# Improving the security of device-independent weak coin flipping protocols

Atul Singh Arora[1,2], Jamie Sikora[3], and Thomas Van Himbeeck[4,5]

[1] *Institute for Quantum Information and Matter, California Institute of Technology, USA*
[2] *Department of Computing and Mathematical Sciences, California Institute of Technology, USA*
[3] *Virginia Polytechnic Institute and State University, USA*
[4] *University of Toronto, Canada*
[5] *Institute of Quantum Computing, University of Waterloo, Canada*

October 2021

**Abstract**

Weak coin flipping is the cryptographic task where Alice and Bob remotely flip a coin but want opposite outcomes. This work studies this task in the device-independent regime where Alice and Bob neither trust each other, nor their quantum devices. The best protocol was devised ten years ago by Silman, Chailloux, Aharon, Kerenidis, Pironio, and Massar with bias $\epsilon \leq 0.33664$, where the bias is a commonly adopted security measure for coin flipping protocols. This work presents some techniques to lower the bias of device-independent weak coin flipping protocols, namely self-testing and abort-phobic compositions. By applying these techniques to the SCAKPM '11 protocol above, we are able to lower the bias to $\epsilon \approx 0.31486$ and assuming a continuity conjecture, we can suppress it further to $\epsilon \approx 0.29104$. In our analysis, we show how to harness the rigidity bounds for the GHZ game in our setting and examine the continuity of an optimisation problem to bound the bias, which may be of independent interest.

## Contents

# 1 Introduction

Coin-flipping is the two-party cryptographic primitive where two parties, henceforth called Alice and Bob, wish to flip a coin, but where, to make things interesting, they do not trust each other. This primitive was introduced by Blum [Blu83] who also introduced the first (classical) protocol. In this work, we concentrate on *weak* coin flipping (WCF) protocols where Alice and Bob desire opposite outcomes. Since then, a series of quantum protocols were introduced with successively improved security. Mochon, in his tour de force, finally settled the question about the limits of the security in the quantum regime by proving the *existence* of quantum protocols with security approaching the ideal limit [Moc07]. This was followed by a flurry of results which achieved diverse cryptographic functionality assuming WCF as a black-box, such as strong coin flipping [CK09], bit commitment [CK11], a variant of oblivious transfer [CGS13], leader election [Gan09] and dice rolling [AS], establishing the importance of WCF in the quantum setting. Returning to Mochon, his work was quite technical and based on the notion of *point games*, a concept introduced by Kitaev. Interestingly, his work was never published—only a preprint was available. Subsequently, a sequence of works have continued the study of point games. In particular, the proof of existence was eventually simplified and peer reviewed [ACG$^+$14] and explicit protocols were reported after more than a decade of Mocohn's work [ARW19, ARV].[1] Yet, we note that all of this work is in the *device-dependent* setting where *Alice and Bob trust their quantum devices*. Very little is known in the *device-independent (DI)* setting where a cheating player is allowed to control an honest player's quantum devices, opening up a plethora of new cheating strategies that were not considered in the previously mentioned references.

We introduce some basic concepts to facilitate further discussion. The prefix *weak* in weak coin flipping refers to the situation where Alice and Bob desire opposite outcomes of the coin. (We have occasion to discuss *strong* coin flipping protocols, where Alice or Bob could try to bias the coin towards either outcome, but it is not the focus of this work.) When designing weak coin flipping protocols, the security goals are as follows.

*Correctness for honest parties:* If Alice and Bob are honest, then they share the same outcome of a protocol $c \in \{0, 1\}$, and $c$ is generated uniformly at random by the protocol.

*Soundness against cheating Alice:* If Bob is honest, then a dishonest (i.e., cheating) Alice cannot force the outcome $c = 0$.

*Soundness against cheating Bob:* If Alice is honest, then a dishonest (i.e., cheating) Bob cannot force the outcome $c = 1$.

The commonly adopted goal of two-party protocol design is to assume perfect correctness and then minimize the effects of a cheating party, i.e., to make it as sound as possible. This way, if no parties cheats, then the protocol at least does what it is meant to still. With this in mind, we need a means to quantify the effects of a cheating party. It is often convenient to have a single measure to determine if one protocol is better than another. For this purpose, we use *cheating probabilities* (denoted $p_B^*$ and $p_A^*$) and *bias* (denoted $\epsilon$), defined as follows.

$p_A^*$: The maximum probability with which a dishonest Alice can force an honest Bob to accept the outcome $c = 0$.

$p_B^*$: The maximum probability with which a dishonest Bob can force an honest Alice to accept the outcome $c = 1$.

$\epsilon$: The maximum amount with which a dishonest party can bias the probability of the outcome away from uniform. Explicitly, $\epsilon = \max\{p_A^*, p_B^*\} - 1/2$.

[1] Interestingly, Miller [Mil20] used techniques from Mochon's proof to show that protocols approaching the ideal limit must have an exponentially increasing number of messages. It is an open question to find how small the bias can be made before any such protocol becomes impractical.

These definitions are not complete in the sense that we have not yet specified what a cheating Alice or a cheating Bob are allowed to do, or of their capabilities. In this work, we study *information theoretic security*—Alice and Bob are only bounded by the laws of quantum mechanics. For example, they are not bounded by polynomial-time quantum computations. In addition to this, we study the security in the *device-independent* regime where we assume Alice and Bob have complete control over the quantum devices when they decide to "cheat".

When studying device-independent (DI) protocols, one should first consider whether or not secure classical protocols are known (since these are not affected by the DI assumption). It was proved that every classical WCF protocol[2] has bias $\epsilon = 1/2$, which is the worst possible value (see [Kit03, HW11]). Thus, it makes sense to study quantum WCF protocols in the DI setting, especially if one with bias $\epsilon < 1/2$ can be found. Indeed, Silman, Chailloux, Aharon, Kerenidis, Pironio, and Massar presented a protocol (see Protocol S) in [SCA$^+$11] with $p_A^* = \cos^2(\pi/8) \approx 0.853$ and $p_B^* = 3/4$. We briefly discuss this protocol because we build on this result but defer the details. Their protocol begins with Alice possessing two *boxes*—physical devices that accept classical inputs and yield classical outputs—and Bob possessing one box which are together supposed to contain the GHZ state and measurements.[3] As the protocol proceeds, they, in addition to exchanging classical information, operate these boxes and exchange them.[4] As is, Protocol S has bias $\epsilon \approx 0.353$ but in [SCA$^+$11], Protocol S is composed many times to lower the bias to $\epsilon \leq 0.33664$.

In this work, we provide two techniques for lowering the bias of weak coin flipping protocols and apply them to Protocol S, mentioned above.

# 2 Results

We state the main result of our work.

**Theorem 1.** *There exist device-independent weak coin flipping protocols with bias, $\epsilon$, approaching* 0.0.31486. *Assuming Conjecture 22 holds, the bias can be lowered to approach $\epsilon \approx 0.29104$.*

We now discuss the key ideas that go into the proof of our main theorem, above. Protocol S was, in fact, a strong coin flipping protocol and we begin by turning it into a weak coin flipping protocol—Protocol W—in a routine manner. Again, we defer the explicit description of the protocol and informally describe the basic idea: since weak coin flipping has the notion of a "winner" (if $c = 0$ Alice wins and if $c = 1$ Bob wins) we have the party who does not win, conduct an additional test.

Our first technique is to add a pre-processing step to Protocol W which *self-tests* the boxes shared by Alice and Bob at the start of the protocol. Our second technique is to compose and analyse the resulting protocols in a new way,[5] which we call *abort-phobic* composition.

## 2.1 First technique: Self-testing

In the original Protocol S and it's WCF variant, Protocol W, a cheating party may control what measurement is performed in the boxes of the other party and how the state of the boxes is correlated to its own quantum memory. This is more general than *device-dependent* protocols, where for instance, the measurements are known to the honest player. However, we employ the concept of self-testing to stop Bob (or Alice) from applying such a strategy. Intuitively, self testing is a powerful property which allows one to, just from certain input-output behaviours of given devices (satisfying minimal assumptions), conclude uniquely which quantum states and measurements constitute the devices (up to relabelling). The GHZ state which was used in Protocols S and W can be self-tested. Clearly, this property has the potential to improve their security.[6]

We define two variants of Protocol W: Protocol P, where Alice self-tests Bob before executing Protocol W, and Protocol Q, where Bob self-tests Alice instead. Skipping the details, the basic construction is almost trivial. Alice and

[2] also holds for strong coin flipping

[3] A GHZ state is a non-local quantum state; we review this in Section 3.

[4] Any protocol described using boxes is readily converted into one where Alice and Bob communicate over an insecure quantum channel; see Section A

[5] The composition in [SCA$^+$11] may also be seen as "abort-phobic" but their analysis doesn't rely on the "abort" probability; their bound essentially neglects the abort event.

[6] In [SCA$^+$11], it was noted that self-testing doesn't help improve the security of Protocol S. Alternatively stated, Protocol S has the curious property that its device dependent variant has the same security as it (the device dependent variant).

Bob start with $n$ triples of boxes and, for instance when Alice self-tests, Alice asks Bob to send all but one randomly selected triple and tests if the GHZ test passes for these. If so, the remaining triple is used for the actual protocol. If $n$ is chosen large enough, then this forces a dishonest Bob to not tamper with the boxes too much, as suggested above. Indeed, this step already allows us to reduce the cheating probabilities.

**Lemma 2** (Informal. See Lemma 10 for a formal statement)**.** *For Protocol P, i.e. where Alice self tests Bob, the cheating probabilities, in the limit of large n, are*

$$p_A^* = \cos^2(\pi/8) \approx 0.85355 \quad \text{and} \quad p_B^* \approx 0.6667. \tag{1}$$

For comparison, recall that for Protocol S (it turns out, also for Protocol W), $p_A^* = \cos^2(\pi/8)$ and $p_B^* = 3/4$. We prove this lemma in two stages. In the *first* stage (see Section 5), we assume perfect self-testing—the self-testing step results in exactly specifying (up to a relabelling) the state and measurements governing Alice's boxes. This may be seen as taking $n \longrightarrow \infty$ in the self-testing step. It is known that for device-dependent protocols, where Alice and Bob trust their devices, the cheating probabilities can be cast as values of semidefinite programs (SDPs) [Kit03, Moc07]. Perfect self-testing allows us to, therefore, express Bob's cheating probabilities as an SDP. Its numerical evaluation yields the quoted value. Analysis for Alice's cheating probability is unchanged from Protocol W. In the *second* stage (see Section 6), we take $n$ to be finite and show that for large $n$, the analysis converges to that of the first stage. While we carry out this analysis for Protocol P, we state an analogous result for Protocol Q assuming a continuity conjecture (Conjecture 22) holds. We give some more details before proceeding to the second technique.

**Self-testing in a cryptographic setting.** Self-testing results can be made *robust*, i.e. in particular, if the success probability in a GHZ test is close to unity, then the states and measurements can be shown to be close to GHZ states and measurements (up to a relabelling), in say trace distance. Robust self-testing results are, however, usually stated in terms of expected success probabilities in tests. This implicitly assumes that multiple identical boxes are available.[7] In our cryptographic setting, such an assumption is unwarranted. Hence, we estimate this expected success probability, by measuring $n - 1$ boxes. This estimate requires, to the best of our knowledge, a novel analysis which we discuss in Section 6.3. This analysis holds for both Protocol P and Protocol Q.

**Continuity argument.** We argued above that under the perfect self-testing assumption, the analysis can be cast as an SDP. However, with a finite number, $n$, of tests, the resulting optimisation problem is neither an SDP nor defined over variables with bounded dimensions. To make a rigorous security guarantee, one needs a continuity result which ensures that in the large $n$ limit, the optimisation problem converges to the aforementioned SDP. Indeed, for Protocol P, we establish this continuity result. We conjecture that an analogous statement holds for Protocol Q.

## 2.2 Second technique: abort-phobic composition

It can happen, that for a given WCF protocol, $p_B^* \neq p_A^*$, in which case we say the protocol is *polarised*. As we saw earlier, it is known (e.g. [SCA+11]) that composing a polarised protocol with itself (or other protocols) can effectively reduce the bias. Our second improvement is a modified way of composing protocols, when there is a positive probability that the honest player catches the cheating player. Let us start by recalling the standard way of composing protocols.

**Standard composition.** For a protocol with cheating probabilities $p_B^*$ and $p_A^*$, we say that it has polarity towards Alice when it satisfies $p_A^* > p_B^*$. Similarly, we say that it has polarity towards Bob when $p_B^* > p_A^*$. Given a polarized protocol $\mathcal{R}$, we may switch the roles of Alice and Bob since the definition of coin-flipping is symmetric. To make the polarity explicit, we define $\mathcal{R}_A$ to be the version of the protocol with $p_A^* > p_B^*$ and $\mathcal{R}_B$ to be the version with $p_B^* > p_A^*$. With this in mind, we can now define a simple composition.

**Protocol 3** (Winner-gets-polarity composition)**.** *Alice and Bob agree on a protocol $\mathcal{R}$.*

1. *Alice and Bob perform protocol $\mathcal{R}$.*

[7] When this iid assumption is dropped, it is usually in the context of extracting randomness or key from the observed statistics, i.e. all the devices are used/measured. In our setting, we need to leave one device unused in the pre-processing step.

2. *If Alice wins, she polarizes the second protocol towards herself, i.e., they now use the protocol $\mathcal{R}_A$ to determine the final outcome.*
3. *If Bob wins, he polarizes the second protocol towards himself, i.e., they now use the protocol $\mathcal{R}_B$ to determine the final outcome.*

The standard composition above is a sensible way to balance the cheating probabilities of a protocol. For instance, if $\mathcal{R}$ has cheating probabilities $p_A^*$ and $p_B^*$ with $p_A^* > p_B^*$, then the composition gets to decide "who gets to be Alice" in the second run. We can easily compute Alice's cheating probability in the composition as

$$(p_A^*)^2 + (1 - p_A^*)p_B^* < p_A^* \tag{2}$$

and Bob's as

$$p_B^* p_A^* + (1 - p_B^*)p_B^* < p_A^*. \tag{3}$$

This does indeed reduce the bias since the maximum cheating probability is now smaller.

**Abort-phobic composition.** The "traditional" way of considering WCF protocols is to view them as only having two outcomes "Alice wins" (when $c = 0$) or "Bob wins" ($c = 1$). This is because Alice can declare herself the winner if she catches Bob cheating. Similarly, Bob can declare himself the winner if he catches Alice cheating.[8] This is completely fine when we consider "one-shot" versions of these protocols, but we lose something when we compose them. For instance, in the simple composition used in Protocol 3, Bob should not really accept to continue onto the second protocol if he catches Alice cheating in the first. That is, if he knows Alice cheated, he can declare himself the winner of the entire protocol. In other words, the cheating probabilities (2) and (3) may get reduced even further. For purposes of this discussion, suppose Bob adopts a cheating strategy which has a probability $v_B$ of him winning ($c = 1$), a probability $v_A$ of him losing ($c = 0$), and a probability $v_\perp$ of Alice catching him cheating. Then his cheating probability in the (abort-phobic) version of the simple composition is now

$$v_B \cdot p_A^* + v_A \cdot p_B^* + v_\perp \cdot 0. \tag{4}$$

This quantity may be a strict improvement if $v_\perp > 0$ when $v_B = p_B^*$.

The concept of abort-phobic composition is simple. Alice and Bob keep using WCF protocols and the winner (at that round) gets to choose the polarity of the subsequent protocol. However, if either party *ever aborts*, then it is game over and the cheating player loses *the entire composite protocol*.

One may think it is tricky to analyse abort-phobic compositions, but we may do this one step at time. To this end, we introduce the concept of *cheat vectors*.

**Definition 4** ($\mathbb{C}_A, \mathbb{C}_B$; Alice and Bob's cheat vectors)**.** Given a protocol $\mathcal{R}$, we say that $(v_A, v_B, v_\perp)$ is a cheat vector for (dishonest) Bob if there exists a cheating strategy where,

$v_B$ is the probability with which Alice accepts the outcome $c = 1$,
$v_A$ is the probability with which Alice accepts the outcome $c = 0$,
$v_\perp$ is the probability with which Alice aborts.

We denote the set of cheat vectors for (dishonest) Bob by $\mathbb{C}_B(\mathcal{R})$. Cheat vectors for (dishonest) Alice and $\mathbb{C}_A(\mathcal{R})$ are analogously defined keeping the notation $v_A$ for her winning, $v_B$ for her losing, and $v_\perp$ for Bob aborting.

In this work, we show how to capture cheat vectors as the feasible region of a semidefinite program, from which we can optimize

$$v_B \cdot p_A^* + v_A \cdot p_B^* + v_\perp \cdot 0. \tag{5}$$

For this to work, we assume we have $p_A^*$ and $p_B^*$ for the protocol that comes in the second round. A simplifying observation is that once we solve for the optimal cheating probabilities in the abort-phobic composition in this way, we can then fix those probabilities and compose again. In other words, we are recursively composing the abort-phobic composition, from the *bottom up*.

By using abort-phobic compositions with Protocol P (where Alice self-tests) one obtains protocols which converge onto a bias of $\epsilon \approx 0.31486$ proving the first part of the main result. For the second part, we place Protocol P at the bottom, and Protocol Q (where Bob self-tests) on higher layers, to obtain protocols whose bias approaches $\epsilon \approx 0.29104$ assuming the continuity conjecture for Protocol Q.

[8] In doing so, we implicitly assume that the protocol has perfect correctness—when both players are honest, the probability of abort is zero.

### 2.3 Applications

The concept of polarity extends beyond finding WCF protocols and, as such, the "winner-gets-polarity" concept allows for WCF to be used in other compositions. Indeed, we can use it to balance the cheating probabilities in *any* polarized protocol for any symmetric two-party cryptographic task for which such notions can be properly defined.

For instance, many *strong* coin-flipping protocols can be thought of as polarized. For an example, Protocol S is indeed a polarized strong coin-flipping protocol. Thus, by balancing the cheating probabilities of that protocol using our DI WCF protocol, we get the following corollary.

**Corollary 5.** *Suppose Conjecture 22 holds. Then, there exist DI strong coin-flipping protocols where no party can cheat with probability greater than* 0.33192.

To contrast, for [SCA+11], the bound on cheating probabilities was 0.336637. There are likely more examples of protocols which can be balanced in a DI way using this idea.

## 3 New protocols using self-testing | First Technique

We start by recalling the DI strong coin flipping protocol introduced in [SCA+11], Protocol S, and introduce its weak coin flipping variant Protocol W. We then describe the new Protocols P and Q, where Alice and Bob respectively perform the self-testing step. We also give more formal security guarantees associated with these. Their proofs constitute Sections 5 and 6.

**Notation** For notational clarity, we often use single calligraphic symbols $\mathcal{S}$, $\mathcal{W}$, $\mathcal{P}$ and $\mathcal{Q}$ to refer to these protocols. When we say, for instance, consider a triple of boxes $\square^A, \square^B, \square^C$, we mean that there is a tripartite quantum state and local measurements associated with these boxes. The input to the box selects the measurement setting and the output is the measurement outcome as governed by quantum theory (see Definition 24). When we speak of Alice and Bob exchanging boxes, we understand that the description of these states and measurement settings are sent over a (possibly insecure) quantum communication channel (see Definitions 25 and 27 in Section A).

We recall the GHZ test before starting our main discussion as this is at the heart of these protocols.

**Definition 6.** Suppose we are given a triple of boxes, $\square^A$, $\square^B$ and $\square^C$, which accept binary inputs $a, b, c \in \{0, 1\}$ and produces binary output $x, y, z \in \{0, 1\}$ respectively. The boxes pass the GHZ test if $a \oplus b \oplus c = xyz \oplus 1$, given the inputs satisfy $x \oplus y \oplus z = 1$.

It is known that no classical triple of boxes can pass the GHZ test with certainty but quantum boxes can.

*Claim* 7. Quantum boxes pass the GHZ test with certainty (even if they cannot communicate), for the state $|\psi\rangle_{ABC} = \frac{|000\rangle_{ABC}+|111\rangle_{ABC}}{\sqrt{2}}$, and measurement[9] $\frac{\sigma_x+\mathbb{I}}{2}$ for input 0 and $\frac{\sigma_y+\mathbb{I}}{2}$ for input 1 (in the notation introduced earlier, $M^A_{0|0} = |+\rangle\langle+|, M^A_{1|0} = |-\rangle\langle-|$ and so on, where $|\pm\rangle = \frac{|0\rangle\pm|1\rangle}{\sqrt{2}}$).

The proof is easier to see in the case where the outcomes are ±1; it follows from the observations that $\sigma_y \otimes \sigma_y \otimes \sigma_y |\psi\rangle = -|\psi\rangle$, $\sigma_x \otimes \sigma_x \otimes \sigma_x |\psi\rangle = |\psi\rangle$ and the anti-commutation of $\sigma_x$ and $\sigma_y$ matrices, i.e. $\sigma_x\sigma_y + \sigma_y\sigma_x = 0$.

In fact a stronger property holds. If a triple of boxes passes the GHZ test with certainty, it can be shown that up to a local isometry, the state and measurements are as in Claim 7 above. While this is, manifestly, a highly idealised setting and we later, in Section 6, see how it works in practice.

**Lemma 8.** *Let $a, b, c, x, y, z \in \{0, 1\}$. Consider a triple of quantum boxes, specified by projectors $\{M^A_{a|x}, M^B_{b|y}, M^C_{c|z}\}$ acting on finite dimensional Hilbert spaces $\mathcal{H}^A, \mathcal{H}^B$ and $\mathcal{H}^C$, and $|\psi\rangle \in \mathcal{H}^A \otimes \mathcal{H}^B \otimes \mathcal{H}^C =: \mathcal{H}^{ABC}$. If the triple pass the GHZ test with probability $1-\epsilon$ (for $1 > \epsilon > 0$), then there exists a local isometry,*

$$\Phi = \Phi^A \otimes \Phi^B \otimes \Phi^C : \mathcal{H}^{ABC} \to \mathcal{H}^{ABC} \otimes \mathbb{C}^{2\times 3}$$

*and a decreasing function of $\epsilon$, $f(\epsilon)$ such that*

$$\|\Phi(|\psi\rangle) - |\chi\rangle \otimes |\mathrm{junk}\rangle\| \le f(\epsilon), \tag{6}$$

$$\left\|\Phi\left(M^D_{d|t}|\psi\rangle\right) - \Pi^D_{d|t}|\mathrm{GHZ}\rangle \otimes |\mathrm{junk}\rangle\right\| \le f(\epsilon) \quad \forall D \in \{A, B, C\}, \text{ and } d, t \in \{0, 1\} \tag{7}$$

[9] we added the identity so that the eigenvalues associated become 0, 1 instead of −1, 1.

*where* $|\mathrm{GHZ}\rangle = \frac{|000\rangle+|111\rangle}{\sqrt{2}} \in \mathbb{C}^{2\times 3}$, $|\mathrm{junk}\rangle \in \mathcal{H}^{ABC}$ *is some arbitrary state and* $\{\Pi^A_{a|x}, \Pi^B_{b|y}, \Pi^C_{c|z}\}$ *are projectors corresponding to* $\sigma_x$ *on the first, second and third qubit of* $|\mathrm{GHZ}\rangle$ *respectively, for* $x = 0$ *and corresponding to* $\sigma_y$ *for* $x = 1$*, as in Claim 7.*

*Proof.* Proofs of robust self-testing for GHZ can be found in [MS13] and [McK14]. □

## 3.1 Original protocols

Protocol S is defined as follows.

**Protocol S** A DI-SCF protocol with $p^*_A = \cos^2 \pi/8$ and $p^*_B = 3/4$ ([SCA+11])

Alice has one box and Bob has two boxes. Each box takes one binary input and gives one binary output and are designed to play the optimal GHZ game strategy. (Who creates and distributes the boxes is not important in the DI setting.)

1. Alice chooses a uniformly random input to her box $x \in_R \{0, 1\}$ and obtains the outcome $a$. She chooses another uniformly random bit $r \in_R \{0, 1\}$ and computes $s = a \oplus (x \cdot r)$. She sends $s$ to Bob.
2. Bob chooses a uniformly random bit $g \in_R \{0, 1\}$ and sends it to Alice. (We may think of $g$ as Bob's "guess" for the value of $x$.)
3. Alice sends $x$ to Bob. They both compute the output $c = x \oplus g$. (This is the outcome of the protocol if no-one abort.)
4. Bob tests Alice

Test 1: Alice sends $a$ to Bob. Bob sees if $s = a$ or $s = a \oplus x$. If this is not the case, he aborts.

Test 2: Bob chooses $y, z \in_R \{0, 1\}$ uniformly at random such that $x \oplus y \oplus z = 1$ and then performs a GHZ using $x, y, z$ as the inputs and $a, b, c$ as the output from the three boxes. He aborts if this test fails.

5. If Bob does not abort, they both accept the value of $c$ as the outcome of the protocol.

We now discuss the correctness and soundness of Protocol S. From Claim 7, it is clear that when both players follow the protocol using GHZ boxes (Definition 6), Bob never aborts and they win with equal probabilities. As for the security, [SCA+11] proved the following.

**Lemma 9** (Security of SCF)**.** *[SCA+11] Let* $\mathcal{S}$ *denote the protocol corresponding to Protocol S. Then, the success probability of cheating Bob,*[10] $p^*_B(\mathcal{S}) \leq \frac{3}{4}$ *and that of cheating Alice,* $p^*_A(\mathcal{S}) \leq \cos^2(\pi/8)$*.*

*Further, both bounds are saturated by a quantum strategy which uses a GHZ state and the honest player measures along the* $\sigma_x/\sigma_y$ *basis corresponding to input* $0/1$ *into the box. Cheating Alice measures along* $\sigma_{\hat{n}}$ *for* $\hat{n} = \frac{1}{\sqrt{2}}(\hat{x}+\hat{y})$ *while cheating Bob measures his first box along* $\sigma_x$ *and second along* $\sigma_y$*.*

Note that both players can cheat maximally assuming they share a GHZ state and the honest player measures along the associated basis. This is why it was asserted that even though the cheating player could potentially tamper with the boxes before handing them to the honest player, exploiting this freedom does not offer any advantage to the cheating player.

Clearly, if we take Protocol S as is and treat it like a weak coin flipping protocol, this conclusion would continue to hold. As motivated in the introduction, we consider a minor, yet crucial, modification to Protocol S. Observe that in Protocol S only Bob performs the test round, while in weak coin flipping there is a notion of Alice winning and Bob winning which may be leveraged. More precisely, if $x \oplus g = 0$, i.e. the outcome corresponding to "Alice wins", we can imagine that Bob continues to perform the test to ensure (at least to some extend) that Alice did not cheat. However, if $x \oplus g = 1$, i.e. the outcome corresponding to "Bob wins", we can require Alice to now complete the GHZ test to ensure that Bob did not cheat. Since we analyse this protocol in detail, we state it as Protocol W, somewhat redundantly below. We have emphasised the changes compared to Protocol S in italics.

[10] For SCF, $P^*_B$ is max{Pr[Bob can force Alice to output 1], Pr[Bob can force Alice to output 0]}; $P^*_A$ is analogously defined.

**Protocol W** Weak Coin Flipping version of Protocol S (Italics indicate the differences with Protocol S)

Alice has one box and Bob has two boxes. Each box takes one binary input and gives one binary output and are designed to play the optimal GHZ game strategy. (Who creates and distributes the boxes is not important in the DI setting.)

1. Alice chooses a uniformly random input to her box $x \in_R \{0, 1\}$ and obtains the outcome $a$. She chooses another uniformly random bit $r \in_R \{0, 1\}$ and computes $s = a \oplus (x \cdot r)$. She sends $s$ to Bob.

2. Bob chooses a uniformly random bit $g \in_R \{0, 1\}$ and sends it to Alice. (We may think of $g$ as Bob's "guess" for the value of $x$.)

3. Alice sends $x$ to Bob. They both compute the output $c = x \oplus g$. This is the outcome of the protocol assuming neither Alice nor Bob aborts.

4. Test rounds:

   (a) *If* $x \oplus g = 0$, Bob tests Alice

   Test 1: Alice sends $a$ to Bob. Bob sees if $s = a$ or $s = a \oplus x$. If this is not the case, he aborts.

   Test 2: Bob chooses $y, z \in_R \{0, 1\}$ uniformly at random such that $x \oplus y \oplus z = 1$ and then performs a GHZ using $x, y, z$ as the inputs and $a, b, c$ as the output from the three boxes. He aborts if this test fails.

   (b) *If* $x \oplus g = 1$, *Alice tests Bob*

   *Test 3: Alice chooses* $y, z \in_R \{0, 1\}$ *uniformly at random such that* $x \oplus y \oplus z = 1$ *and sends them to Bob. Bob inputs* $y, z$ *into his boxes, obtains and sends* $b, c$ *to Alice. Alice tests if* $x, y, z$ *as inputs and* $a, b, c$ *as outputs, satisfy the GHZ test. She aborts if this test fails.*

5. If Alice and Bob do not abort, they both accept the value of $c$ as the outcome of the protocol.

While it is not surprising that $p_A^*(\mathcal{W}) = p_A^*(\mathcal{P}) = \cos^2(\pi/8)$, it turns out that $p_B^*(\mathcal{W}) = p_B^*(\mathcal{P}) = 3/4$, despite the additional test that Alice performs i.e. $P_B^*$ for Protocol W is not lowered. Yet, this is not quite a setback—one can show that the best cheating strategy now deviates from the GHZ state and measurements for the honest player, suggesting that a cheating player *does* benefit from tampering with the boxes. Consequently, adding a self-testing step before initiating Protocol W, may potentially improve its security and as we shall see in the following subsections, it indeed does.

A remark about the limitation of self-testing in this setting. We note that no self-testing scheme can be concocted which simultaneously self-tests Alice and Bob's boxes. More precisely, no such procedure can ensure that Alice and Bob share a GHZ state (Alice one part, Bob the other two, for instance) because this would mean perfect (or near perfect) SCF is possible which, recall, is forbidden even in the device dependent case.[11]

## 3.2 Alice self-tests | Protocol $\mathcal{P}$

We begin with explicitly stating Protocol P—where Alice self-tests her boxes before initiating Protocol W. In the honest implementation, the triple of boxes used in Protocol P are characterised by the GHZ setup (see Claim 7).

To state the associated security condition, we need the definition of cheat vectors from the introduction (see Definition 4).

**Lemma 10.** *Let* $\mathcal{P}$ *denote Protocol* P*. Then Alice's cheating probability* $p_A^*(\mathcal{P}) \leq \cos^2(\pi/8) \approx 0.852$*. Further, let* $c_0, c_1, c_\perp \in \mathbb{R}$*, and* $\mathbb{C}_B(\mathcal{P})$ *be the set of cheat vectors for Bob (see Figure 1a). Then, as* $N \to \infty$*, the solution to the optimisation problem* $\max(c_0\alpha + c_1\beta + c_\perp\gamma)$ *over* $\mathbb{C}_B(\mathcal{P})$ *approaches that of an SDP over variables of constant dimension (wrt* $N$*). In particular, i.e. for* $c_0 = c_\perp = 0$ *and* $c_1 = 1$*,* $p_B^*(\mathcal{P}) \approx 0.667$ *(in the limit).*

We defer the proof to Section 5.1 and Section 6.1. As remarked in the introduction, the value for $p_B^*(\mathcal{P})$ is lower than $p_B^*(\mathcal{W})$ and was obtained by numerically solving the corresponding SDP while the analysis for cheating Alice

[11] More precisely, Kitaev [Kit03] showed that for any SCF protocol, $\epsilon \geq \frac{1}{\sqrt{2}} - \frac{1}{2}$.

**Protocol P** Alice self-tests

Alice starts with $n$ boxes, indexed from $1_1$ to $1_n$. Bob starts with $2n$ boxes, the first half indexed by $2_1$ to $2_n$ and the last half indexed by $3_1$ to $3_n$. The triple of boxes $(1_i, 2_i, 3_i)$ is meant to play the optimal GHZ game strategy.

1. Alice selects a uniformly random index $i \in \{1, \ldots, n\}$ and asks Bob to send her all the boxes *except* those indexed by $2_i$ and $3_i$.

2. Alice performs $n-1$ GHZ tests using the $n-1$ triples of boxes she has, making sure there is no communication between any of them, e.g. by shielding the boxes (in the relativistic settings, coin flipping is possible).

3. Alice aborts if *any* of the GHZ tests fail. Otherwise, she announces to Bob that they can use the remaining boxes for Protocol W.

is the same as that of the original protocol. The fact that optimising linear functions in Bob's cheat vectors is an SDP becomes useful in Section 4 when we compose these protocols.

### 3.3 Bob self-tests | Protocol $Q$

We analogously define Protocol Q—where Bob self-tests his boxes before initiating Protocol W.

**Protocol Q** Bob self-tests

Alice starts with $n$ boxes, indexed from $1_1$ to $1_n$. Bob starts with $2n$ boxes, the first half indexed by $2_1$ to $2_n$ and the last half indexed by $3_1$ to $3_n$. The triple of boxes $(1_i, 2_i, 3_i)$ is meant to play the optimal GHZ game strategy.

1. Bob selects a uniformly random index $i \in \{1, \ldots, n\}$ and asks Alice to send him all the boxes *except* those indexed by $1_i$.

2. Bob performs $n-1$ GHZ tests using the $n-1$ triples of boxes he has, making sure there is no communication between any of them.

3. Bob aborts if *any* of the GHZ tests fail. Otherwise, he announces to Alice that they can use the remaining boxes for Protocol W.

Consider Protocol W and Protocol S. Suppose Bob is honest while Alice is malicious, and that at step 3, she sends an $x$ s.t. $x \oplus g = 0$. Under these conditions, observe that Bob's actions are identical in both Protocol W and Protocol S. Since it is already known from Lemma 9 that Alice doesn't gain anything from tampering with Bob's boxes, the same conclusion holds for Protocol W. Thus, we do not expect any improvement in Bob's security, viz. $p_A^*(Q) = p_A^*(\mathcal{W})$ given that Protocol Q only ensures Alice doesn't tamper with Bob's boxes. It is also immediate that $p_B^*(Q) = p_B^*(\mathcal{W})$. This means that we do not see any advantage of self-testing at this stage but, analogously to Protocol P, optimisation of linear functions of Alice's cheat vectors now becomes an SDP and we reap the benefits of this simplification in the next section.

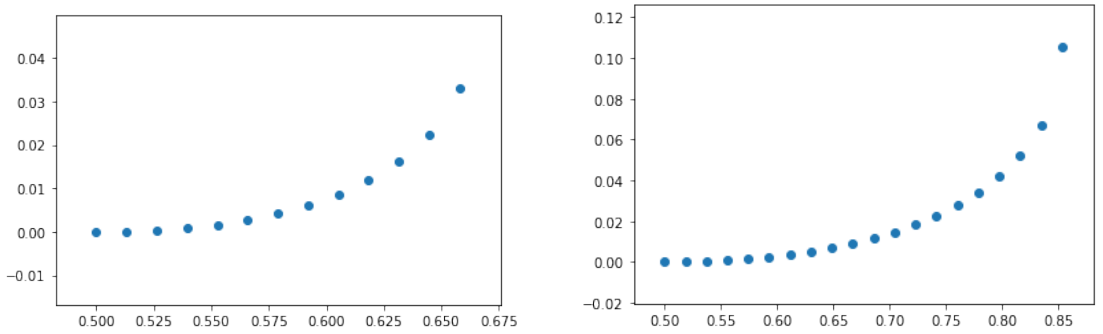
**Lemma 11.** *Let $Q$ denote Protocol Q. Then, Alice's cheating probability, $p_A^*(Q) \leq 3/4$ and Bob's cheating probability, $p_B^*(Q) \leq \cos^2(\pi/8)$ (which are the same as those in Lemma 9). Further, let $c_0, c_1, c_\perp \in \mathbb{R}$, and $\mathbb{C}_A(Q)$ be the set of cheat vectors for Alice (see Figure 1b) and suppose Conjecture 22 holds. Then, as $N \longrightarrow \infty$, the solution to the optimisation problem $\max(c_0\alpha + c_1\beta + c_\perp\gamma)$ over $(\alpha, \beta, \gamma) \in \mathbb{C}_A(Q)$ approaches that of an SDP of constant dimension (wrt $N$).*

The proof is again deferred; see Section 5.2 and Section 6.2.

## 4 Compositions | Second Technique

In this section we assume that $N$ is large enough so that the cheating probabilities/vectors of the individual protocols Protocol P and Protocol Q are close to their asymptotic values.

Central to the discussion in this section, will be the notion of *polarity* introduced in Section 2.2 and our results will apply to polarised protocols. Note that $\mathcal{W}, \mathcal{P}$, and $Q$ are all polarised. We begin by recalling that in Section 2.2,



(a) Cheat vectors $\mathbb{C}_B(\mathcal{P})$. The $x$-axis represents $v_B$ and the $y$-axis represents the smallest $v_\perp$, given $v_B$.

(b) Cheat vectors $\mathbb{C}_A(\mathcal{Q})$. The $x$-axis represents $v_B$ and the $y$-axis represents the smallest $v_\perp$, given $v_B$.

Figure 1: Cheat vectors for Protocol P and Protocol Q. Observe how compared to Figure 1a, the abort probabilities in Figure 1b are higher making it more suitable for abort-phobic compositions.

again, we had introduced a "standard composition"—the simplest implementation of the "winner gets polarity" idea. Here, we restate this composition with more precision and introduce the notation we use for the more involved cases.

## 4.1 Composition

The following simply formalises the following—execute protocol $\mathcal{X}$ to determine who gets to choose the polarity of protocol $\mathcal{Y}$. We use $C$ with two parameters, as in $C(\mathcal{X},\mathcal{Y})$, to denote a single composition described above. We use $C(\mathcal{X})$ to denote repeated compositions of $\mathcal{X}$.

**Definition 12** ($C(\cdot,\cdot)$ and $C(\cdot)$)**.** Given two polarised WCF protocols, $\mathcal{X}$ and $\mathcal{Y}$, let $\mathcal{X}_A,\mathcal{X}_B$ and $\mathcal{Y}_A,\mathcal{Y}_B$ be their polarisations (see Section 2.2). Define $C(\mathcal{X},\mathcal{Y})$ as follows:

1. Alice and Bob execute $\mathcal{X}_A$ and obtain outcome $X \in \{A,B,\perp\}$.
2. (a) If $X=A$, execute $\mathcal{Y}_A$ and obtain outcome $Y \in \{A,B,\perp\}$, else
   (b) if $X=B$, execute $\mathcal{Y}_B$ and obtain outcome $Y \in \{A,B,\perp\}$, and finally
   (c) if $X=\perp$, set $Y=\perp$.

   Output $Y$.

Let $\mathcal{Z}^{i+1} := C(\mathcal{X},\mathcal{Z}^i)$ for $i \geq 1$, and $\mathcal{Z}^1 := \mathcal{X}$. Then, formally, define $C(\mathcal{X}) := \lim_{i\to\infty}\mathcal{Z}^i$.[12]

The study of such composed protocols is simplified by assuming that in an honest run, neither player outputs $\perp$ (abort), i.e. they either output $A$ or $B$. We take a moment to explain this.

Consider any protocol $\mathcal{R}$ where, when both players are honest, the probability of abort is zero. The protocols we have described so far, satisfy this property, so long as we assume that honest players can prepare perfect GHZ boxes. Such protocols are readily converted into protocols where an honest player never outputs abort.

For instance, suppose that in the execution of the aforementioned protocol $\mathcal{R}$ (with no-honest-abort), Alice behaves honestly but Bob is malicious. Suppose after interacting with Bob, Alice reaches the conclusion that she must abort. Since she knows that if Bob was honest, the outcome abort could not have arisen, she concludes that Bob is cheating and declares herself the winner, i.e. she outputs $A$. Similarly, when Bob is honest and after the interaction, reaches the outcome abort, he knows Alice cheated and can therefore declare himself the winner, i.e. output $B$.

Whenever we modify a protocol so that an honest Alice (Bob) replaces the outcome abort with Alice (Bob) winning, we say Alice (Bob) is *lenient*. This is motivated by the fact that when we compose protocols, if Alice can conclude that Bob is cheating, and she still outputs $A$ instead of aborting, she is giving Bob further opportunity to cheat—she is being lenient.

[12] This is just to facilitate notation. This way the cheating probabilities $p_A^*$ and $p_B^*$ converge and numerically this only takes a few compositions to reach in our case.

**Definition 13** ($\mathcal{R}$ with lenient players)**.** Suppose $\mathcal{R}$ is a WCF protocol such that when both players are honest, the probability of outcome abort, $\perp$, is zero. Then by "*$\mathcal{R}$ with lenient Alice (Bob)*", which we denote by $\mathcal{R}^{L\perp}$ ($\mathcal{R}^{\perp L}$), we mean that Alice (Bob) follows $\mathcal{R}$ except that the outcome $\perp$ replaced with $A$ ($B$). Finally, by "*lenient $\mathcal{R}$*", which we denote by $\mathcal{R}^{LL}$, we mean $\mathcal{R}$ with lenient Alice and Bob.

For clarity and conciseness, we define $C^{LL}$ to be compositions with lenient variants of the given protocols. We work out some examples of such protocols and analyse their security in the following section. These can be improved by considering $C^{L\perp}$ and $C^{\perp L}$—compositions where only one player is lenient. We discuss those afterwards.

**Definition 14** ($C^{LL}$, $C^{\perp L}$ and $C^{L\perp}$)**.** Suppose a WCF protocol $\mathcal{X}$ can be transformed into its *lenient* variants (see Definition 13). Then define

$$\begin{aligned} C^{LL}(\mathcal{X},\mathcal{Y}) &:= C(\mathcal{X}^{LL},\mathcal{Y}), \\ C^{\perp L}(\mathcal{X},\mathcal{Y}) &:= C(\mathcal{X}^{\perp L},\mathcal{Y}), \quad \text{and} \\ C^{L\perp}(\mathcal{X},\mathcal{Y}) &:= C(\mathcal{X}^{L\perp},\mathcal{Y}). \end{aligned}$$

In words, $C^{LL}$ is referred to as a *standard* composition, while $C^{\perp L}$ and $C^{L\perp}$ are referred to as *abort-phobic* compositions. The single argument versions are analogously defined, i.e. $C^{LL}(\mathcal{X}) := C(\mathcal{X}^{LL})$, $C^{L\perp}(\mathcal{X}) := C(\mathcal{X}^{L\perp})$ and $C^{\perp L}(\mathcal{X}^{\perp L})$.

## 4.2 Standard Composition | $C^{LL}$

We begin with the simplest case, standard composition, $C^{LL}$. Let us take an example. Let $\mathcal{P}$ denote Protocol P and recall (see Lemma 10)

$$\begin{aligned} p_A^*(\mathcal{P}_A) &=: \alpha \approx 0.852, \\ p_B^*(\mathcal{P}_A) &=: \beta \approx 0.667. \end{aligned}$$

Note that therefore $p_A^*(\mathcal{P}_B) = \beta$ and $p_B^*(\mathcal{P}_B) = \alpha$. Further, let $\mathcal{P}' := C^{LL}(\mathcal{P},\mathcal{P})$, i.e. Alice and Bob (who are both lenient) first execute $\mathcal{P}_A$ and if the outcome is $A$, they execute $\mathcal{P}_A$, while if the outcome is $B$, they execute $\mathcal{P}_B$. This is illustrated in Figure 2 where note that the event abort doesn't appear due to the leniency assumption. This allows us to evaluate the cheating probabilities for the resulting protocol as

$$\begin{aligned} p_A^*(\mathcal{P}') &= \alpha\alpha + (1-\alpha)\beta =: \alpha^{(1)}, \quad \text{and} \\ p_B^*(\mathcal{P}') &= \beta\alpha + (1-\beta)\beta =: \beta^{(1)}. \end{aligned} \tag{8}$$

To see this, consider Equation (8). Alice knows that if she wins the first round, her probability of winning is $\alpha > \beta$. She knows that in the first round, she can force the outcome $A$ with probability $\alpha$. From leniency, she knows that Bob would output $B$ with the remaining probability.[13]

A side remark: one consequence of this simplified analysis is that[14] $\alpha^{(1)} > \beta^{(1)}$. Intuitively, it means that polarity is preserved by the composition procedure. Proceeding similarly, i.e. defining $\mathcal{P}'' := C^{LL}(\mathcal{P},\mathcal{P}')$, and repeating $k+1$ times overall, one obtains[15]

$$\begin{aligned} \alpha^{(k+1)} &= \alpha\alpha^{(k)} + (1-\alpha)\beta^{(k)} \\ \beta^{(k+1)} &= \beta\alpha^{(k)} + (1-\beta)\beta^{(k)}. \end{aligned}$$

In the limit of $k \longrightarrow \infty$, one obtains

$$p_A^*(C^{LL}(\mathcal{P})) = p_B^*(C^{LL}(\mathcal{P})) = \lim_{k\to\infty} \alpha^{(k)} = \lim_{k\to\infty} \beta^{(k)} \approx 0.8199.$$

[13] Without leniency, this probability could have been shared between the outcomes $\perp$ (abort) and $B$. Consequently, only upper bounds could be obtained on $p_A^*(\mathcal{P}')$ and $p_B^*(\mathcal{P}')$ using only $\alpha$ and $\beta$ as security guarantees for $\mathcal{P}_A$. Upper bounds, however, would not be enough to determine the polarity of $\mathcal{P}'$ and stymie an unambiguous repetition of the composition procedure (at least as it is defined). One could nevertheless compose by hoping that the upper bounds can be used to accurately represent the polarity. Regardless, this would still yield a protocol and the same calculation would yield correct bounds but the composition itself might be sub-optimal.

[14] $\alpha^{(1)} - \beta^{(1)} = (\alpha-\beta)\alpha - (\alpha-\beta)\beta = (\alpha-\beta)^2 > 0$

[15] Again, note that $\alpha^{(k+1)} - \beta^{(k+1)} = (\alpha^{(k)} - \beta^{(k)})(\alpha-\beta) > 0$.

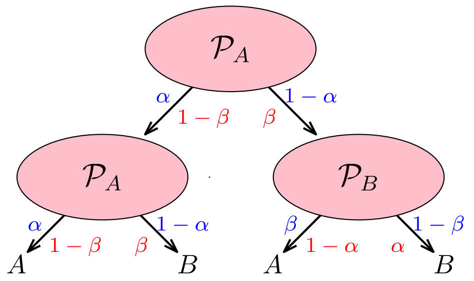


Figure 2: Standard composition of weak coin flipping protocols. Subprotocols only have two outcomes depending on the coin flip. Labels indicate probabilities of outcomes for cheating Alice (blue) and cheating Bob (red)

Proceeding similarly, one obtains for $X \in \{A, B\}$ and $\mathcal{X} \in \{\mathcal{I}, \mathcal{Q}\}$,

$$p_X^*(C^{LL}(\mathcal{X})) \approx 0.836$$

We thus have the following.

**Theorem 15.** *Let $\mathcal{W}$ and $\mathcal{Q}$ denote Protocol W and Protocol Q respectively. Suppose $X \in \{A, B\}$ and $\mathcal{X} \in \{\mathcal{W}, \mathcal{Q}\}$. Then, for a large enough $N$, one has $p_X^*(C^{LL}(\mathcal{X})) \leq 0.836$ and, assuming Conjecture 22 holds, one has $p_X^*(C^{LL}(\mathcal{P})) \leq 0.8199$.*

### 4.3 Abort Phobic Compositions | $C^{L\perp}, C^{\perp L}$

We now look at the case of abort phobic compositions, $C^{L\perp}$ and $C^{\perp L}$. We work through essentially the same example as above and see what changes in this setting. As usual, let $\mathcal{P}$ denote Protocol P and recall that as before

$$p_A^*(\mathcal{P}_A) =: \alpha \approx 0.852,$$
$$p_B^*(\mathcal{P}_A) =: \beta \approx 0.667.$$

In addition, we know from Lemma 10 that cheat vectors for Bob, $(v_A, v_B, v_\perp) \in \mathbb{C}_B(\mathcal{P}_A)$ admit a nice characterisation courtesy of the self testing step. Let $\mathcal{P}' := C^{\perp L}(\mathcal{P}, \mathcal{P})$, i.e. Alice and Bob execute $\mathcal{P}_A$ and if the outcome is $A$, they execute $\mathcal{P}_A$ while if the outcome is $B$, they execute $\mathcal{P}_B$. Bob is assumed to be lenient so an honest Bob never outputs abort, $\perp$. However, an honest Alice can output abort, $\perp$ so we keep that output in the illustration, Figure 3. Our goal is to find $p_A^*(\mathcal{P}')$ and $p_B^*(\mathcal{P}')$. The former is the same as before because Bob is lenient:

$$p_A^*(\mathcal{P}') = \alpha \cdot \alpha + (1-\alpha) \cdot \beta.$$

Clearly, $p_B^*(\mathcal{P}') \leq \beta\alpha + (1-\beta)\beta$ but this bound may not be tight because $(1-\beta)$ is the combined probability of Alice aborting and Alice outputting $A$. However, we can use cheat vectors to obtain

$$p_B^*(\mathcal{P}') = \max_{(v_A, v_B, v_\perp) \in \mathbb{C}_B} v_B\alpha + v_A\beta$$

which is an SDP one can solve numerically. Unlike the previous case, the polarity of the resulting protocol, $\mathcal{P}'$, might have flipped (compared to the polarity of $\mathcal{P}$).

Repeating this procedure, one can consider $\mathcal{P}'' := C^{\perp L}(\mathcal{P}, \mathcal{P}')$ and obtain $p_A^*(\mathcal{P}'')$ directly as illustrated above and numerically solve for $p_B^*(\mathcal{P}'')$ using the cheat vectors. Numerically, we found that ten-fifteen repetitions caused the cheating probabilities to converge to approximately 0.81459. We saw that the abort probabilities associated with $\mathcal{P}$ were quite small and therefore one could hope that $\mathcal{Q}$ (which denotes Protocol Q) fares better. Proceeding analogously and considering $\mathcal{Q}' := C^{L\perp}(\mathcal{Q}, \mathcal{Q})$, $\mathcal{Q}'' := C^{L\perp}(\mathcal{Q}, \mathcal{Q}')$, etc., the cheating probabilities converge to approximately 0.822655.

**Theorem 16.** *Let $\mathcal{P}$ and $\mathcal{Q}$ denote Protocol P and Protocol Q respectively. Suppose $X \in \{A, B\}$. Then for a large enough $N$, one has*

$$p_X^*(C^{\perp L}(\mathcal{P})) \leq 0.81459$$

*and assuming Conjecture 22 one has*

$$p_X^*(C^{L\perp}(\mathcal{Q})) \leq 0.822655.$$

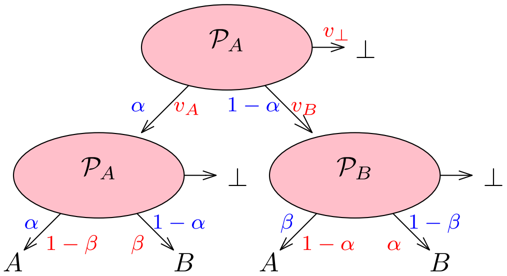


Figure 3: Abort phobic compositing for weak coin flipping protocols. Subprotocols have three possible outcomes including an abort symbol. Aborting in any subprotocol directly leads to aborting the whole protocol. Labels indicate probabilities of outcomes for cheating Alice (blue) and cheating Bob (red). In the security analysis of cheating Bob, we need to optimise over the cheat vectors $(v_A, v_B, v_\perp) \in \mathbb{C}_B$.

While by itself $\mathcal{Q}$ doesn't seem to help, one can suppress the bias further, by noting that at the very last step, only the cheating probabilities $p_A^*(\mathcal{Q})$ and $p_B^*(\mathcal{Q})$ played a role (i.e. the fact that the cheating vectors $\mathbb{C}_A$ for $\mathcal{Q}$ had an SDP characterisation was not used). Further, we know that $p_A^*(\mathcal{P}) = p_A^*(\mathcal{Q})$ but $p_B^*(\mathcal{P}) < p_B^*(\mathcal{Q})$, i.e. using $\mathcal{P}$ at the very last step will result in a strictly better protocol.

**Theorem 17.** *Let $X \in \{A, B\}$,*

$$\mathcal{Z}^1 := C^{L\perp}(\mathcal{Q}, \mathcal{P}), \quad \text{and}$$
$$\mathcal{Z}^{i+1} := C^{L\perp}(\mathcal{Q}, \mathcal{Z}^i) \quad i > 1.$$

*Then for large enough $N$,*

$$\lim_{i\to\infty} p_X^*(\mathcal{Z}^i) \leq 0.791044.$$

# 5 Security Proof | Asymptotic limit

The goal of this section is to analyse the security of Protocol P (Protocol Q) in the simplest setting—assuming that Alice's box (Bob's boxes) indeed correspond to the GHZ state and measurements. Readers familiar with analysis of device dependent protocols in the distrustful cryptography setting may prefer to skip to Section 6 where we analyse Protocol P in the general case (and give a partial analysis for Protocol Q).

To be more precise, in this section, we prove the security under the following assumption.

**Assumption 18.** *In protocol $\mathcal{P}$ ($\mathcal{Q}$), Alice (Bob) does not perform the box verification step and instead it is assumed that her box is (his boxes are) taken from a triple of boxes which win the GHZ game with certainty.*

Lemma 8 asserts that when the winning probability is exactly one (i.e. $\epsilon = 0$ in the lemma), the states and measurements are the same as the GHZ state and $\sigma_x, \sigma_y$ measurements, up to local isometries and this allows us to use semi definite programming.

## 5.1 SDP when Alice self-tests | $\mathcal{P}$

*Asymptotic proof of Lemma 10.* We prove Lemma 10 under Assumption 18. We begin by making two observations.

First, note that in the protocol, if Alice applies an isometry on her box *after* she has inputted $x$, obtained the outcome $a$ (and has noted it somewhere), the security of the resulting protocol is unchanged because the rest of the protocol only depends on $x$ and $a$, and Alice's isometry only amounts to relabelling of the post measurement state. This freedom allows us to simplify the analysis.

Second, in the analysis, we cannot model Alice's random choice, say for $x$, as a mixed state because Bob can always hold a purification and thus know $x$. Therefore, we model the randomness using pure states and measure them in the end.

Notation: Other than $PQR$, all other registers store qubits.

We proceed step by step.

1. We can model (justified below) Alice's act of inputting a random $x$ and obtaining an outcome $a$ from her box through the state
$$|\Psi_0\rangle := \frac{1}{2}\sum_{x,a\in\{0,1\}} |xa\rangle_{XA}\,|\Phi(x,a)\rangle_{IJ} \tag{9}$$
where $X$ represents the random input and $A$ the output. Here, $|\Phi(x,a)\rangle_{IJ}$ are Bell states (see Equation (11)) and the registers $IJ$ are held by Bob. Alice's act of choosing $r$ at random, computing $s = a \oplus x.r$ is modelled as
$$|\Psi_1\rangle := \frac{1}{2\sqrt{2}}\sum_{x,a,r\in\{0,1\}} |xa\rangle_{XA}\,|\Phi(x,a)\rangle_{IJ}\,|r\rangle_R\,|a\oplus x.r\rangle_S\,. \tag{10}$$
Finally, Alice's act of sending $s$ is modelled as Alice starting with the state
$$\mathrm{tr}_{IJS}\left[|\Psi_1\rangle\langle\Psi_1|\right] \in XAR.$$

**Justification for starting with $|\Psi_0\rangle$.**
To see why we start with the state $|\Psi_0\rangle$, model Alice's choice of $x$ as $|+\rangle_X$, suppose her measurement result is stored in $|0\rangle_A$, the state of the boxes before measurement is $|\psi\rangle_{PQR}$ and Alice holds $P$, i.e.
$$\left|\Psi_0'\right\rangle := |+\rangle_X\,|0\rangle_A\,|\psi\rangle_{PQR}\,.$$
Let $\{M^P_{a|x}\}$ be the measurement operators corresponding to Alice's box. The measurement process is unitarily modelled as
$$\left|\Psi_1'\right\rangle := U_{\text{measure}}\left|\Psi_0'\right\rangle = \frac{1}{\sqrt{2}}\sum_{x,a\in\{0,1\}} |x\rangle_X\,|a\rangle_A\,M^P_{a|x}\,|\psi\rangle_{PQR}$$
where
$$U_{\text{measure}} = \sum_{x\in\{0,1\}} |x\rangle\langle x|_X \otimes \left[\mathbb{I}_A\otimes M^P_{0|x} + X_X\otimes M^P_{1|x}\right]\otimes \mathbb{I}_{QR}.$$
Now we harness the freedom of applying an isometry to the post measured state (as observed above). We choose the local isometry in Lemma 8. Without loss of generality, we can assume that Bob had already applied his part of the isometry before sending the boxes (because he can always reverse it when it is his turn). We thus have,
$$\begin{aligned}\left|\Psi_2'\right\rangle := \Phi_{PQR}\left|\Psi_1'\right\rangle &= \frac{1}{\sqrt{2}}\sum_{x,a\in\{0,1\}} |x\rangle_X\,|a\rangle_A\,\Pi^H_{x|a}\,|\mathrm{GHZ}\rangle_{HIJ}\otimes|\mathrm{junk}\rangle_{PQR}\\ &= \frac{1}{2}\sum_{x,a\in\{0,1\}} |x\rangle_X\,|a\rangle_A\,U^H(x,a)\,|0\rangle_H\,|\Phi(x,a)\rangle_{IJ}\otimes|\mathrm{junk}\rangle_{PQR}\end{aligned}$$
where
$$|\Phi(x,a)\rangle_{IJ} = \frac{|00\rangle + (-1)^a(i)^x\,|11\rangle}{\sqrt{2}} \tag{11}$$
and $U^H(x,a)\,|0\rangle_H$ is $\frac{|0\rangle+(-1)^a(i)^x|1\rangle}{\sqrt{2}}$. Since the state of register $H$ is completely determined by registers $X$ and $A$, we can drop it from the analysis without loss of generality. Finally, since $|\mathrm{junk}\rangle_{PQR}$ is completely tensored out, we can drop it too without affecting the security. Formally, we can assume that Alice gives Bob the register $P$ at this point.

2. Bob sending $g$ is modelled by introducing $\rho_2 \in XARG$ satisfying $\mathrm{tr}_{IJS}\left[|\Psi_1\rangle\langle\Psi_1|\right] = \mathrm{tr}_G(\rho_2)$.

3. At this point, either $x\oplus g$ is zero, in which case Alice's output is fixed or $x\oplus g$ is one, and in that case Bob will already know $x$ because he knows $g$ (he sent it) and Alice will proceed to testing Bob. Formally, therefore, we needn't do anything at this step.

4. Assuming $x\oplus g = 1$, Alice sends $y, z$ to Bob such that $x\oplus y\oplus z = 1$. However, since Bob already knows $x$, he can deduce $z$ from $y$. We thus only need to model Alice sending $y$ and Bob responding with $d = b\oplus c$ (because Alice will only use $b\oplus c$ to test the GHZ game, so it suffices for Bob to send $d$). This amounts to introducing $\rho_3 \in XARGYD$ satisfying $\rho_2\otimes\frac{\mathbb{I}_Y}{2} = \mathrm{tr}_D(\rho_3)$.

5. Since we postponed the measurements to the end, we add this last step. Alice now measures $\rho_3$ to determine $x \oplus g$ and if it is one, whether the GHZ test passed. Let

$$\Pi_i := \sum_{x,y\in\{0,1\}:x\oplus g=i} |xg\rangle\langle xg|_{XG} \otimes \mathbb{I}_{ARYD}, \tag{12}$$

$$\Pi^{\text{GHZ}} := \sum_{\substack{x,y\in\{0,1\},\\ a,d\in\{0,1\}:a\oplus d\oplus 1=xy\cdot(1\oplus x\oplus y)}} |xyad\rangle\langle xyad|_{XYAD} \otimes \mathbb{I}_{RG}. \tag{13}$$

Then, we can write the cheat vector for Alice, i.e. the tuple of probabilities that Alice outputs 0, 1 and abort (see Definition 4), as

$$(\alpha, \beta, \gamma) = (\text{tr}(\Pi_0\rho_3), \text{tr}(\Pi_1\Pi^{\text{GHZ}}\rho_3), \text{tr}(\Pi_1\bar{\Pi}^{\text{GHZ}}\rho_3))$$

where $\bar{\Pi} := \mathbb{I} - \Pi$.

To summarise, the final SDP is as follows: let $|\Psi_1\rangle \in XAIJRS$ be as given in Equation (10), $\rho_2 \in XARG$ and $\rho_3 \in XARGYD$

$$\max \quad \text{tr}([c_0\Pi_0 + \Pi_1(c_1\Pi^{\text{GHZ}} + c_\perp\bar{\Pi}^{\text{GHZ}})]\rho_3)$$

subject to

$$\text{tr}_{IJS}\left[|\Psi_1\rangle\langle\Psi_1|\right] = \text{tr}_G(\rho_2)$$

$$\rho_2 \otimes \frac{\mathbb{I}_Y}{2} = \text{tr}_D(\rho_3)$$

where the projectors are defined in Equation (13). □

## 5.2 SDP when Bob self-tests | $\mathcal{Q}$

*Asymptotic proof of Lemma 11.* Denote by $\mathcal{S}$ the protocol corresponding to Protocol S.

It is evident that $p_B^*(\mathcal{Q}) \leq p_B^*(\mathcal{S})$ because compared to $\mathcal{S}$, in $\mathcal{Q}$ Alice performs an extra test. However, it is not hard to see that the inequality is saturated, i.e. $p_B^*(\mathcal{Q}) = p_B^*(\mathcal{S})$. Consider ... (TODO: recall/re-construct the cheating strategy for Bob that lets him win with the same 3/4 probability).

From Lemma 9, it is also clear that $p_A^*(\mathcal{Q}) = p_A^*(\mathcal{S})$ because the only difference between Bob's actions in $\mathcal{Q}$ and $\mathcal{S}$ is that Bob self-tests to ensure his boxes are indeed GHZ. However, the optimal cheating strategy for $\mathcal{S}$ can be implemented using GHZ boxes.

This establishes the first part of the lemma. For the second part, i.e. establishing that optimising $c_0\alpha + c_1\beta + c_\perp\gamma$ over $(\alpha, \beta, \gamma) \in \mathbb{C}_A$ is an SDP, we proceed as follows. Suppose Assumption 18 holds. Then we can assume that Bob starts with the state

$$\rho_0 := \text{tr}_H(|\text{GHZ}\rangle\langle\text{GHZ}|_{HIJ}) \tag{14}$$

and the effect of measuring the two boxes can be represented by the application of projectors of pauli operators $X$ and $Z$.

The justification is similar to that given in the former proof. Suppose Bob holds registers $QR$ of $|\psi\rangle_{PQR}$ which is the combined state of the three boxes. Suppose his measurement operators are $\{M^Q_{b|y}, M^R_{c|z}\}$. Then using the isometry in Lemma 8, Bob can relabel his state (and without loss of generality, we can suppose Alice also relabels according to the aforementioned isometry) to get $\Phi_{PQR}|\psi\rangle_{PQR} = |\text{GHZ}\rangle_{HIJ} \otimes |\text{junk}\rangle_{PQR}$. Further, since $\Phi_{PQR}(M^Q_{b|y} \otimes M^R_{c|z}|\psi\rangle_{PQR}) = \Pi^I_{b|y}\Pi^J_{c|z}|\text{GHZ}\rangle_{HIJ} \otimes |\text{junk}\rangle_{PQR}$ Bob's act of measurement, in the new labelling, corresponds to simply measuring the GHZ state in the appropriate Pauli basis.

1. Bob receiving $s$ from Alice is modelled by introducing $\rho_1 \in SIJ$ satisfying $\text{tr}_S(\rho_1) = \rho_0$.

2. Bob sending $g \in_R \{0, 1\}$ can be seen as appending a mixed state: $\rho_1 \otimes \frac{1}{2}\mathbb{I}_G$.

3. Alice sending $x$ (and $a$) can be modelled as introducing $\rho_2 \in AXSIJG$ satisfying $\text{tr}_A(\rho_2) = \rho_1 \otimes \frac{\mathbb{I}_G}{2}$.

4. To model the GHZ test, introduce a register $Y$ in the state $\frac{|0\rangle_Y + |1\rangle_Y}{\sqrt{2}}$. Recall that to perform the GHZ test, we need $x \oplus y \oplus z = 1$ i.e. $z = 1 \oplus y \oplus x$. Further introduce registers $B$ and $C$ to hold the measurement results, define

$$U := \sum_{y,x\in\{0,1\}} |y\rangle\langle y|_Y\, |x\rangle\langle x|_X \otimes (\mathbb{I}_B \otimes \Pi^I_{0|y} + X_B \otimes \Pi^I_{1|y}) \otimes (\mathbb{I}_C \otimes \Pi^J_{0|(1\oplus y\oplus x)} + X_C \otimes \Pi^J_{1|(1\oplus y\oplus x)}) \otimes \mathbb{I}_{ASG}. \quad (15)$$

By construction, $\rho_3 := U\,(|+\rangle\langle+|_Y \otimes |00\rangle\langle 00|_{BC} \otimes \rho_2)\,U^\dagger \in YBCAXSIJG$ models the measurement process.

5. Since we postponed the measurements to the end, we add this step. Define

$$\Pi_i := \sum_{x,g\in\{0,1\}:x\oplus g=i} |xg\rangle\langle xg|_{XG} \otimes \mathbb{I}_{YABSIJ}$$

to determine who won. Define

$$\Pi^{\text{sTest}} := \sum_{s,a,x\in\{0,1\}:s=a\vee s=a\oplus x} |sax\rangle\langle sax|_{SAX} \otimes \mathbb{I}_{GYBCIJ}$$

to model the first test, i.e. $s$ should either be $a$ or $a\oplus x$. Define

$$\Pi^{\text{GHZ}} := \sum_{\substack{x,y\in\{0,1\},\\ a,b,c\in\{0,1\}:a\oplus b\oplus c\oplus 1=xy\cdot(1\oplus x\oplus y)}} |xyabc\rangle\langle xyabc|_{XYABC} \otimes \mathbb{I}_{GSIJ}$$

to model the GHZ test. Let

$$\Pi^{\text{Test}} := \Pi^{\text{GHZ}}\Pi^{\text{sTest}}, \quad \bar{\Pi}^{\text{Test}} := \mathbb{I} - \Pi^{\text{Test}}. \quad (16)$$

One can then write the cheat vector for Bob, i.e. the tuple of probabilities that Bob outputs 0, 1 and abort (see Definition 4), as

$$(\alpha,\beta,\gamma) = (\text{tr}(\Pi_0\Pi^{\text{Test}}\rho_3), \text{tr}(\Pi_1\rho_3), \text{tr}(\Pi_0\bar{\Pi}^{\text{Test}}\rho_3)).$$

To summarise, the final SDP is as follows: let $\rho_0 \in IJ$ be as defined in Equation (14), $\rho_1 \in SIJ$ and $\rho_2 \in AXSIJG$. Then,

$$\eta := \max \quad \text{tr}\left([\Pi_0(c_0\Pi^{\text{Test}} + c_\perp\bar{\Pi}^{\text{Test}}) + c_1\Pi_1]U\,(|{+}00\rangle\langle{+}00|_{YBC} \otimes \rho_2)\,U^\dagger\right) \quad (17)$$

subject to

$$\text{tr}_S(\rho_1) = \rho_0$$
$$\text{tr}_A(\rho_2) = \frac{1}{2}\rho_1 \otimes \mathbb{I}_G$$

where $U$ is as defined in Equation (15) and the projectors as in Equation (16).

□

# 6 Security Proofs | Finite $N$

In this section, we prove Lemma 10 (and state Conjecture 22 for Lemma 11), accounting for the fact that $N$ is finite. We first show that if the GHZ winning probability is $1-\epsilon$ then the optimisation problem over the cheat vectors converges to the asymptotic SDP discussed in the previous section, as $\epsilon \to 0$. We then show how this winning probability can be estimated by testing $N-1$ triples of boxes and bound the probability of reaching an erroneous conclusion.

## 6.1 Analysis when Alice self-tests | $\mathcal{P}$

**Lemma 19.** *Let $c_0, c_1, c_\perp$ be non-negative. Denote by $v_0, v_1, v_\perp$ the probability that Alice outputs $0, 1, \perp$ respectively when Protocol P is executed against some cheating strategy of Bob. Let $\epsilon(N)$ denote the winning probability of the GHZ boxes deduced by Alice (via Proposition 23) when she tests $N-1$ boxes. Then, the objective $c_0v_0 + c_1v_1 + c_\perp v_\perp$ is bounded by the value, $\eta_\epsilon$, of the following optimisation program.*

$$\eta_\epsilon := \text{maximize: } tr([c_0\Pi_0 + \Pi_1(c_1\Pi^{\text{GHZ}} + c_\perp\bar{\Pi}^{\text{GHZ}})]\rho_3) \tag{18}$$

$$\text{subject to: } \|y_{a,x} - \Pi^H_{a|x}\left|GHZ\right\rangle_{HIJ} \otimes \left|junk\right\rangle_{H'I'J'}\|_{tr} \leq \epsilon \tag{19}$$

$$\left|\phi\right\rangle = \frac{1}{2}\sum_{a,x,r}\left|x,a,r\right\rangle\left|a \oplus x.r\right\rangle \otimes y_{a,x} \tag{20}$$

$$\rho_1 = \left|\phi\right\rangle\left\langle\phi\right| \tag{21}$$

$$tr_{S,I,I',J,J'}(\rho_1) = tr_G(\rho_2) \tag{22}$$

$$\rho_2 \otimes \frac{\mathbb{I}_Y}{2} = tr_D(\rho_3) \tag{23}$$

$$\|y_{a,x}\|_2 \leq 1. \tag{24}$$

*Proof.* The state at the beginning of the protocol can be written as

$$\left|\phi_0\right\rangle = \frac{1}{\sqrt{2}}\sum_x\left|x\right\rangle\left|\psi\right\rangle. \tag{25}$$

When Alice measures, she creates the state

$$\left|\phi_1\right\rangle = \frac{1}{\sqrt{2}}\sum_{x,a}\left|x,a\right\rangle M_{a|x}\left|\psi\right\rangle. \tag{26}$$

Then we have

$$\Phi\left(\left|\phi\right\rangle\right) = \Phi\left(\frac{1}{2}\sum_{x,a,r}\left|x,a,r\right\rangle\left|a \oplus x.r\right\rangle \otimes M_{a|x}\left|\psi\right\rangle\right) = \frac{1}{2}\sum_{x,a,r}\left|x,a,r\right\rangle\left|a \oplus x.r\right\rangle \otimes \Phi(M_{a|x}\left|\psi\right\rangle). \tag{27}$$

From self-testing, we know that

$$\|\Phi(M_{a|x}\left|\psi\right\rangle) - \Pi^H_{a|x}\left|GHZ\right\rangle_{HIJ} \otimes \left|junk\right\rangle_{H'I'J'}\|_{tr} \leq \epsilon. \tag{28}$$

Since $M_{a|x}$ and $\left|\psi\right\rangle$ is under Bob's control, we can relax the problem to allowing Bob to control *the entire vector* $\Phi(M_{a|x}\left|\psi\right\rangle)$, and we just call this vector $y_{a,x}$ (which is subnormalized). We now have the optimization problem

$$\text{maximize: } \text{tr}([c_0\Pi_0 + \Pi_1(c_1\Pi^{\text{GHZ}} + c_\perp\bar{\Pi}^{\text{GHZ}})]\rho_3) \tag{29}$$

$$\text{subject to: } \|y_{a,x} - \Pi^H_{a|x}\left|GHZ\right\rangle_{HIJ} \otimes \left|junk\right\rangle_{H'I'J'}\|_{tr} \leq \epsilon \tag{30}$$

$$\Phi(\left|\phi\right\rangle) = \frac{1}{2}\sum_{a,x,r}\left|x,a,r\right\rangle\left|a \oplus x.r\right\rangle \otimes y_{a,x} \tag{31}$$

$$\rho_1 = \left|\phi\right\rangle\left\langle\phi\right| \tag{32}$$

$$\text{tr}_{S,I,I',J,J'}(\rho_1) = \text{tr}_G(\rho_2) \tag{33}$$

$$\rho_2 \otimes \frac{\mathbb{I}_Y}{2} = \text{tr}_D(\rho_3) \tag{34}$$

$$\|y_{a,x}\|_2 \leq 1. \tag{35}$$

Now, by multiplying each of the constraints involving $\rho_1$, $\rho_2$, and $\rho_3$ on each side by $\Phi$, we preserve feasibility. This proves the result. □

**Lemma 20** (Continuity). *We have* $\lim_{\epsilon\downarrow 0}\eta_\epsilon = \eta$*, where* $\eta$ *is defined as the following SDP:*

$$\begin{aligned}\text{maximize: } & tr([c_0\Pi_0 + \Pi_1(c_1\Pi^{\mathrm{GHZ}} + c_\perp\bar{\Pi}^{\mathrm{GHZ}})]\rho_3) && (36)\\ \text{subject to: } & \rho_1 = |\phi\rangle\langle\phi| && (37)\\ & tr_{S,I,I',J,J'}(\rho_1) = tr_G(\rho_2) && (38)\\ & \rho_2 \otimes \frac{\mathbb{I}_Y}{2} = tr_D(\rho_3), && (39)\end{aligned}$$

*where* $|\phi\rangle = \frac{1}{2}\sum_{a,x,r}|x,a,r\rangle|a\oplus x.r\rangle \otimes \Pi^H_{a|x}|GHZ\rangle_{HIJ} \otimes |junk\rangle_{H'I'J'}$.

*Proof.* First, note that $\lim_{\epsilon\downarrow 0}\eta_\epsilon$ exists since $\eta_\epsilon \le \eta_{\epsilon'}$ when $\epsilon' \ge \epsilon$ and the entire sequence is bounded between 0 and $c_0 + c_1 + c_\perp$. Let

$$X^\epsilon := (\rho_1^\epsilon, \rho_2^\epsilon, \rho_3^\epsilon, y_{a,x}^\epsilon) \tag{40}$$

be an optimal solution to the $\eta_\epsilon$ SDP (which exists by compactness). Then, again by compactness, the sequence $\{(X^{1/t}) : t \in \mathbb{N}\}$ has an accumulation point (a limit of a convergent subsequence). Let $X' = (\rho_1', \rho_2', \rho_3', y_{a,x}')$ be this accumulation point. Since the objective function is also continuous in $X$, we have that $\eta \ge \lim_{\epsilon\downarrow 0}\eta_\epsilon$. On the other hand, let $X = (\rho_1, \rho_2, \rho_3, y_{a,x})$ be an optimal solution of the above SDP. This is feasible in any of the $\eta_\epsilon$ SDPs. Thus, $\eta_\epsilon \ge \eta$ for all $\epsilon$. This, we have our result. □

**Lemma 21** (Tidying it up). *The SDP above may be expressed as*

$$\begin{aligned}\eta = \text{maximize: } & tr([c_0\Pi_0 + \Pi_1(c_1\Pi^{\mathrm{GHZ}} + c_\perp\bar{\Pi}^{\mathrm{GHZ}})]\rho_3) && (41)\\ \text{subject to: } & \rho_1 = |\phi'\rangle\langle\phi'| && (42)\\ & tr_{S,I,J}(\rho_1) = tr_G(\rho_2) && (43)\\ & \rho_2 \otimes \frac{\mathbb{I}_Y}{2} = tr_D(\rho_3), && (44)\end{aligned}$$

*where* $|\phi'\rangle = \frac{1}{2\sqrt{2}}\sum_{a,x,r}|x,a,r\rangle|a\oplus x.r\rangle \otimes |\psi_{a,x}\rangle_{IJ}$ *where* $|\psi_{a,x}\rangle_{IJ}$ *is the post-measured state on Bob's side conditioned on Alice obtaining outcome* $a$ *using measurement* $x$.

*Proof.* It is easy to see that any feasible $\rho_1$ must be of the form $\rho_1' \otimes |junk\rangle\langle junk|$. Similarly, any feasible $\rho_2$ must be of the form $\rho_2' \otimes \mathrm{tr}_{I'JJ'}(|junk\rangle\langle junk|)$. Similarly, any feasible $\rho_3$ must be of the form $\rho_3' \otimes \mathrm{tr}_{II'JJ'}(|junk\rangle\langle junk|)$. Thus, we can reduce these variables. We now remove the $HH'$ spaces. Note that we have

$$\Pi_{a|x}|GHZ\rangle = \sqrt{\mathrm{prob}(a|x)}\,|\psi'_{x,a}\rangle_H\,|\psi_{x,a}\rangle_{IJ}, \tag{45}$$

where the states above are the post-measured state for both Alice and Bob, respectively. Now, Alice can apply the following controlled unitary

$$U_{XAH} = \sum_{x,a}|x,a\rangle\langle x,a| \otimes U_{x,a} \tag{46}$$

where $U_{x,a}$ maps $|\psi'_{x,a}\rangle$ to $|0\rangle$. Since applying this unitary on both sides of each constraint does does not affect feasibility, and it does not change the objective function value, we can redefine the variables to apply this unitary. The idea is that the $H$ space is not utilized by Alice again throughout the protocol (and so she can "clean it up" and discard it). □

## 6.2 Conjecture when Bob self-tests | $Q$

**Conjecture 22** (Continuity conjecture for $Q$). *Let* $c_0, c_1, c_\perp$ *be non-negative. Denote by* $v_0, v_1, v_\perp$ *the probability that Bob outputs* $0, 1, \perp$ *respectively when Protocol* $Q$ *is executed against some cheating strategy of Alice. Let* $\epsilon(N)$ *denote the winning probability of the GHZ boxes deduced by Bob (via Proposition 23) when he tests* $N-1$ *triples of boxes. Consider the objective* $\eta_\epsilon = c_0v_0 + c_1v_1 + c_\perp v_\perp$*. Then,* $\lim_{N\to\infty}\eta_\epsilon = \eta$ *where* $\eta$ *is the value of the SDP programme in Equation (17).*

**Protocol GHZ** Estimation of the GHZ value

1. Pick a box $J \in [n]$ uniformly at random.
2. For $i \in [n] \backslash J$, play the GHZ game with box $i$, denote outcome of game as $X_i \in \{0, 1\}$
3. If

$$\Omega: \quad X_i = 1, \text{ for all } i \in [n] \backslash J \tag{47}$$

4. Then conclude that the remaining box satisfies

$$T: \quad E[X_J|J, \Omega] \geq 1 - \delta \tag{48}$$

## 6.3 Estimation of GHZ winning probability

In the interest of clarity, we restate the self-testing procedure. We assume that the $3n$ boxes are described by some joint quantum state and local measurement operators. After playing the GHZ game with $3(n-1)$ of them, and verifying that they all pass, we want to make a statement about the remaining box, whose state $\tilde{\rho}$ is conditioned on the passing of all the other test.

The expectation value of $E[X_J|J, \Omega]$ accurately describes the expected GHZ value associated to the state of the remaining boxes $J$, conditioned on having measuring some outcome sequence in the other boxes which passes all the GHZ tests. Note that the conditioning in $J$ is important because otherwise we would get a bound on the GHZ averaged over all boxes, but we are only interested in the remaining box.

**Proposition 23** (Security of Protocol GHZ)**.** *For any implementation of the boxes and choice of $\delta > 0$ the joint probability that that the test $\Omega$ passes and that the conclusion $T$ is false is small* $\Pr[\Omega \cap \bar{T}] \leq \frac{1}{1-\delta+n\delta} \leq \frac{1}{n\delta}$*, where the first upper-bound is tight.*

This is the correct form of the security statement. It is important to bound the joint distribution of $\Omega$ and $\overline{T}$, and not $\Pr[\overline{T}|\Omega]$, conditioning on passing the test $\Omega$. Indeed in the latter case, it would not be possible to conclude anything of value about the remaining box $J$, as there could be some implementation of the boxes which has a very low expectation value of GHZ, but which passes the test with small but non-zero probability. Consider a hypothetical ideal protocol, which after having chosen $J$, only passes when $T$ is true. In that case, $\Pr[\Omega \cap \overline{T}] = 0$. Then the actual protocol is equivalent the ideal one, except that it fails with probability $\epsilon = \frac{1}{1-\delta+m\delta}$, and so it is $\epsilon$-close to the ideal algorithm.

*Proof.* For a given implementation of the boxes, let $p(x_1, \cdots x_n)$ denote the joint probability distribution of passing the GHZ games. Let $S = \{j|\ E[X_j|J = j, \Omega] < 1 - \delta\} \subset [n]$ be the set of boxes that have an expectation value for GHZ (conditioned on passing in the other boxes) below our target threshold and let $m = |S|$ be the number of such boxes. The value of $m$ is unknown, so we will need to maximise over it in the end.

Let $\alpha = \Pr(\{X_i\}_i = 1)$ and $\beta_i = \Pr(\{X_i\}_{i \neq j} = 1 \cap X_j = 0)$ be respectively the probabilities of the events where all the tests pass, or they all pass except for the $j$th test. This allows us to rewrite $E[X_j|J = j, \Omega] = \Pr(\{X_i\}_i = 1)/\Pr(\{X_i\}_{i \neq j} = 1) = \alpha/(\alpha + \beta_j)$, and so, by definition of $S$, we have $\alpha/(\alpha + \beta_j) < (1 - \delta)$, for $j \in S$, which is equivalent to $\beta_j > \frac{\delta}{1-\delta}\alpha$.

The aim of the proof is to bound the probability $\Pr[\Omega \cap \overline{T}]$. If we condition and summed over the different values of $J$, we can rewrite it as

$$\Pr(\Omega \cap \overline{T}) = \sum_j \frac{1}{n} \Pr(\Omega \cap \overline{T}|J = j) = \sum_{j \in S} \frac{1}{n} \Pr(\{X_i\}_{i \neq j} = 1) = \frac{1}{n} \sum_{j \in S} (\alpha + \beta_i)\,, \tag{49}$$

where we have kept the round $j \in S$ ones, conditioned on which $T$ is false. We are thus left with the optimisation problem

$$\max_{\alpha\geq 0,(\beta_i)_i\geq 0} \quad \frac{1}{n}\left(\sum_{j\in S}\alpha+\beta_j\right) \tag{50}$$

$$\text{subject to} \quad \alpha+\sum_{j\in S}\beta_j \leq 1 \tag{51}$$

$$\beta_j \geq \frac{\delta}{1-\delta}\alpha, \text{ for } j\in S \tag{52}$$

This is a linear problem. Simplifying it by defining $\Sigma = \sum_{j\in S}\beta_j$, gives

$$\max_{\alpha\geq 0,\Sigma\geq 0} \quad \frac{1}{n}(m\alpha+\Sigma) \tag{53}$$

$$\text{subject to} \quad \alpha+\Sigma\leq 1 \tag{54}$$

$$\Sigma \geq m\frac{\delta}{1-\delta}\alpha \tag{55}$$

It is easily shown that the maximum is attained for $(\alpha,\Sigma)=\left(\frac{1-\delta}{1-\delta+m\delta},\frac{m\delta}{1-\delta+m\delta}\right)$ which gives the upper-bound

$$\Pr[\Omega\cap\overline{T}] \leq \frac{1}{n}\max_m \frac{m}{1-\delta+m\delta} = \frac{1}{1-\delta+n\delta} \tag{56}$$

We note that the upper-bound is an increasing function of $m$ and so the maximum is attained for $m=n$. This yield the desired upper-bound. From the converse statement, we note that from the present proof we can construct a probability distribution $p(x_1,\cdots x_n)$, which saturates all inequalities, and so the upper-bound $\frac{1}{1-\delta+n\delta}$ is tight. □

# References

[ACG+14] Dorit Aharonov, André Chailloux, Maor Ganz, Iordanis Kerenidis, and Loïck Magnin, *A simpler proof of existence of quantum weak coin flipping with arbitrarily small bias*, SIAM Journal on Computing **45** (2014), no. 3, 633–679.

[ARV] Atul Singh Arora, Jérémie Roland, and Chrysoula Vlachou, *Analytic quantum weak coin flipping protocols with arbitrarily small bias*, pp. 919–938.

[ARW19] Atul Singh Arora, Jérémie Roland, and Stephan Weis, *Quantum weak coin flipping*, Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing - STOC 2019, ACM Press, 2019.

[AS] N. Aharon and J. Silman, *Quantum dice rolling: A multi-outcome generalization of quantum coin flipping.*

[Blu83] Manuel Blum, *Coin flipping by telephone a protocol for solving impossible problems*, SIGACT News **15** (1983), no. 1, 23–27.

[CGS13] André Chailloux, Gus Gutoski, and Jamie Sikora, *Optimal bounds for semi-honest quantum oblivious transfer.*

[CK09] André Chailloux and Iordanis Kerenidis, *Optimal Quantum Strong Coin Flipping*, 50th FOCS, 2009, pp. 527–533.

[CK11] ______, *Optimal Bounds for Quantum Bit Commitment*, 52nd FOCS, 2011, pp. 354–362.

[Gan09] Maor Ganz, *Quantum Leader Election.*

[HW11] Esther Hänggi and Jürg Wullschleger, *Tight bounds for classical and quantum coin flipping*, Theory of Cryptography (Berlin, Heidelberg) (Yuval Ishai, ed.), Springer Berlin Heidelberg, 2011, pp. 468–485.

[Kit03] Alexei Kitaev, *Quantum coin flipping*, Talk at the 6th workshop on Quantum Information Processing, 2003.

[McK14] Matthew McKague, *Self-testing graph states*, Theory of Quantum Computation, Communication, and Cryptography (Berlin, Heidelberg) (Dave Bacon, Miguel Martin-Delgado, and Martin Roetteler, eds.), Springer Berlin Heidelberg, 2014, pp. 104–120.

[Mil20] Carl A. Miller, *The impossibility of efficient quantum weak coin flipping*, Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (New York, NY, USA), STOC 2020, Association for Computing Machinery, 2020, pp. 916–929.

[Moc07] Carlos Mochon, *Quantum weak coin flipping with arbitrarily small bias*, arXiv:0711.4114 (2007).

[MS13] Carl A. Miller and Yaoyun Shi, *Optimal Robust Self-Testing by Binary Nonlocal XOR Games*, 8th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2013) (Dagstuhl, Germany) (Simone Severini and Fernando Brandao, eds.), Leibniz International Proceedings in Informatics (LIPIcs), vol. 22, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013, pp. 254–262.

[SCA+11] J. Silman, A. Chailloux, N. Aharon, I. Kerenidis, S. Pironio, and S. Massar, *Fully distrustful quantum bit commitment and coin flipping*, Physical Review Letters **106** (2011), no. 22.

# 7 Acknowledgements

ASA is grateful for the hospitality of Perimeter Institute where part of this work was carried out. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Economic Development, Job Creation and Trade. ASA further acknowledges the FNRS (Belgium) for support through the FRIA grants, 3/5/5 – MCF/XH/FC – 16754 and F 3/5/5 – FRIA/FC – 6700 FC 20759. A part of this work was carried out while ASA was at the Université libre de Bruxelles, Belgium.

# A Device Independence and the Box Paradigm

We describe device independent protocols as classical protocols with one modification: we assume that the two parties can exchange boxes and that the parties can shield their boxes (from the other boxes i.e. the boxes can't communicate with each other once shielded).

**Definition 24** (Box)**.** A *box* is a device that takes an input $x \in \mathcal{X}$ and yields an outputs $a \in \mathcal{A}$ where $\mathcal{X}$ and $\mathcal{A}$ are finite sets. Typically, a set of $n$ boxes, taking inputs $x_1, x_2, \dots x_n$ and yielding outputs $a_1, a_2 \dots a_n$ are *characterised* by a joint conditional probability distribution, denoted by

$$p(a_1, a_2 \dots a_n | x_1, x_2 \dots x_n).$$

Further, if $p(a_1, a_2 \dots a_n | x_1, x_2 \dots x_n) = \text{tr}\left[M^1_{a_1|x_1} \otimes M^2_{a_2|x_2} \cdots \otimes M^n_{a_n|x_n} \rho\right]$ then we call the set of boxes, *quantum boxes*, where $\{M^i_{a'|x'}\}_{a' \in \mathcal{A}_i}$constitute a POVM for a fixed $i$ and $x'$, $\rho$ is a density matrix and their dimensions are mutually consistent.

Henceforth, we restrict ourselves to quantum boxes.

**Definition 25** (Protocol in the box formalism)**.** A generic two-party protocol in the box formalism has the following form:

1. Inputs:
   (a) Alice is given boxes $\square^A_1, \square^A_2 \dots \square^A_p$ and Bob is given boxes $\square^B_1, \square^B_2, \dots \square^B_q$.
   (b) Alice is given a random string $r^A$ and Bob is given a random string $r^B$ (of arbitrary but finite length).
2. Structure: At each round of the protocol, the following is allowed.
   (a) Alice and Bob can locally perform arbitrary but finite time computations on a Turing Machine.
   (b) They can exchange classical strings/messages and boxes.

A protocol in the box formalism is readily expressed as a protocol which uses a (trusted) classical channel (i.e. they trust their classical devices to reliably send/receive messages), untrusted quantum devices and an untrusted quantum channel (i.e. a channel that can carry quantum states but may be controlled by the adversary).

**Assumption 26** (Setup of Device Independent Two-Party Protocols)**.** *Alice and Bob*

1. *both have private sources of randomness,*
2. *can send and receive classical messages over a (trusted) classical channel,*
3. *can prevent parts of their untrusted quantum devices from communicating with each other, and*
4. *have access to an untrusted quantum channel.*

We restrict ourselves to a "measure and exchange" class of protocols—protocols where Alice and Bob start with some pre-prepared states and subsequently, only perform classical computation and quantum measurements locally in conjunction with exchanging classical and quantum messages. More precisely, we consider the following (likely restricted) class of device independent protocols.

**Definition 27** (Measure and Exchange (Device Independent Two-Party) Protocols)**.** A *measure and exchange (device independent two-party) protocol* has the following form:

1. Inputs:
   (a) Alice is given quantum registers $A_1, A_2, \dots A_p$ together with POVMs[16]

$$\{M^{A_1}_{a|x}\}_a, \{M^{A_2}_{a|x}\}_a, \dots \{M^{A_p}_{a|x}\}_a$$

[16] For concreteness, take the case of binary measurements. By $\{M^{A_1}_{a|x}\}_a$, for instance, we mean $\{M^{A_1}_{0|x}, M^{A_1}_{1|x}\}$ is a POVM for $x \in \{0, 1\}$.

which act on them and Bob is, analogously, given quantum registers $B_1, B_2, \dots B_q$ together with POVMs

$$\{M_{b|y}^{B_1}\}_b, \{M_{b|y}^{B_2}\}_b, \dots, \{M_{b|y}^{B_q}\}_b.$$

Alice shields $A_1, A_2, \dots A_p$ (and the POVMs) from each other and from Bob's lab. Bob similarly shields $B_1, B_2 \dots B_q$ (and the POVMs) from each other and from Alice's lab.

(b) Alice is given a random string $r^A$ and Bob is given a random string $r^B$ (of arbitrary but finite length).

2. Structure: At each round of the protocol, the following is allowed.

   (a) Alice and Bob can locally perform arbitrary but finite time computations on a Turing Machine.

   (b) They can exchange classical strings/messages.

   (c) Alice (for instance) can

      i. send a register $A_l$ and the encoding of her POVMs $\{M_i^{A_l}\}_i$ to Bob, or

      ii. receive a register $B_m$ and the encoding of the POVMs $\{M_i^{B_m}\}_i$.

   Analogously for Bob.

It is clear that a protocol in the box formalism (Definition 25) which uses only quantum boxes (Definition 24) can be implemented as a measure and exchange protocol (Definition 27).