In Appendix A.1 we show a simple generalization of the BT attack that succeeds with probability $1 - f$, and uses $O\left(nr^2\theta_{min}^{-2} f^{-1}\right)$ validations.

The PM (protective measurement) attack: What if there are infinitely many states per qubit (which implies an infinite number of verifications, since $\theta_{min} = 0$)? In this case, the previous attack fails (see Appendix A.2), and we present an alternative tomographic attack based on protective measurements [5]. The *protective measurement* (PM) attack allows us to estimate the expectation value $\langle A \rangle = \langle \psi | A | \psi \rangle$ of an operator A in the state $|\psi\rangle$, without disturbing the state much, by preparing a probe in the initial state $|0\rangle$, choosing $\delta = \frac{c}{N}$ for some constant c and repeating the following procedure N times

1. Weakly couple the probe and the system.

2. Send the state to the bank for validation.

We aim for the following to happen:

$$\begin{aligned}|0\rangle|\psi\rangle &\xrightarrow{e^{-i\delta(\sigma_x\otimes A)}} \approx |0\rangle|\psi\rangle - i\delta|1\rangle A|\psi\rangle \\ &\xrightarrow{\text{bank measures }\{|\psi\rangle\langle\psi|,\mathbb{I}-|\psi\rangle\langle\psi|\}} \approx \left(e^{-i\delta\langle A\rangle\sigma_x}|0\rangle\right)\otimes|\psi\rangle \\ &\xrightarrow{\text{repeat } N \text{ times}} \approx \left(e^{-iN\delta\langle A\rangle\sigma_x}|0\rangle\right)\otimes|\psi\rangle.\end{aligned}$$

By measuring the probe and using standard parameter estimation techniques, we can approximate $\langle A\rangle$ and thus $|\psi\rangle$ to the desired precision (depending on n, the number of qubits in the unknown state $|\psi\rangle$).

In Def. 2 we present a more general question regarding the cost, accuracy and confidence of *protective tomography*, i.e. where we have a single copy of a state $|\psi\rangle$ and access to a validation procedure $\{|\psi\rangle\langle\psi|, 1-|\psi\rangle\langle\psi|\}$ that destroys the state when the validation fails. As far as we know, our analysis in Sec. 4, gives the first quantitative answer to this question.

Using this tomography approach, without any assumptions on the states, for any constant ϵ, using $O\left(n^5\ln^2(n)\right)$ bank validations, we get caught with probability $1-\epsilon$; and conditioned on the event that we are not caught, we find a (classical description of a) quantum state ρ such that $F(\rho, |\$\rangle\langle\$|) \geq 1-\epsilon$, with probability at least $1-\epsilon$. Note that $1-F^2(\rho, |\$\rangle\langle\$|)$ is precisely the probability of getting caught when providing a fake state ρ instead of the legitimate money $|\$\rangle$.

Discussion and applications. Our attack applies in the strict-testing regime, where good banknotes are returned or passed on, while failed tests result in confiscation of the banknote (or us being sent to jail). However, the attack does not work if after a valid test, a new quantum bill with a new serial number are returned to the owner. Does this affect the advantages of Wiesner's scheme? In order to answer this question, we first need to understand these advantages.

The two main advantages of Wiesner's scheme over other money schemes are the following (see a more detailed analysis in Ref. [12]):

- The data needed for validation is static and classical. Therefore, after the money has been issued, many bank branches can validate the money without any need for communication.
- The hardware requirements (single-qubit memory and single-qubit measurements) are less demanding than other schemes such as Farhi et al. [17] and Aaronson and Christiano [13] which require a fully scalable quantum computer and quantum memory which can hold a large number of entangled qubits.

These advantages remain when the quantum money is replaced with a new one after each successful validity test. In some sense, such quantum money resembles one time tokens, which are destroyed after each usage.

Where do our results lead? First, we believe that the greatest potential of this work is in the context of weak measurements. The framework of weak measurement has been proven an important concept in numerous cases [14, 15] including protective measurements and precision metrology. Our "bomb" attack shares a lot of the properties of weak measurements, but not all. We believe that further investigation of these different approaches in a broad perspective will yield practical applications for weak measurement methods. Second, our adaptive attack is interesting on its own as a pedagogical device for exhibiting the counter-intuitive properties of quantum mechanics, and more specifically, for weak

& protective measurements. Third, it raises an important warning about quantum money constructions – we need to be cautious reusing valid bills (even though false bills are destroyed). Fourth, one can think about a failed test simply as an undesirable scenario, and try to apply the technique even if no actual strict-validation box exists. This could find use in tomography and state discrimination using several copies of unknown states. Fifth, after finishing this work we learned that our preprint and the strict-testing model has inspired interesting questions in query complexity [16].

Wiesner's money in a noisy environment. In this paper, we have focused on Wiesner's money in a noiseless environment. That is, the bank rejects the money if even a single qubit is measured incorrectly. In a more realistic setting, we have to deal with noise, and the bank would want to tolerate a limited amount of errors in the quantum state [11], say 10 %. (All concrete numbers in this section are examples.) Additionally, the natural design choice would be that the bank repairs the incorrect qubits so that the quantum state does not deteriorate more and more. But then, a simple attack exists [7, 8]: The owner replaces the first qubit of the money state by $|0\rangle$, keeping the original qubit. Then he submits the money state for validation. The bank repairs the replaced qubit. Now the user has a copy of the first qubit, and the original money state. By repeating this process, the user may get copies of all qubits, thus getting two copies of the money state. In light of this attack, it seems obvious that the bank must issue a fresh, independent money state (with a new serial number) even in case of a successful validation. This is, of course, the same recommendation as we are making in this paper: "Issue a new money state after each validation."

We are aware of the following possible criticism of our result, saying we are attacking an obvious loophole: "Since in the case of noise, there are obvious reasons why the bank should re-issue the money, this attack is not relevant since the bank would prepare new money states anyway." However, this is not the case, our attack is relevant even in a setting with noise. The bank surely should not simply repair the incorrectly measured, noisy qubits. However, it is not strictly necessary to always re-issue a new state either. Instead, the bank could do the following:

- If at most 5% of the qubits were measured incorrectly, the bank hands back the state. (No repairs.)
- If 5%–10% of the qubits were measured incorrectly, the bank re-issues a new money state.
- If more than 10% of the qubits were measured incorrectly, the bank informs the police of a forgery attempt.

Why would this be a reasonable thing to do? First, this approach takes care of the degradation of the quantum money by re-issuing the money when the error exceeds 5%. However, it also saves resources: the money is not re-issued upon each validation. Hence, fewer serial numbers need to be allocated, reducing the storage needed by the bank.[a] Finally, it is no longer possible for the attacker to just replace qubits to get a second copy of the money state. In fact, none of the prior attacks we know of applies to this scheme. The attack proposed in this paper, however, applies to this scheme without any modification. Thus our attack does indeed constitute a new attack vector, and needs to be taken into

[a] This becomes particularly relevant when we do not use a pseudorandom function to map serial numbers to money states [2] (because we want information-theoretical security) and we distribute the whole database with the serial numbers up front to the branches of the bank (because we want to allow the different branches to validate money independently without communication).

account in the design of quantum money protocols (and possibly other quantum protocols). We stress that the above protocol example is just that – an illustrative example. Its purpose is to illustrate that there can be many settings in which our attack applies, the above is just the simplest one we could come up with.

Structure of the paper. We introduce the quantum (Zeno effect based) Elitzur-Vaidman bomb quality tester in Section 2 and use it as a tool to enlighten the rest of the analysis in Section 3, where we show our adaptive BT (bomb-testing) attack on Wiesner's scheme. We present and analyze the PM (protective measurement) attack in Section 4 which lets us perform single-copy tomography with the help of strict testing, for general states, or for a generalization of Wiesner's scheme which was mentioned before. Finally in Section 5, we briefly compare the two attacks. In Appendix A we extend the BT attack to deal with simple generalizations of Wiesner's scheme. However, in Appendix A.2 we show that unlike the PM attack, the BT attack does not always work – and it is instructive to learn why.

2 Elitzur-Vaidman's bomb quality tester

We usually think that in order to measure a quantum system we must interact with it. However, sometimes there is a possibility of an *interaction-free* measurement detecting some property of a system without disturbing it. The prime example of this is Elitzur-Vaidman's bomb tester [3], a probabilistic test that can certify a property of an object (a working trigger) by detecting a photon that never "interacted" with the object. Using the quantum Zeno effect, this test has been improved [4] so that one can be sure about the system's property, while the probability of disturbing the object goes to zero (we do not want any explosions).

We now present a quantum information variant of this approach, which we will use in the next section as an adaptive attack against Wiesner's quantum money scheme.

The goal is to test whether a "quantum bomb" is a dud or an actual bomb.[b] If we have a dud, it remains in the $|0\rangle$ state when we interact with it. On the other hand, we can flip the state of a live bomb to $|1\rangle$, which makes it explode. The trick is how not to trigger a live bomb. The safe bomb quality testing procedure is illustrated in Figure 1. We pick a large number N, choose a small angle δ and label R_δ a counterclockwise rotation by this angle:

$$\delta = \frac{\pi}{2N}, \qquad R_\delta = \begin{bmatrix} \cos\delta & -\sin\delta \\ \sin\delta & \cos\delta \end{bmatrix}. \tag{1}$$

We will use two registers (the probe and the system), and apply the *controlled interaction*

$$C_P = |0\rangle\langle 0| \otimes \mathbb{I} + |1\rangle\langle 1| \otimes P. \tag{2}$$

When there is an active bomb, this operation is a controlled-X (a CNOT), while for a dud $P = \mathbb{I}$, so the operation C_P is just an identity. The testing procedure starts with the first register in the state $|0\rangle$ and applies the following steps N times

1. Prepare the second (system) register in $|0\rangle$.

[b] Note that we differ somewhat from the original treatment of the bomb-tester, as our trigger is a controlled quantum gate (I or X in Figure 1) instead of a transparent object for a dud and a photon detector for a working bomb. However, mathematically, our presentation is equivalent to the original one, and it turns out to be more convenient for understanding Section 3.

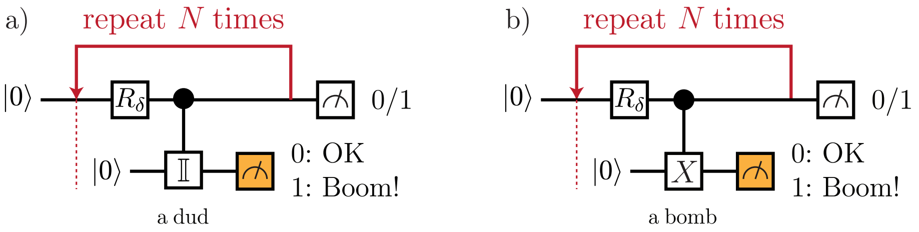

Fig. 1. A quality-testing procedure for bombs: run N rounds and end with a measurement of the first register. a) A dud can't explode, and the first register slowly rotates from $|0\rangle$ to $|1\rangle$. b) With a live bomb, we can really trigger the bomb by flipping the second register to $|1\rangle$. This does not happen often as δ is small, and we are much more likely to measure $|0\rangle$ on the second register. The first register is then projected back to $|0\rangle$.

2. Rotate the first (probe) register using R_δ.
3. Apply the controlled interaction C_P.
4. Measure the second (system) register. If we get $|1\rangle$, we fail. If we succeed, we reuse the first register and go back to step 1.

Once we are done with N repetitions of these steps, we measure the first register in the computational basis.

First, let us look at the case with a dud (Figure 1a). The controlled interaction does nothing, and the state after the first round is $(\cos\delta|0\rangle + \sin\delta|1\rangle)\,|0\rangle$. We thus measure $|0\rangle$ in the second register, and continue. In the following rounds, the first register is repeatedly rotated by δ, finally ending up in the state $|1\rangle$. Thus, without a bomb, we finally end up measuring $|1\rangle$ in the first register.

Second, what if there is an actual bomb? Now $P = X$, and the controlled interaction is the CNOT operation. If we actually tested the bomb (flipped the second register), we would die (measure $|1\rangle$ in the second register). However, we choose to test the bomb in superposition, with a small angle δ of the control register state. In the first round, the state before the measurement is $\cos\delta|0\rangle|0\rangle + \sin\delta|1\rangle|1\rangle$, and the probability of an explosion is $\sin^2\delta$. Detecting an explosion is a measurement. Thus, if nothing is heard, we project both registers back to $|0\rangle|0\rangle$. The state of the system after a successful round is just what it was in the beginning! If the bomb never explodes, the control register remains in the state $|0\rangle$ throughout the N rounds. The probability of getting no explosion in these N steps is

$$\left(1-\sin^2\delta\right)^N \geq \left(1-\frac{\pi^2}{4N^2}\right)^N \geq 1 - N\frac{\pi^2}{4N^2} = 1-\frac{\pi^2}{4N}. \tag{3}$$

Thus, we will live through the N steps and measure $|0\rangle$ in the first register with probability approaching 1, and conclude that there is a "live" bomb in the system.

To conclude, we can "safely" discern the quality of a "quantum bomb" by relying on the quantum Zeno effect (in other words, because a watched quantum pot never boils). Things get more interesting when we apply this test to other input states besides $|0\rangle$ in the second register. In the next section we show this results in a successful attack against Wiesner's quantum money scheme.

3 An adaptive attack on Wiesner's quantum money

We now show that if validated banknotes are forwarded to a new owner (or returned to the original owner), Wiesner's quantum money is vulnerable to an adaptive attack. The attack works also when

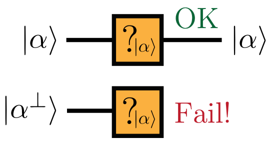

Fig. 2. A strict (unforgiving) money tester accepts a valid state and hands it back to us (or to whom we pay). However, if we hand in an orthogonal state, we fail and rot in jail.

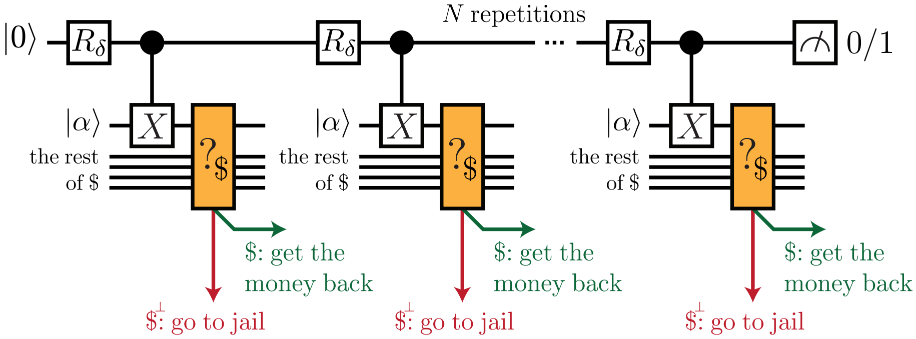

Fig. 3. An adaptive attack on Wiesner's quantum money with a strict testing procedure. We can identify whether the qubit $|\alpha\rangle$ is in the state $|+\rangle$ without going to jail (being detected). If we do not identify it, we can use controlled-$(-X)$ instead to test for $|-\rangle$. If we do not detect it either, we just measure the qubit in the computational basis.

the bank destroys bad bills, or sends us to jail if we try to validate a bill that does not pass the test (see Figure 2).[c]

Failing a validation test is undesirable, analogous to a "bomb explosion" from the previous Section. Motivated by the success of Elitzur-Vaidman's bomb tester, we will try to extract information about our system (learn in which of the 4 states the i^{th} qubit of the quantum money state is in) without triggering the "bomb" (being sent to jail).

Let us have a single valid banknote with the n-qubit state $|\alpha_1\rangle \otimes |\alpha_2\rangle \otimes \cdots \otimes |\alpha_n\rangle$ with $|\alpha_i\rangle \in \{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$. We show how to determine any particular single qubit state $|\alpha\rangle$ from the banknote by following the attack in Figure 3. There are two differences from Figure 1. First, the input state $|\alpha\rangle$ is unknown – a qubit from the banknote can be any one of the four possible states $|0\rangle, |1\rangle, |+\rangle, |-\rangle$. Second, instead of doing measurements in the $|0\rangle, |1\rangle$ basis on the second register, the bank measures the qubit in the (unknown to us) $|\alpha\rangle, |\alpha^\perp\rangle$ basis, verifying whether we gave it the state $|\alpha\rangle$ and the rest of the undisturbed banknote qubits. The test is unforgiving, sending us to jail if the money tests false, i.e. when the unknown qubit projects to $|\alpha^\perp\rangle$.

What happens in Figure 3 when we flip the four possible qubit states using an X operation?

1. First, we analyze the case that the unknown state is $|0\rangle$ or $|1\rangle$. Flipping maps the states $|0\rangle \leftrightarrow |1\rangle$ to each other, and the validity tester rejects the flipped state. Thus, we can view these two qubit states as the "live bomb" case in Figure 1b). Measuring the validity of the bill will result in

[c] As mentioned in the introduction, the scenario where the bank returns the bad bills, that is, when the attacker has access to the post-measurement result in failed cases, was already analyzed and successfully attacked by Lutomirski [7], Aaronson [8], and Farhi et al. [9].

projecting the control qubit back to $|0\rangle$ every time. In the end, the control (first register) will remain $|0\rangle$, and we will pass through the procedure with probability arbitrarily close to one (3).

2. Second, when the unknown qubit is $|+\rangle$, a flip does nothing to it. Thus, it behaves like the case with a "dud" in Figure 1a). In Figure 3, this results in a rotation of the the control qubit to $|1\rangle$ in N steps. We are never caught doing anything illegal in this case.

3. Finally, when the input state is $|-\rangle$, a bit flip gives it a minus sign, which makes things more interesting. The initial state is $|0\rangle|-\rangle$. We apply the rotation R_δ and a CNOT, obtaining

$$R_\delta \otimes \mathbb{I}: \ \left((\cos\delta)|0\rangle + (\sin\delta)|1\rangle\right)|-\rangle, \tag{4}$$

$$\text{CNOT}: \ \left((\cos\delta)|0\rangle - (\sin\delta)|1\rangle\right)|-\rangle. \tag{5}$$

   The quantum money still passes the test perfectly. However, the relative sign of the first register state is now negative, and its angle with the $|0\rangle$ state is $-\delta$. Let us do the second iteration. We rotate the first register from $-\delta$ to $|0\rangle$, and the following CNOT does not do anything. The third round is just like the first round, the fourth like the second, etc. After an even number of rounds, the state of the first register will be $|0\rangle$. Meanwhile, all the tests will have passed perfectly, and we are without any danger of being caught.

Therefore, if the qubit $|\alpha\rangle$ was in the $|+\rangle$ state, we can identify it accurately, with impunity. How to identify the other three cases? We can test for the state $|-\rangle$ using the controlled-$(-X)$ operation in Figure 3. Hence, we can rule out (or certify) $|-\rangle$ as well, conclude that $|\alpha\rangle \in \{|0\rangle, |1\rangle\}$, and safely measure it in the computational basis.

Wiesner's strict testing (good money returning, bad money confiscating), 4-state $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$ scheme is thus vulnerable to an adaptive attack. Suppose we want our attack to fail with probability at most f for $f \ll 1$. For a bill with n secret qubits, we choose $N = \frac{\pi^2 n}{2f}$, and run the procedure in Figure 3 at most twice per qubit – once for every qubit i in order to distinguish whether the i^{th} qubit is in the state $|+\rangle$, and if it was not, run the alternative test to distinguish whether it is in the state $|-\rangle$. If it isn't – the state is a computational basis state, and we can now safely measure it. Thus, $2N$ verifications are necessary to identify a single qubit. For each run of the N-step procedure in Figure 3, the probability of failing is at most $\frac{\pi^2}{4N}$, according to Eq. (3). Therefore,

$$\Pr(\text{attack succeeds}) \geq \left(1 - \frac{\pi^2}{4N}\right)^{2n} \geq 1 - \frac{\pi^2 n}{2N} = 1 - f, \tag{6}$$

Thus, we can identify all n qubits using at most $2n \times N$ adaptive questions to a strict (unforgiving) tester.

However, we can submit a banknote for verification with all n qubits slightly changed (in parallel, with one probe qubit per system qubit). This way, each verification fails with probability $O\left(n\delta^2\right)$, and running N successful rounds fails with probability $O\left(Nn\delta^2\right) = O\left(nN^{-1}\right)$. Thus, if we set f for our final failure probability, it is enough to use $N = O\left(nf^{-1}\right)$, which translates to only

$$2N = \frac{\pi^2 n}{f} \tag{7}$$

verification rounds, saving a factor of n.

We can try to salvage Wiesner's strict testing money scheme from this attack by adding other states for the banknote qubits. Adding the y-basis states does not help, as the CNOT flips them to orthogonal states, and they correspond to a "live bomb" case again. To detect them, we would perform tests with a controlled-Y or controlled-$(-Y)$ instead of a CNOT. We could also try to use a collection of unrelated single-qubit states. If the list of possible states is finite (and known), we can use a similar procedure, outlined in Appendix A. However, we can also deal with the most general case (where the list of possible states is infinite, or is unknown) as shown in the next Section.

4 Another way to attack: a protective measurement

In the BT attack above, the quantum Zeno effect kept us safe from explosions (our attack being detected by the bank). It worked nicely because of the special relationship between the four states $|0\rangle, |1\rangle, |+\rangle, |-\rangle$. They were flipped or kept intact. However, what would happen if we analyzed a different list of single-qubit quantum money states, on which the CNOT operation did something else? The behavior of the probe system and the probabilities of failure are calculated in detail in Appendix A. We see there that if two of the possible states are very close to each other, the BT attack is not weak enough, and we cannot get a satisfactory upper bound on the cumulative probability of failure. We now present a different attack which ensures that we are safe enough in any round (and in summary, in an N-round procedure). Below we show that it is possible to construct an approximate state ρ such that the fidelity[d] $F(\rho, |\alpha\rangle) > 1 - \epsilon$ and estimate the running time, confidence levels and probability of success for the procedure. The fidelity squared gives the probability that the bank will accept our counterfeit ρ as valid.

The basic building block of this method is to ensure *weak interaction* between the probe and the system. The method is reminiscent of the *quantum random walk* and *protective measurement* ideas of [18, 19]. We let a probe system interact weakly with the bill at each step while maintaining coherence of the bill state. The probe state will evolve as a linear function of the weakness parameter δ, while the probability of a failed validation will be quadratic in δ. The procedure is called a protective measurement since the validation step protects the money state by projecting it back to its original state with high probability.

The protective measurement scheme was originally derived as a method to fully measure the wavefunction of an unknown protected quantum state (essentially performing tomography) without disturbing the state. The probe is usually a continuous variable that can be used to record expectation values with the desired accuracy. Since the motivation was conceptual rather than practical, there has never been any attempt to quantify the resources required for full tomography using this scheme. Apart from our use of protective measurement in a practical scenario, we describe a protective measurement scheme with a qubit probe and bound its running time.

First, we will describe how the validation procedure of the bank can be used to estimate the expectation value of any dichotomic observable (i.e., an observable with eigenvalues ± 1)

$$A = P - P^{\perp}, \tag{8}$$

where P is a projector on its $+1$ eigenspace (in this work, A will always be one of the Pauli matrices),and $P^{\perp} = I - P$.

[d] The definition of the fidelity is $F(\rho, \sigma) = \mathrm{Tr}\left(\sqrt{\rho^{1/2}\sigma\rho^{1/2}}\right)$. It can be easily seen that if one of the two states is pure, the fidelity satisfies: $F(\rho, |\psi\rangle) = \sqrt{\langle\psi|\rho|\psi\rangle}$.

The basic idea for estimating $\langle A\rangle$ is to use a weak interaction between a probe in some state $|\varphi_0\rangle$ and the money state. Measuring the money state will then affect the probe state to a small degree. By repeating the weak interaction and validation, we finally (approximately) transform the probe state into:

$$|\varphi_N\rangle \approx e^{-ic\langle A\rangle\sigma_x}|\varphi_0\rangle, \tag{9}$$

for some constant c. The following is a formal description of the above intuition.

Definition 1 (Protective Measurement) *You are given a single copy of an unknown state $|\alpha\rangle \in \mathbb{C}^d$, and access to a two outcome von Neumann measurement $\{\Pi = |\alpha\rangle\langle\alpha|, I - \Pi\}$, the validation. We say that a protocol is a protective measurement of a dichotomic observable A with running time N, accuracy ϵ, and failure probability f when (a) The protocol makes at most N uses of the validation. (b) With probability that all the outcomes are Π is at least $1 - f$. In this case, the procedure maps $|\varphi\rangle|\alpha\rangle \to \left[e^{-i\frac{\pi}{8}\langle A\rangle\sigma_x}|\varphi\rangle + O(\epsilon)|\varphi'\rangle\right]|\alpha\rangle$ for all $|\varphi\rangle \in \mathbb{C}^2$.*

We note that this definition is slightly different then Aharonov and Vaidman's original definition [19]. In particular they use continuous variables for the meter and consider other possible protection methods.

We usually think of a measurement as a mapping which takes a quantum state to a probabilistic classical result (and perhaps the post-measurement state). A protective measurement, on the other hand, maps a quantum state to quantum state.

Lemma 1 *For any dichotomic observable A there exists a protective measurement protocol with running time N, accuracy $O(1/N)$ and failure probability $O(1/N)$.*

We prove this Lemma in Sec. 4.1 below. The procedure used in the proof can be slightly modified to identify one of the four Wiesner states (see Sec. 4.2).

By generating many copies of $|\varphi_N\rangle$ and measuring in the σ_y basis we get an estimate of $\langle A\rangle$:

Lemma 2 (Statistics from protective measurement) *For any $\nu, \eta, f > 0$, it is possible to use a protective measurement protocol to estimate $\langle A\rangle$ with precision at least ν, confidence at least $1 - \eta$, probability of failure $O(f)$ and running time $O\left(f^{-1}\nu^{-4}\ln^2(\eta^{-1})\right)$.*

The proof of this lemma in Sec. 4.3 can be generalized to protective measurement protocols with a wider range of parameters.

Estimating $\langle A\rangle$ will allow us to perform tomography to get a classical description of the money state $|\alpha\rangle$ and ultimately produce its (approximate) copy – and similarly for each unknown qubit. Formally,

Definition 2 (Protective Tomography) *You are given a single copy of an unknown state $|\alpha\rangle \in \mathbb{C}^d$, and access to a two outcome von Neumann measurement $\{\Pi = |\alpha\rangle\langle\alpha|, I - \Pi\}$ the validation. We say that a protocol achieves protective tomography with infidelity ϵ, confidence $1-\eta$, failure probability f and running time t if it outputs a classical description of a mixed state ρ such that: (a) the probability of failure, i.e. that at some step of the algorithm the outcome of the measurement is $I - \Pi$, is $O(f)$. (b) if the algorithm does not fail, with probability at least $1 - \eta$, we have $F(|\alpha\rangle, \rho) \geq 1 - \epsilon$, (c) the algorithm uses at most t validations.*

If the state $|\alpha\rangle$ is a product state of n qubits, such as Wiesner's scheme and its extension (with infinitely many possible states per qubit, instead of 4), We can repeat the above procedure with adjusted parameters, and get an approximate classical description for $|\alpha\rangle$, i.e.

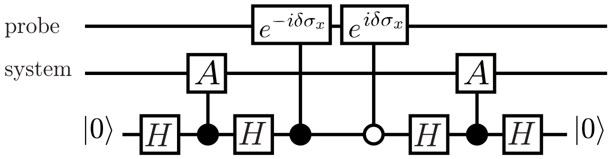

Fig. 4. An implementation of U from (10), involving projections on the ± 1 eigenspaces of the observable A.

Theorem 1 *There exists a protective tomography protocol for a qubit system ($d = 2$) with running time scaling as $t = O\left(f^{-1}\epsilon^{-4}\ln^2(\eta^{-1})\right)$.*

We present the proof in Section 4.4. Note that it can be easily extended to qudits of dimension d, with the running time now scaling as $t = O\left(d^{12}f^{-1}\epsilon^{-4}\ln^2(d^2\eta^{-1})\right)$. Next, we want to be able to perform protective tomography for a composite n-qubit product state. Theorem 1 implies the following

Corollary 1 *There exists a protective tomography protocol for n-qubit states of the form $|\alpha\rangle = \bigotimes_{i=1}^{n}|\alpha_i\rangle$, with running time $t = O\left(n^5 f^{-1}\epsilon^{-4}\ln^2(n\eta^{-1})\right)$.*

We prove this Corollary in Section 4.5. In terms of the attack on Wiesner's quantum money, this is the final step producing the approximate classical description of $|\alpha\rangle$, with the probability of not being caught at least $1 - f$. This in turn allows us to create as many counterfeit bills ρ as we want. We also have a guarantee that with probability at least $1 - \eta$, these counterfeit bills are going to be pretty good approximates of the original bill. Finally, if ρ is a good approximation of $|\alpha\rangle$, then the first time each counterfeit bill is used, it can fail the bank's validation procedure with probability at most 2ϵ.

4.1 Proof of lemma 1: Coupling to the expectation value of a dichotomic observable

The crucial difference from the approach of Section 2 is that, instead of rotating by δ and applying a CNOT to the probe/system, we use the unitary coupling operation

$$U = e^{-i\delta(\sigma_x\otimes A)} = e^{-i\delta\left(\sigma_x\otimes P-\sigma_x\otimes P^{\perp}\right)} = e^{-i\delta\sigma_x\otimes P}e^{i\delta\sigma_x\otimes P^{\perp}} = e^{-i\delta\sigma_x}\otimes P + e^{i\delta\sigma_x}\otimes P^{\perp}, \quad (10)$$

expressible in terms of the projector P. In this formula, we assume that system containing the probe is on the left, and the system containing the unknown state on the right of the tensor product.) This works because we can divide the Hilbert space into two subspaces: the kernel and the range of P. These subspaces are invariant under the unitary U. Its action in each of the subspaces is then easily expressible. Note that when A is unitary with eigenvalues ± 1, the unitary U can be implemented as in Figure 4.

Choosing a small $\delta = \frac{c}{N}$ (and tuning the constant c for optimal performance), we ensure that the interaction of the probe and the tested system is always weak – independent of the relationship of P and the unknown state $|\alpha\rangle$. This is in contrast with the method in Section 3, where if the probe was close to the state $|1\rangle$, the controlled interaction could change the state of the system a lot.

We start with the probe qubit in an arbitrary state $|\varphi_0\rangle$. After k rounds (assuming we have not yet been caught) the probe register will be in some state $|\varphi_k\rangle$ and the second register will hold the unknown state $|\alpha\rangle$. Let us apply U again, and get the state $U|\varphi_k\rangle|\alpha\rangle$. This is followed by the validation, which gives us the unnormalized state

$$W|\varphi_k\rangle = \left(\mathbb{I}\otimes\langle\alpha|\right)U|\varphi_k\rangle|\alpha\rangle = \sqrt{p_k}|\varphi_{k+1}\rangle, \quad (11)$$

where the normalization constant p_k is the probability of avoiding detection in the k-th step and

$$\begin{aligned} W &= \langle\alpha|P|\alpha\rangle e^{-i\delta\sigma_x} + \langle\alpha|P^\perp|\alpha\rangle e^{i\delta\sigma_x} \\ &= (\cos\delta)\langle\alpha|(P+P^\perp)|\alpha\rangle\mathbb{I} - i(\sin\delta)\langle\alpha|(P-P^\perp)|\alpha\rangle\sigma_x \\ &= (\cos\delta)\mathbb{I} - i(\sin\delta)\langle A\rangle\sigma_x, \end{aligned} \tag{12}$$

as $P + P^\perp = \mathbb{I}$ and $\langle A\rangle = \langle\alpha|A|\alpha\rangle = \langle\alpha|P - P^\perp|\alpha\rangle$, recalling 8. The matrix W has eigenvalues $\lambda_\mp = \cos\delta \mp i\langle A\rangle \sin\delta$ and eigenstates $|+\rangle, |-\rangle$.

Because the above holds for any k, we have

$$W^N|\varphi_0\rangle = \left(\prod_{k=0}^{N-1}\sqrt{p_k}\right)|\varphi_N\rangle = \sqrt{p_{pass}}|\varphi_N\rangle, \tag{13}$$

where p_{pass} is the probability to pass all N validation steps. Let us look at what this becomes for large N, recalling $\delta = \frac{c}{N}$.

$$\lambda_\mp^N = (\cos(\delta) \mp i\sin(\delta)\langle A\rangle)^N = \left(e^{\mp i\delta\langle A\rangle} + O\left(\delta^2\right)\right)^N = \left(e^{\mp i\delta\langle A\rangle}\left(1 + O\left(\delta^2\right)\right)\right)^N \tag{14}$$

$$= e^{\mp iN\delta\langle A\rangle}\left(1 + N \times O\left(\delta^2\right)\right) = e^{\mp ic\langle A\rangle} + O\left(N^{-1}\right). \tag{15}$$

When we choose N to be large (thus, small δ), we get

$$W^N = e^{-ic\langle A\rangle\sigma_x} + O\left(\frac{1}{N}\right), \tag{16}$$

meaning that we have approximately rotated the probe system by an amount proportional to $\langle A\rangle$. Furthermore,

$$\sqrt{p_{pass}}|\varphi_N\rangle = e^{-ic\langle A\rangle\sigma_x}|\varphi_0\rangle + O\left(\frac{1}{N}\right)|\tilde{\varphi}\rangle, \tag{17}$$

with $|\tilde{\varphi}\rangle$ some normalized state. By Eq. (13) this means the probability that we failed some validation is small:

$$p_{pass} = 1 - O\left(\frac{1}{N}\right), \tag{18}$$

and the final state can be also rewritten as

$$|\varphi_N\rangle = e^{-ic\langle A\rangle\sigma_x}|\varphi_0\rangle + O\left(\frac{1}{N}\right)|\varphi'\rangle, \tag{19}$$

where $|\varphi'\rangle$ is some normalized state. We have demonstrated that we can apply the transformation W^N with a low probability of failure and high precision, i.e. $f = O(1/N)$ and $\eta = O(1/N)$. To prove Lemma 1 we set $c = \frac{\pi}{8}$.

4.2 A simple protective measurement: Identifying the 4 Wiesner money states

Before describing the solution to the problem of protective tomography we give a simple application of the previous subsection for the case of the four Wiesner states.

To identify of the four states $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$ we can use a slightly modified version of the procedure described above. We choose $A = \sigma_x$, $c = \frac{\pi}{2}$, $|\varphi_0\rangle = |0\rangle$ and recall $\langle 0|\sigma_x|0\rangle = \langle 1|\sigma_x|1\rangle = 0$ and $\langle +|\sigma_x|+\rangle = -\langle -|\sigma_x|-\rangle = 1$,

If the money qubit $|\alpha\rangle$ was initially $|+\rangle$ or $|-\rangle$, the operation W in Eq. (12) is exactly $e^{\mp i\delta\sigma_x}$. It is unitary, which means we never fail a verification. The final probe state in these two cases will be $W^N|0\rangle = \mp i|1\rangle$.

On the other hand, if the money state is $|0\rangle$ or $|1\rangle$, then W is approximately the identity and the probe will remain close to $|0\rangle$.

Thus, when we measure the probe after N rounds of interaction/verification in the computational basis, we find out whether the unknown state is an x-basis or z-basis state, allowing us to uniquely identify $|\alpha\rangle$ by measuring it in this basis.

4.3 Proof of Lemma 2: Single copy estimation of an expectation value

The procedure for estimating the expectation value $\langle A\rangle$ is based on running the protocol in Sec. 4.1 for $m = 336\ln(2\eta^{-1})\nu^{-2}$ times, with $N = m/f$. We assume, w.l.o.g. that the precision parameter ν is smaller than some universal constant (if not, replace ν with that constant, and the result is improved). We can already see using Lemma 1 that the total running time is $mN = O(\ln^2(\eta^{-1})\nu^{-4}f^{-1})$, and that the overall failure probability is $O(\frac{m}{N}) = O(f)$. By applying the protective measurement procedure successfully m times we obtain m copies of the state

$$= |\varphi_N\rangle = \cos\left(\frac{\pi}{8}\langle A\rangle\right)|0\rangle - i\sin\left(\frac{\pi}{8}\langle A\rangle\right)|1\rangle + O\left(\frac{1}{N}\right)|\varphi'\rangle, \tag{20}$$

with some unknown normalized (error) state $|\varphi'\rangle$. We can collect the desired statistics for estimating $\langle A\rangle$ by measuring each of these m copies in the σ_y basis. Let $\bar{p}_{y_+}$ be the probability for getting the result $+1$ and $p^{(m)}_{y_+}$ be the empirical frequency of a $+1$ result in these m measurements.

Using $\langle y_+| = \frac{1}{\sqrt{2}}[\langle 0| - i\langle 1|]$, we get

$$\bar{p}_{y_+} = |\langle y_+|\varphi_N\rangle|^2 = \frac{1}{2}\left|\cos\left(\frac{\pi}{8}\langle A\rangle\right) - \sin\left(\frac{\pi}{8}\langle A\rangle\right) + O\left(\frac{1}{N}\right)\right|^2 \tag{21}$$

$$= \frac{1}{2}\left(1 - 2\sin\left(\frac{\pi}{8}\langle A\rangle\right)\cos\left(\frac{\pi}{8}\langle A\rangle\right)\right) + O\left(\frac{1}{N}\right) \tag{22}$$

$$= \frac{1}{2}\left(1 - \sin\left(\frac{\pi}{4}\langle A\rangle\right)\right) + O\left(\frac{1}{N}\right). \tag{23}$$

Next, we will show that we can estimate $\langle A\rangle$ by $\frac{4}{\pi}\arcsin\left(1 - 2p^{(m)}_{y+}\right)$, with precision ν and confidence $1 - \eta$.

Theorem 2 (Chernoff bound, see, e.g. [20, p. 66]) *Let $X_1, \ldots, X_m$ be independent Bernoulli trials, $X = \frac{1}{m}\sum_{i=1}^m X_i$, $\mu = \mathbb{E}[X]$. For $0 < \tilde{\nu} < 1$,*

$$\Pr(|X - \mu| \geq \tilde{\nu}\mu) \leq 2\exp\left(-\frac{\tilde{\nu}^2 m\mu}{3}\right).$$

Let us choose $\tilde{\nu} = \frac{\nu}{4}$ and recall that $m = \frac{336\ln(2/\eta)}{\nu^2}$. Since $|\langle A\rangle| \leq 1$, the expectation value $\bar{p}_{y_+}$ is well bounded away from zero: $\bar{p}_{y_+} \geq \frac{1}{2} - \frac{1}{\sqrt{8}} + O\left(\frac{1}{N}\right) \geq \frac{1}{7}$. Therefore,

$$m = \frac{336\ln(2/\eta)}{\nu^2} \geq \frac{3\ln(2/\eta)}{\tilde{\nu}^2 \bar{p}_{y+}}$$

By the Chernoff bound, the probability that $\left|p_{y_+}^{(m)} - \bar{p}_{y_+}\right| \geq \tilde{\nu}\bar{p}_{y_+}$ is at most

$$2\exp\left(-\frac{\tilde{\nu}^2 m \bar{p}_{y+}}{3}\right) \leq \eta.$$

Our next goal is to show that we can estimate $\langle A\rangle$ easily and well using $p_{y_+}^{(m)}$. We know that with confidence $1-\eta$, the value $p_{y_+}^{(m)}$is within $\frac{\nu}{4}\bar{p}_{y_+} \leq \frac{\nu}{4}$ of $\bar{p}_{y_+} = \frac{1}{2}\left(1 - \sin\left(\frac{\pi}{4}\langle A\rangle\right)\right) + O\left(\frac{1}{N}\right)$ (see Eq. (23)) . Moving the terms around, we get that $\sin\left(\frac{\pi}{4}\langle A\rangle\right)$ is within $\frac{\nu}{2} + O\left(\frac{1}{N}\right)$ of $1 - 2p_{y_+}^{(m)}$, or that

$$\arcsin\left(1 - 2p_{y_+}^{(m)} - \frac{\nu}{2} - O\left(\frac{1}{N}\right)\right) \leq \frac{\pi}{4}\langle A\rangle \leq \arcsin\left(1 - 2p_{y_+}^{(m)} + \frac{\nu}{2} + O\left(\frac{1}{N}\right)\right), \tag{24}$$

with confidence $1-\eta$.

How far can $\frac{\pi}{4}\langle A\rangle$ be from $\arcsin\left(1 - 2p_{y_+}^{(m)}\right)$? We can use the Taylor series expansion $\arcsin(x + \delta) = \arcsin(x) + \frac{\delta}{\sqrt{1-x^2}} + O(\delta^2)$ for $x = 1 - 2p_{y_+}^{(m)}$ and $\delta = \frac{\nu}{2} + O\left(\frac{1}{N}\right)$, assuming small ν and large N. It now proves useful that we chose our procedure so that $\bar{p}_{y_+}$ in (23) is in the range $\frac{1}{2} - \frac{1}{\ }\sqrt{8} - O(\frac{1}{N}) \leq \bar{p}_{y_+} \leq \frac{1}{2} + \frac{1}{\sqrt{8}} + O(\frac{1}{N})$. Because of this, for reasonably small ν and big enough N, we can bound $|x| = \left|1 - 2p_{y_+}^{(m)}\right| \leq \frac{1}{\sqrt{2}} + O(\nu) + O\left(\frac{1}{N}\right) \leq \frac{3}{4}$, and $\frac{1}{\sqrt{1-x^2}} \leq \frac{4}{\sqrt{7}}$. For small enough δ, we can have the $O(\delta^2)$ term in the Taylor series smaller in magnitude than $\frac{|\delta|}{20}$.[e]We can then conclude that with probability $1-\eta$, the value of $\frac{\pi}{4}\langle A\rangle$ is within $\left(\frac{4}{\sqrt{7}} + \frac{1}{20}\right)\delta \leq 0.781\nu$ (see Footnote e) of $\arcsin\left(1 - 2p_{y_+}^{(m)}\right)$, or alternatively,

$$= \left|\langle A\rangle - \frac{4}{\pi}\arcsin\left(1 - 2p_{y_+}^{(m)}\right)\right| \leq \nu. \tag{25}$$

This concludes our proof of Lemma 2.

4.4 Proof of Theorem 1: tomography of a single qubit

Using Lemma 2, let $\langle\tilde{\sigma}_j\rangle$ for $j \in \{x, y, z\}$ be the approximated expectation value of each of the three Pauli operators $\langle\sigma_j\rangle = \langle\alpha|\sigma_j|\alpha\rangle$, with precision parameters $\tilde{\nu} = \epsilon/6,\ \tilde{\eta} = \eta/3, \tilde{f} = f$. The running time for getting all three values is $3 \cdot O\left(\tilde{f}^{-1}\tilde{\nu}^{-4}\ln^2(\tilde{\eta}^{-1})\right) = O\left(f^{-1}\epsilon^{-4}\ln^2(\eta^{-1})\right)$ as required and the failure probability is $f \leq 3\tilde{f}$ by the union bound. Conditioned that there were no failures, by using the union bound again, we get the required confidence value

$$= \Pr\left(\bigcap_{j\in\{x,y,z\}} |\langle\tilde{\sigma}_j\rangle - \langle\sigma_j\rangle| \leq \tilde{\nu}\right) \geq 1 - 3\tilde{\eta} = 1 - \eta. \tag{26}$$

[e] This is guaranteed by the assumption we made in the beginning of the proof that ν is smaller than some universal constant.

To finish the proof, all we need is to show how, in the event there were no failures, we can construct the state ρ from the three approximate expectation values $\langle\tilde{\sigma}_j\rangle$ such that ρ is an approximation of $|\alpha\rangle\langle\alpha| = \mathbb{I}/2 + \sum_{j\in\{x,y,z\}}\langle\sigma_j\rangle\sigma_j$ with fidelity at least $1-\epsilon$ and confidence at least $1-\eta$. Let $\tilde{\rho} = \mathbb{I}/2 + \sum_{j\in\{x,y,z\}}\langle\tilde{\sigma}_j\rangle\sigma_j$. This Hermitian trace one operator is not necessarily positive semidefinite. We therefore choose ρ to be the closest state to $\tilde{\rho}$, that is $\rho = \arg\min_\tau D(\tilde{\rho},\tau)$, where τ runs over all single qubit mixed states and $D(.,.)$ is the trace distance $D(\alpha,\beta) = \frac{1}{2}||\alpha-\beta||_{tr}$ and $||A||_{tr} = Tr(\sqrt{AA^\dagger}))$. Using the triangle inequality and the definition of ρ, we obtain

$$= D(\rho,|\alpha\rangle\langle\alpha|) \leq D(\rho,\tilde{\rho}) + D(\tilde{\rho},|\alpha\rangle\langle\alpha|) \leq 2D(\tilde{\rho},|\alpha\rangle\langle\alpha|). \tag{27}$$

Then the fidelity of the final state satisfies that with probability at least $1-\eta$,

$$\begin{aligned} F(\rho,|\alpha\rangle\langle\alpha|) &\geq 1 - D(\rho,|\alpha\rangle\langle\alpha|) && (28)\\ &\geq 1 - 2D(\tilde{\rho},|\alpha\rangle\langle\alpha|) && (29)\\ &\geq 1 - 2\sum_{j\in\{x,y,z\}}\frac{1}{2}\left\|\left(\langle\sigma_j\rangle - \langle\tilde{\sigma}_j\rangle\right)\sigma_j\right\|_{tr} && (30)\\ &= 1 - 2\sum_{j\in\{x,y,z\}}\left|\langle\sigma_j\rangle - \langle\tilde{\sigma}_j\rangle\right| \geq 1 - 6\tilde{\nu} = 1-\epsilon, && (31) \end{aligned}$$

where in the first step, we used the fact that if one of the states α or β is pure, then $F(\alpha,\beta) \geq 1 - D(\alpha,\beta)$; in the second step we used Eq. (27); in the third step we used the triangle inequality for the trace norm; we used Eq. (26) in the last inequality; and the definition of $\tilde{\nu} = \epsilon/6$ in the last step. The properties of the trace distance which we used are shown, for example, in the textbook [21].

This completes the proof of Theorem 1.

4.5 Proof of Corollary 1: Protective tomography of n qubits.

We use Theorem 1 to apply tomography to each of the n qubits, with parameters $\tilde{\epsilon} = \epsilon/n$, $\tilde{\eta} = \eta/n$, $\tilde{f} = f/n$. By the union bound, the failure probability is at most $n\tilde{f} = f$, and the error probability is at most $n\tilde{\epsilon} = \epsilon$, as required. Let $\rho = \rho_1 \otimes \ldots \otimes \rho_n$, where ρ_i is the $\tilde{\epsilon}$ approximation of $|\alpha_i\rangle$, as provided by the theorem.

$$F(\rho,|\alpha\rangle\langle\alpha|) = \prod_{i=1}^{n} F(\rho_i,|\alpha_i\rangle\langle\alpha_i|) \geq (1-\tilde{\epsilon})^n \geq 1 - n\tilde{\epsilon} = 1-\epsilon$$

where the first step follows from $F(\alpha\otimes\beta,\gamma\otimes\delta) = F(\alpha,\gamma)F(\beta,\delta)$.

By applying this procedure directly, we get a total running time of $n \cdot O\left(n^5 f^{-1}\epsilon^{-4}\ln^2(n\eta^{-1})\right)$. But we can save a factor of n to get a total running time of $O\left(n^5 f^{-1}\epsilon^{-4}\ln^2(n\eta^{-1})\right)$, as required, by running the procedures in parallel, using the same idea that was explained in the previous section, see the analysis preceding Eq. (7). This completes the proof of Corollary 1.

5 Discussion: Comparing the two attacks

We presented two different attacks for counterfeiting quantum money in the strict testing regime. Both attacks are based on known methods to exploit the quantum Zeno effect in order to learn a quantum state without disturbing it. Despite their similarities, the methods are conceptually different and are not always interchangeable. In Appendix 1, we discuss a generalized version of the BT attack and show that it fails in the most general scenario. There, we have to choose the PM attack.

One may ask why (or whether) the PM attack is less useful than BT in other scenarios. The first possible advantage of the BT attack is that it identifies a state from a given list of states, instead of producing only an estimate of the state. However, a simple modification of the PM protocol can also identify the state, in a similar fashion as the BT attack (just like the PM attack in Section 4.2). Another advantage might be in terms of resources. Although our analysis in both cases is not necessarily optimal, in all the cases we checked the PM attack (or simple adaptations of it) does not use more queries than the BT attack. To conclude, our analysis suggests that the PM attack has both qualitative and quantitative advantages compared to the BT attack.

One way to gain efficiency may be to use phase estimation on the transformation for W in Eq. (12), which is approximately a rotation. We could improve the efficiency even further by using multiple probes per qubit of the money state. The two attacks also lead to different behaviors of the probe qubit: in the BT attack, the probe qubit moves only if the money state lies in a very narrow window of angles around the reflected state (see Fig. A.1); in the PM attack, the probe qubit rotates by an angle proportional to the expectation value that we want to measure.

In both attacks, we have two systems, the probe and the money. The fundamental idea is to cause a minimal kick to the money state in each iteration. Let us use Q_i to describe the completely positive trace preserving channel acting on the money state in the ith iteration and $V_\$$ to indicate the bank's verification step that follows Q_i. Ideally, we want $Q_i(|\$\rangle\langle\$|) \approx |\$\rangle\langle\$|$ for all i. This could happen for one of two reasons, either $|\$\rangle\langle\$|$ is a fixed state for the channels Q_i, or $Q_i \approx \mathbb{I}$ (this is what we call a weak channel). In general, the channel is not fixed, since it depends on the state of the pointer. For both methods, the initial input state is unknown, and therefore the channel has to be initially weak. Does it remain this way?

In the case of the PM attack, the interaction term is weak, so the corresponding channel is also always weak. Moreover, the attack is set up so that the channel is invariant in time to a good approximation. The BT attack is more interesting. Initially, the channel is weak, because the pointer is very close to the $|0\rangle$ state. For most input states, the pointer stays very close to its initial position (see detailed analysis in Appendix 1). However, for a small family of states around one of the fixed points, the pointer starts to move towards the state $|1\rangle$ and the channel becomes stronger. This strange behavior of the channel that goes from weak to strong as a result of different input states has not been previously noticed, as far as we know. We believe that further insight into these types of measurements can lead to interesting quantum information protocols beyond those mentioned here.

Acknowledgments

We would like to thank Scott Aaronson, Dorit Aharonov, David Gosset, Guy Kindler, Carl Miller, Lev Vaidman and Stephen Wiesner for valuable discussions, and the Simons institute Quantum Hamiltonian Complexity program, during which a substantial part of this work was done. DN has received funding from the People Programme (Marie Curie Actions) EU's 7th Framework Programme under REA grant agreement No. 609427, and this research has been further co-funded by the Slovak Academy of Sciences. He also thanks the Slovak Research and Development Agency grant APVV-14-0878 QETWORK. OS also thanks the ARO Grant W922NF-09-1-0440 and NSF Grant CCF-0905626. DU was supported by the Estonian ICT program 2011-2015 (3.2.1201.13-0022), the European Union through the European Regional Development Fund through the sub-measure "Supporting the development of R&D of info and communication technology", by the European Social Fund's Doctoral Studies and Internationalisation Programme DoRa, and by the Estonian Centre of Excellence in Computer Science, EXCS. AB is supported by NSERC, Industry Canada and CIFAR.

References

1. S. Wiesner, Conjugate coding, *ACM Sigact News*, **15** (1), 78–88 (1983).
2. C.H. Bennett, G. Brassard, S. Breidbart, and S. Wiesner, Quantum cryptography, or unforgeable subway tokens, *Advances in Cryptology*, 267–275 (Springer, 1983).
3. A.C. Elitzur and L. Vaidman, Quantum mechanical interaction-free measurements, *Foundations of Physics*, **23** (7), 987–997 (1993).
4. P. Kwiat, H. Weinfurter, T. Herzog, A. Zeilinger, and M.A. Kasevich, Interaction-free measurement, *Physical Review Letters*, **74** (24), p4763 (1995).
5. Y. Aharonov, J. Anandan, and L. Vaidman, Meaning of the wave function, *Physical Review A*, **47** (6), p4616 (1993).
6. W.K. Wootters and W.H. Zurek, A single quantum cannot be cloned, *Nature*, **299** (5886), 802–803 (1982).
7. A. Lutomirski, An online attack against Wiesner's quantum money, *arXiv preprint arXiv:1010.0256*, (2010).
8. S. Aaronson,Quantum copy-protection and quantum money, *Conference on Computational Complexity*, 229–242 (IEEE 2009).
9. E. Farhi, D. Gosset, A. Hassidim, A. Lutomirski, D. Nagaj, Daniel and P. Shor, Quantum state restoration and single-copy tomography for ground states of hamiltonians, *Physical review letters*, **105** (19), p190503 (2010).
10. A. Molina, T. Vidick, and J. Watrous, Optimal counterfeiting attacks and generalizations for Wiesners quantum money, *Theory of Quantum Computation, Communication, and Cryptography*, 45–64 (Sprinter 2013).
11. F. Pastawski, N.Y. Yao, L. Jiang, M.D. Lukin, and J.I. Cirac, *Proceedings of the National Academy of Sciences*, **109** (40), 16079-16082 (2012).
12. D. Gavinsky, Quantum money with classical verification, *IEEE 27th Annual Conference on Computational Complexity*, 42–52 (2012).
13. S. Aaronson and P. Christiano, *Proceedings of the 44th Symposium on Theory of Computing*, 41–60 (ACM 2012).
14. J. Dressel, M. Malik, F.M. Miatto, A.N. Jordan, and R.W. Boyd, *Rev. Mod. Phys.*, **86** (1), 307-316 (2014).
15. Y. Aharonov and L. Vaidman, *Time in Quantum Mechanics*, 399–447 (Springer 2007).
16. C. Y-Y Lin and H-H Lin, *Proceedings of the 30th Conference on Computational Complexity*, CCC '15 (30), 537–566 (2015).
17. E. Farhi, D. Gosset, A. Hassidim, A. Lutomirski, and P. Shor, Quantum money from knots, In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 276–289 (ACM 2012).
18. Y. Aharonov, L. Davidovich, and N. Zagury, Quantum random walks, *Physical Review A*, **48** (2), p1687 (1993).
19. Y. Aharonov and L. Vaidman, Measurement of the schrödinger wave of a single particle, *Physics Letters A*, **178** (1), 38–42 (1993).
20. M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*, (Cambridge University Press, 2005).
21. M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, (Cambridge university press, 2010).

Appendix A The basic bomb-tester and semi-faulty bombs

Let us now discuss the BT attack of Section 3, and demonstrate that it works efficiently and safely for a given list of hidden states, but stops working when the secret states are completely unknown, or chosen too close to each other.

A.1 Using a longer list of secret states is still insecure against the bomb-testing adaptive attack

Counterfeiting quantum banknotes with arbitrary secret single-qubit states becomes more complicated. Let us continue what we did before and try to identify an unknown qubit state $|\alpha\rangle$, if we have access to a strict validation procedure (see Figure 1). Our approach is to try to find a unitary R which acts as $R|\alpha\rangle = |\alpha\rangle$.

Let us assume that the unknown state comes from a list of possible states $|\alpha\rangle \in \mathcal{S}$. We can just pick one of the states $|\beta\rangle$ from this list, and use our previous attack with the controlled reflection $R = 2|\beta\rangle\langle\beta| - \mathbb{I}$ instead of the controlled X (CNOT) in Figure 3.

First, if the unknown state $|\alpha\rangle$ is exactly our tested $|\beta\rangle$, we have the "dud case", which we identify without fail by measuring $|1\rangle$ on the first register (as $R|\alpha\rangle = |\alpha\rangle$).

Second, if $|\alpha\rangle \neq |\beta\rangle$, the operation R does "something" to $|\alpha\rangle$. We can choose the phases of the vectors involved to give us a real angle $0 \leq \theta \leq \frac{\pi}{2}$ with $\cos\theta = \langle\alpha|\beta\rangle$ and

$$R|\alpha\rangle = \cos(2\theta)|\alpha\rangle + \sin(2\theta)|\alpha^{\perp}\rangle. \tag{A.1}$$

Let us now look at our testing procedure in Figure 3. If we haven't been caught after some number k of rounds, the control register contains a state parametrized by an angle φ_k, i.e. $|\varphi_k\rangle \propto (\cos\varphi_k)|0\rangle + (\sin\varphi_k)|1\rangle$. We have the valid quantum money state in the second register. After rotating the first register by δ, the state of the system becomes

$$\left(\cos(\varphi_k + \delta)|0\rangle + \sin(\varphi_k + \delta)|1\rangle\right)\,|\alpha\rangle. \tag{A.2}$$

We apply the controlled probe C_R (instead of CNOT) to both registers, and obtain

$$\cos(\varphi_k + \delta)|0\rangle|\alpha\rangle + \cos(2\theta)\sin(\varphi_k + \delta)|1\rangle|\alpha\rangle + \sin(2\theta)\sin(\varphi_k + \delta)|1\rangle|\alpha^{\perp}\rangle. \tag{A.3}$$

We now measure the second register. The probability of being caught in this round is $\sin^2(2\theta)\sin^2(\varphi_k + \delta)$. More importantly, after a successful test, we are left with the unnormalized state

$$\left(\cos(\varphi_k + \delta)|0\rangle + \cos(2\theta)\sin(\varphi_k + \delta)|1\rangle\right)\,|\alpha\rangle. \tag{A.4}$$

It means that the state of the first register as a two component (unnormalized) vector is transformed as

$$|\varphi_{k+1}\rangle = \begin{bmatrix} 1 & 0 \\ 0 & \cos(2\theta) \end{bmatrix} R_\delta|\varphi_k\rangle = \underbrace{\begin{bmatrix} \cos\delta & -\sin\delta \\ q\sin\delta & q\cos\delta \end{bmatrix}}_{T}|\varphi_k\rangle = T|\varphi_k\rangle, \tag{A.5}$$

where we labeled

$$q = \cos(2\theta). \tag{A.6}$$

When we pass all the tests, the unnormalized state of the first register at the end of the protocol is $|\varphi_N\rangle = T^N|0\rangle$. We can get the probability of passing all the tests by taking the norm squared of this vector.

There are two simple special cases. First, when we guess correctly, we have $\theta = 0$ ($q = 1$), which gives us $T = R_\delta$ and $\varphi_{k+1} = \varphi_k + \delta$ (the "dud" case). Then $\langle 1|T^N|0\rangle = 1$, and in N steps the first qubit is rotated to $|1\rangle$, plus we cannot get caught as since we are doing nothing to the banknote. Second, for $\theta = \frac{\pi}{2}$ ($q = -1$), we have $T^2 = \mathbb{I}$, and the system behaves like the $|-\rangle$ case in Section 3, with the first register remaining in the state $|0\rangle$ after an even number of rounds, i.e. $\langle 0|T^N|0\rangle = 1$. We are also never caught doing anything illegal.

Third, we have a much more interesting general case $\theta_{min} \leq \theta < \frac{\pi}{2}$, meaning $|q| < 1$. We claim that we will pass all the tests successfully, and measure $|0\rangle$ in the first register at the end, with probability close to 1. We are choosing a large N, which means small δ, for which we can write

$$T = \begin{bmatrix} 1 & -\delta \\ q\delta & q \end{bmatrix} + \Delta T, \tag{A.7}$$

with a small error term $\|\Delta T\| = O(\delta^2)$. Let us now calculate the state of the first register after N rounds using (A.7), keeping track of the errors we accumulate:

$$\begin{aligned}
T^N|0\rangle = T^N \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= T^{N-1} \begin{bmatrix} 1 \\ \delta q \end{bmatrix} + \underbrace{T^{N-1}\Delta T \begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{|v_1\rangle} \\
&= T^{N-2} \begin{bmatrix} 1 \\ \delta\left(q+q^2\right) \end{bmatrix} + \underbrace{T^{N-2}\left(\begin{bmatrix} -\delta^2 q \\ 0 \end{bmatrix} + \Delta T \begin{bmatrix} 1 \\ \delta q \end{bmatrix} \right)}_{|v_2\rangle} + |v_1\rangle \\
&= T^{N-3} \begin{bmatrix} 1 \\ \delta\left(q+q^2+q^3\right) \end{bmatrix} + \underbrace{T^{N-3}\left(\begin{bmatrix} -\delta^2\left(q+q^2\right) \\ 0 \end{bmatrix} + \Delta T \begin{bmatrix} 1 \\ \delta\left(q+q^2\right) \end{bmatrix} \right)}_{|v_3\rangle} + |v_2\rangle + |v_1\rangle \\
&= \begin{bmatrix} 1 \\ \delta\left(q+q^2+\cdots+q^N\right) \end{bmatrix} + |v_N\rangle + \cdots + |v_1\rangle, \qquad \text{(A.8)}
\end{aligned}$$

where the error vectors and their norms are

$$|v_k\rangle = T^{N-k}\left(\begin{bmatrix} -\delta^2\left(q+q^2+\cdots+q^{k-1}\right) \\ 0 \end{bmatrix} + \Delta T \begin{bmatrix} 1 \\ \delta\left(q+q^2+\cdots+q^{k-1}\right) \end{bmatrix} \right), \tag{A.9}$$

$$\||v_k\rangle\| \leq O\left(\delta^2\left(1+q+q^2+\cdots+q^{k-1}\right)\right) \leq O\left(\frac{\delta^2}{1-q}\right). \tag{A.10}$$

because we have $\|T\| \leq 1$ in (A.5) implying $\left\|T^k|w\rangle\right\| \leq \||w\rangle\|$ for any $k \geq 1$, as well as $\|\Delta T\| = O\left(\delta^2\right)$. We can now look at (A.8) and lower bound the amplitude

$$\langle 0|T^N|0\rangle \geq 1 - N\||v_N\rangle\| \geq 1 - O\left(\frac{N\delta^2}{1-q}\right) \geq 1 - O\left(N^{-1}\theta_{min}^{-2}\right), \tag{A.11}$$

using $1-q = 1-\cos(2\theta) = \Omega\left(\theta^2\right)$ for small θ. Therefore, there exists a constant c such that when we choose

$$N = cf^{-1}\theta_{min}^{-2} = O\left(f^{-1}\theta_{min}^{-2}\right), \tag{A.12}$$

the probability of passing all N tests while measuring $|0\rangle$ at the end will be

$$\left|\langle 0|T^N|0\rangle\right|^2 \geq 1 - f. \tag{A.13}$$

This procedure rules out (or identifies) $|\beta\rangle$, one of the possible states from $\mathcal{S}$. We can now proceed to eliminate or identify further states, also getting rid of the pesky smallest possible angles θ_{min} which make us use large N, so that the procedure gets more efficient later on.

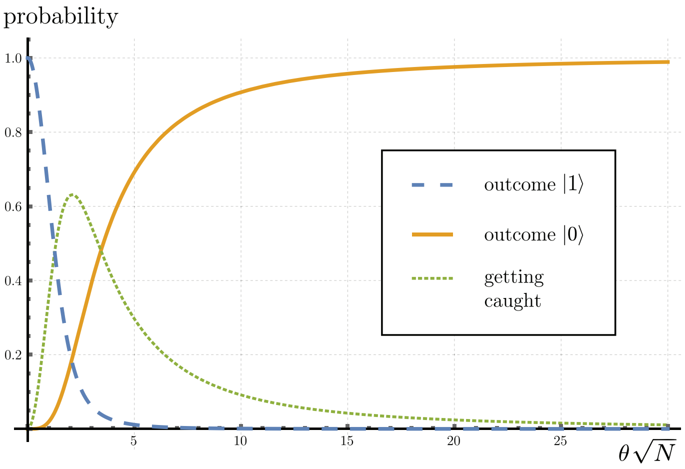

Fig. A.1. The probability of each of the outcomes of the protocol from Section 1 as a function of θ. Note that the x-axis is $\theta\sqrt{N}$. The probability of an getting caught by the bank, shown in the dotted green line, goes to 0 as $\theta\sqrt{N} \to \infty$, as well as with $\theta\sqrt{N} \to 0$.

However, in general, in the worst case, in order to identify a single qubit, we need to repeat this procedure $r := |\mathcal{S}|$ times. Moreover, we need to identify all n qubits. We can do it safely by choosing a much smaller $f' = f/(nr)$, resulting in $N' = O\left(nr\theta_{min}^{-2}f^{-1}\right)$, which makes the entire attack succeed with probability $1 - O\left(f\right)$. The total number of queries in the attack is at most $n \times r \times N' = O\left(n^2r^2\theta_{min}^{-2}f^{-1}\right)$. However, just as in Section 3, we can do the attack in parallel, modifying all n banknote qubits slightly (using n individual probe qubits) and submitting them for verification. In this case, we save a factor of n, and will require only $O\left(nr^2\theta_{min}^{-2}f^{-1}\right)$ rounds of verification.

A.2 The bomb testing attacks does not work for general unknown states

Let us relax our assumptions and look at what happens if θ_{min} is not bounded from below, i.e. if the states that we are receiving are completely unknown, or perhaps from a continuous range of states. We run into a new problem, because what we do can no longer always be interpreted as a weak measurement. Let us illustrate this point.

When the set of possible money states is dense, there is no minimal angle between two possible states. Similarly, when we do not know what the state could be, we can only guess, and attempt to do a controlled reflection about a random axis. Thus, the counterfeiter has to choose some θ_{min} since the other parameters N, δ explicitly depend on θ_{min}. This introduces a fourth option to the list above: $0 < \theta < \theta_{min}$.

Before explaining *why*, we first explain *what* happens. Using Mathematica, we analytically computed the outcome of the protocol from Section 1, as a function of $\theta\sqrt{N}$, in the limit $N \to \infty$, which is depicted in Fig. A.1.

It can be seen that we do not get caught with high probability only if either $\theta \ll \frac{1}{\sqrt{N}}$ (in which case, the outcome of the measurement of the probe qubit is $|1\rangle$) or if $\theta \gg \frac{1}{\sqrt{N}}$ (in which case, the outcome of the measurement of the probe qubit is $|0\rangle$). If θ is in the order of $\frac{1}{\sqrt{N}}$ there is a constant probability of getting caught by the bank. If we choose a random state $|\beta\rangle$ that we want to identify, the probability to land in the safe region $\theta \ll \frac{1}{\sqrt{N}}$ (which allows us to identify the money state as close to $|\beta\rangle$), is much smaller than the probability of landing in the dangerous zone $\theta = \Theta(\frac{1}{\sqrt{N}})$. We need many calls to the protocol, therefore, we cannot tolerate a constant probability of being caught. To conclude, independent of how N is chosen, we will most likely get caught by the bank before getting an approximation to a single qubit of the bank note (and we need to succeed in that n times). Therefore, the BT attack fails when the set of possible states is infinite or unknown.

The basic building block of the protocol we analyzed in Section 1 is a rotation followed by an interaction, followed by validation. We want the validation to succeed with high probability. Prior to the controlled rotation (the interaction stage) the money state is the original $|\alpha\rangle$. What happens after the interaction? There are two cases:

1. The money state is an eigenstate of R (i.e. $\theta \in \{0, \pi/2\}$), so the unknown system stays factorized and validation always succeeds.

2. $\theta \in (0, \pi/2)$. We can be assured of success (not being detected) when the interaction stage is "weak enough".

The interaction is "weak enough" for our purposes whenever we do not disturb the system much, so that we remain undetected even as we test it many times. We want to safely run $O(N)$ rounds, so it is enough if the probability to be caught in any particular round is $o\left(N^{-1}\right)$; clearly $O\left(\delta^2\right) = O\left(N^{-2}\right)$ is enough.

The critical case appears sneakily, when the state $|\alpha\rangle$ is almost, but not quite a $+1$ eigenstate of R, which means $q \lesssim 1$ in (A.6). In this case, R does not disturb the unknown system much, and the transformation T (A.7) is approximately

$$T_{q\lesssim 1} \approx \begin{bmatrix} \cos\delta & -\sin\delta \\ \sin\delta & \cos\delta \end{bmatrix}, \tag{A.14}$$

resulting in a rotation of the state $|\varphi_k\rangle$ to $|\varphi_k + \delta\rangle$. In this way, in a constant fraction of N of rounds, our probe rotates to a state where its overlap with $|1\rangle$ becomes a constant, so we actually start to significantly apply $R_{q\lesssim 1}$ to the system. Although $R_{q\lesssim 1}$ does not disturb the system much, if $q = 1 - O\left(\sqrt{\delta}\right)$, the *cumulative* probability of detection can grow as we repeat many rounds; this is our ultimate doom. The the probe/system would become $\cos\varphi_k|0\rangle|\psi\rangle + \cos\varphi_k\cos\theta|1\rangle|\psi\rangle + \cos\varphi_k\sin\theta|\psi^\perp\rangle$, so the probability of failure in this round is $\cos^2\varphi_k\sin^2\theta = O\left(\delta\right)$ for $\Theta\left(\varphi_k\right) = 1$. Note that we can nicely avoid this problem when using the protective measurement method from Section 4.1, ensuring a sufficient upper bound on the disturbance of the unknown system in any round.

Let us finish by saying why this is not a problem for q well bounded away from 1, in which case our interaction with the system can be always "weak enough". At the start of a particular round, the probe qubit is in the state $|\varphi_k\rangle = \cos\varphi_k|0\rangle + \sin\varphi_k|1\rangle$. We then rotate it by δ to $|\varphi_k + \delta\rangle$. Next, we perform the controlled reflection, run through a validation step and when we succeed, we end up in the unnormalized state (A.5). Normalizing it, we find the angle φ_k from the beginning of the round changed to φ_{k+1}, which for positive[f] q is less than $\varphi_k + \delta$. It turns out that this iterative process has an upper bound. Moreover, for reasonable q this upper bound is on the order of δ, as can be learned from (A.8). Thus, the probability of failure in any round is upper bounded by $O\left(\delta^2\right)$, and the measurement

[f] For $q \leq 0$, we can bound $|\varphi_k| \leq \delta$, so the measurement is always weak.

we do is weak enough for our purposes. We will not be detected in N rounds with probability more than on the order of $\delta = O\left(\frac{1}{N}\right)$.