Advantage distillation for device-independent quantum key distribution

Ernest Y.-Z. Tan,1 Charles C.-W. Lim,2,3 and Renato Renner1

1Institute for Theoretical Physics, ETH Zürich, Switzerland
2Department of Electrical & Computer Engineering, National University of Singapore, Singapore
3Centre for Quantum Technologies, National University of Singapore, Singapore

We derive a sufficient condition for advantage distillation to be secure against collective attacks in device-independent quantum key distribution (DIQKD), focusing on the repetition-code protocol. In addition, we describe a semidefinite programming method to check whether this condition holds for any probability distribution obtained in a DIQKD protocol. Applying our method to various probability distributions, we find that advantage distillation is possible up to depolarising-noise values of \( q \approx 9.1\% \) or limited detector efficiencies of \( \eta \approx 89.1\% \) in a 2-input 2-output scenario. This exceeds the noise thresholds of \( q \approx 7.1\% \) and \( \eta \approx 90.9\% \) respectively for DIQKD with one-way error correction using the CHSH inequality, thereby showing that it is possible to distill secret key beyond those thresholds.

I. INTRODUCTION

In quantum key distribution, the goal is to extract a key from correlations obtained by measuring some quantum systems. Device-independent quantum key distribution (DIQKD) is based on the observation that when these correlations violate a Bell inequality, a secure key can be extracted even if the users’ devices are not fully characterised [1–4]. Specifically, when establishing the security of DIQKD protocols, one does not assume that the users’ devices are working exactly according to their specifications. Instead, it is only required that the devices do not broadcast their outputs directly to the adversary [5], or signal to other components except when foreseen by the protocol [3]. This is in contrast to more traditional QKD protocols [6], which are device-dependent in the sense that they assume the devices are implementing some specified quantum operations, at least within some tolerance ranges [7]. Implementations of such QKD protocols have been attacked by a variety of methods [8–10], typically by exploiting imperfections which cause the devices to operate outside of the prescribed models. By working with fewer assumptions on the devices, DIQKD offers the possibility of secure key distribution with less detailed characterisation of the devices, which would reduce the scope of possible attacks.

Unfortunately, the prospect of increased security offered by DIQKD comes at the cost of more stringent requirements on the physical implementations, such as requiring rather low noise levels to achieve positive key rates. There has been substantial difficulty in finding security proofs for DIQKD protocols with sufficient noise robustness for experimental realisation. One approach towards improving the robustness may be to further investigate the error-correction step. In QKD, the raw data of the two users is typically not perfectly correlated, and they need to agree on a shared secret key using public communication. Existing DIQKD security proofs [1–3, 11–13] have mainly focused on using one-way communication for error correction. However, for classical key reconciliation [11, 12] and device-dependent QKD protocols such as the BB84 and six-state protocols [13–16], it has been found that the noise threshold for one-way error correction can be surpassed by using two-way communication, a concept sometimes referred to as advantage distillation. In this work, we address the question of whether this may also be true for DIQKD, focusing on the repetition-code protocol [11–14]. We mainly aim to find the noise thresholds that advantage distillation can tolerate, rather than optimising the key rate at a given noise value.

The paper is organised as follows. In Sec. II, we describe the family of DIQKD scenarios considered in this work, and derive a sufficient condition for advantage distillation to successfully produce a secure key in such protocols. We also describe and apply an SDP method to check whether this condition holds for several possible DIQKD protocols, thereby obtaining lower bounds on their noise tolerance. In Sec. III, we describe an improvement to this condition that can be applied for 2-input 2-output DIQKD protocols, and show that this improved condition allows for noise tolerance beyond that of one-way error correction. Finally, in Sec. IV, we further discuss these results and consider possibilities for future work.

II. GENERAL CONDITION

We consider a DIQKD protocol where Alice has \( \mathcal{X} \) possible measurements \( A_0, A_1, \ldots, A_{\mathcal{X}-1} \), and similarly Bob has \( \mathcal{Y} \) possible measurements \( B_0, B_1, \ldots, B_{\mathcal{Y}-1} \), with \( A_0 \) and \( B_0 \) taken to be binary-outcome measurements from which a raw key is generated. Following the collective-attack model, we assume that Alice and Bob’s state and possible measurements across the rounds are in IID tensor-product form. In each round, Eve is allowed to hold some purification.
of the Alice-Bob state $\rho_{AB}$, and since all purifications are isometrically equivalent, we can assume the purified state $\rho_{ABE}$ is IID as well. Given the IID structure, Alice and Bob can carry out parameter estimation to arbitrary accuracy, so we shall assume the outcome probabilities $\Pr_{A|E}[xy](ab|xy)$ for each pair of measurements $A_xB_y$ on the state are fully characterised in the protocol. (In the rest of this work, we will suppress the subscripts for probability distributions when the relevant distribution is clear from context.) We focus on cases where the key-generating measurements have symmetrised outcomes, in the sense $\Pr(01|00) = \Pr(10|00) = \epsilon/2$ and $\Pr(00|00) = \Pr(11|00) = (1 - \epsilon)/2$ for some $\epsilon < 1/2$ (if $\epsilon > 1/2$, simply swap the labels of Bob’s outcomes). This can be enforced in any binary-outcome scenario via a symmetrisation step, where Alice generates a uniformly random bit $T$ in each round and publicly sends it to Bob, with both of them flipping their measurement outcomes if and only if $T = 1$. \footnote{In this work, when symmetrisation is implemented, we take it to be applied for all measurement pairs.}

The repetition-code protocol \cite{11-14} for advantage distillation is based on the outputs of a block of $n$ rounds in which $A_0B_0$ was measured (we shall denote the output bitstrings as $A_0$, $B_0$, and Eve’s side-information across all the rounds as $E$). Alice privately generates a uniformly random bit $C$, and sends the message $M = A_0 \oplus (C, C', ..., C)$ to Bob via a public authenticated channel. Bob replies to Alice to accept the block if and only if $B_0 \oplus M = (C', C'', ..., C')$ for some $C' \in \mathbb{Z}_2$. If the resulting systems satisfy

$$H(C|EM; \text{accept}) - H(C|C'; \text{accept}) > 0,$$

then repeating this procedure over many $n$-round blocks would allow a secret key to be distilled asymptotically from the bitstrings $C, C'$ in the accepted blocks, via a one-way error correction protocol \cite{17}. Excluding rounds used for parameter estimation, the asymptotic key rate will be $(H(C|EM; \text{accept}) - H(C|C'; \text{accept}))/(e^n + (1 - e)^n)/n$ \cite{13}. We now prove the following theorem (where $F(\rho, \sigma) = \left\|\sqrt{\rho} - \sqrt{\sigma}\right\|_1$ is the root-fidelity):

**Theorem 1.** For a DIQKD protocol as described above, a sufficient condition for Eq. (1) to hold for large $n$ is

$$F(\rho_{E|00}, \rho_{E|11})^2 > \frac{\epsilon}{1 - \epsilon},$$

where $\rho_{E|a_0b_0}$ is Eve’s single-round state conditioned on $A_0B_0$ being measured with outcome $a_0b_0$.

**Proof.** All density matrices denoted in this proof are normalised. To bound $H(C|EM; \text{accept})$, we first observe that since $H(X|Y; Z) = \sum_z \Pr_Z(z)H(X|Y; Z = z)$ for classical $Z$, it suffices to bound $H(C|E; M = m \land \text{accept})$ for arbitrary messages $m$. Starting from the initial $A_0B_0E$ state

$$\rho_{A_0B_0E} = \sum_{a_0, b_0} \Pr_{A_0B_0}(a_0, b_0) |a_0, b_0\rangle \langle a_0, b_0| \otimes \rho_{E|a_0b_0},$$

a straightforward calculation shows that conditioned on the block being accepted and $M = m$, the $CE$ state takes the form $\rho_{CE|M=m \& \text{accept}} = \sum_c (1/2) |c\rangle \langle c| \otimes \omega_c$ with

$$\omega_0 = \frac{\Pr_{A_0B_0}(m, m)\rho_{E|m m} + \Pr_{A_0B_0}(m, \overline{m})\rho_{E|m \overline{m}}}{\Pr_{A_0B_0}(m, m) + \Pr_{A_0B_0}(m, \overline{m})}, \quad \omega_1 = \frac{\Pr_{A_0B_0}(\overline{m}, m)\rho_{E|\overline{m}m} + \Pr_{A_0B_0}(\overline{m}, \overline{m})\rho_{E|\overline{m}\overline{m}}}{\Pr_{A_0B_0}(\overline{m}, m) + \Pr_{A_0B_0}(\overline{m}, \overline{m})},$$

where $\overline{m} = m \oplus 1$. We now consider a state $\tilde{\rho}_{CE}$ defined as follows:

$$\tilde{\rho}_{CE} = (1/2) (|0\rangle \langle 0| \otimes \rho_{E|m m} + |1\rangle \langle 1| \otimes \rho_{E|\overline{m}\overline{m}}).$$

With symmetrised IID outcomes, we have $\Pr_{A_0B_0}(m, m) = \Pr_{A_0B_0}(\overline{m}, \overline{m}) = (1 - \epsilon)^n$ and $\Pr_{A_0B_0}(m, \overline{m}) = \Pr_{A_0B_0}(\overline{m}, m) = \epsilon^n$, so

$$d(\rho_{CE|M=m \& \text{accept}}, \tilde{\rho}_{CE}) \leq \delta_n, \text{ where } \delta_n = \frac{\epsilon^n}{e^n + (1 - \epsilon)^n}.$$

Applying a continuity bound for conditional von Neumann entropy \cite{18} then yields

$$H(C|E; M = m \land \text{accept}) \geq H(C|E) - \delta_n - (1 + \delta_n) h_2\left(\frac{\delta_n}{1 + \delta_n}\right),$$

where

$$h_2(x) = -x\log_2 x - (1 - x)\log_2 (1 - x).$$
where \( h_2 \) is the binary entropy function. The \( H(C|E)_{\hat{\rho}} \) term is bounded by

\[
H(C|E)_{\hat{\rho}} \geq 1 - h_2 \left( \frac{1 - F(\rho_{E|mm}, \rho_{E|mm})}{2} \right) = 1 - h_2 \left( \frac{1 - F(\rho_{E|00}, \rho_{E|11})}{2} \right),
\]

(8)

using the IID assumption. As for \( H(C|C'; \text{accept}) \), it can be seen that \( \Pr(C \neq C'|\text{accept}) = \delta_n \) for IID outcomes, so

\[
H(C|C'; \text{accept}) = h_2(\delta_n).
\]

(9)

Combining these results, we conclude

\[
\frac{H(C|EM; \text{accept})}{H(C|C'; \text{accept})} \geq \left( 1 - h_2 \left( \frac{1 - F(\rho_{E|00}, \rho_{E|11})}{2} \right) \right) - \delta_n - (1 + \delta_n)h_2 \left( \frac{\delta_n}{1 + \delta_n} \right) h_2(\delta_n)^{-1},
\]

(10)

and it can be shown (see Appendix A) that when Eq. (2) holds, then the right-hand side of Eq. (10) limits to \( +\infty \) as \( n \to \infty \). Therefore, Eq. (1) will hold for sufficiently large \( n \).

Given specific values or bounds for \( F(\rho_{E|00}, \rho_{E|11}), \epsilon, n \), one can substitute them into Eq. (10) to get an explicit bound on the asymptotic key rate (see Eq. (1)). Eq. (2) is in fact the same condition as that obtained in [12] for device-dependent QKD, but here we obtain it with less detailed knowledge of the state. We now require a device-independent bound on \( F(\rho_{E|00}, \rho_{E|11}) \). One way to do so is to use the Fuchs-van de Graaf inequality [21] together with the operational interpretation of trace distance,

\[
F(\rho_{E|00}, \rho_{E|11}) \geq 1 - d(\rho_{E|00}, \rho_{E|11}) = 2(1 - P_g(\rho_{E|00}, \rho_{E|11})),
\]

(11)

where \( P_g(\rho_{E|00}, \rho_{E|11}) \) is Eve’s maximum probability of guessing \( C \) given the \( E \) part of a c-q state \( \sigma_{CE} = \sum_{n}(1/2) |c\rangle \langle c| \otimes \rho_{E|cc} \). We now observe that in a DIQKD protocol as described above, \( P_g(\rho_{E|00}, \rho_{E|11}) \) can be viewed as Eve’s guessing probability of the outcome of \( AB_0 \), conditioned on the outcome being either 00 or 11. A device-independent method to bound such guessing probabilities based on the distribution \( \Pr_{AB|XY} \) has been described in [23], using the family of SDPs known as the NPA hierarchy [24]. We can hence simply apply this method to find whether Eq. (2) holds for various distributions. The method can also be adapted for a scenario where the probabilities \( \Pr(\text{ab}|\text{xy}) \) are bounded by intervals instead of exactly known; we leave this for future work.

We now separately consider two possible noise models for binary-outcome distributions. The first is depolarising noise parametrised by \( q \in [0, 1/2] \):

\[
\Pr(\text{ab}|\text{xy}) = (1 - 2q)\Pr_{\text{target}}(\text{ab}|\text{xy}) + q/2,
\]

(12)

where \( \Pr_{\text{target}} \) is some ideal distribution. If \( \Pr_{\text{target}}(00|00) = \Pr_{\text{target}}(11|00) = 1/2 \), then \( \epsilon = q \), and if \( \Pr_{\text{target}} \) has symmetrise outcomes, then the resulting noisy distribution does as well. The second noise model is a limited-detector-efficiency model parametrised by \( \eta \in [0, 1] \), where we start with some \( \Pr_{\text{target}} \) and then subject all outcomes to independent Z-channels that flip 1 to 0 with probability \( 1 - \eta \) [1]. This is a simplistic model for a photonic setup where detectors fail to detect photons with probability \( 1 - \eta \) and the outcome 0 is assigned whenever this occurs, in order to avoid the postselection loophole. This noise model does not generally produce symmetrised outcomes for \( AB_0 \), and hence a symmetrisation step must be explicitly implemented when this model is used.

In Table I we summarise the results obtained by applying our method to various distributions subject to either of the noise models. All computations were performed at level 3 of the NPA hierarchy [23] except scenarios (i)-(ii) which were at level 4. We find that scenario (iii) has the best robustness against depolarising noise, while scenario (ii) and (iii) have the best robustness against limited detector efficiency. Scenario (ii) is somewhat different from the others in that it does not have symmetrisated outcomes, and hence its noise tolerance in the limited-detector-efficiency model is affected by the outcome labelling choice (the labelling chosen here appeared to be the optimal choice for those measurement angles). However, these results narrowly fail to outperform the one-way error correction protocol in [1], which uses the same \( \Pr_{\text{target}} \) as scenario (i) (with Alice and Bob interchanged) but can tolerate noise up to \( q_1 \approx 7.1\% \) and \( \eta_1 \approx 92.4\% \) (or \( \eta_1 \approx 99.9\% \) if the state and measurements are optimised to maximise CHSH violation for each value of \( \eta \) [24]). The approach in this section does at least have the benefit of being quite generally applicable, hence allowing us to find whether the repetition-code protocol is secure against collective attacks in a wide variety of DIQKD scenarios. In the following section, we instead restrict ourselves to 2-input 2-output protocols, and in doing so find an improvement to Theorem 1 that allows us to outperform one-way error correction.

---

2 Alternatively, one could also use \( H(C|E)_{\hat{\rho}} \geq 1 - d(\rho_{E|mm}, \rho_{E|mm}) \) [20] or even the weaker bound \( H(C|E)_{\hat{\rho}} \geq H_{\text{min}}(C|E)_{\hat{\rho}} = -\log((1 + d(\rho_{E|mm}, \rho_{E|mm}))/2) \), followed by using the Fuchs-van de Graaf inequality to bound \( d(\rho_{E|mm}, \rho_{E|mm}) \) in terms of \( F(\rho_{E|mm}, \rho_{E|mm}) \). This results in slightly weaker bounds on \( H(C|EM; \text{accept}) \) than Eq. (10), but yields the same final statement as in Theorem 1.

3 In this work, by level \( M \) of the NPA hierarchy we always mean “global” level \( M \), i.e. the NPA matrix is indexed by operators consisting of products of up to \( M \) measurement operators in total. This is in contrast to “local” level \( M \), where the index operators are products of up to \( M \) measurement operators per party.
TABLE I. In each row, $q_t$ is the maximum noise that the distribution $P_{\text{target}}$ can tolerate in a depolarising-noise model, such that we can still show Eq. (2) is satisfied and hence positive key rate is achievable. Analogously, $\eta_t$ is the minimum efficiency which can be tolerated when we instead consider a limited-detection-efficiency model. Scenario (ii) and all limited-detector-efficiency models do not have symmetrised outcomes, so a symmetrisation step was carried out after applying noise.

| Scenario (i) | Description of $P_{\text{target}}$ | State and measurements for $P_{\text{target}}$ | $q_t$ | $\eta_t$ |
|-------------|------------------------------------|-----------------------------------------------|------|--------|
| (i)         | Maximally violates CHSH with the measurements $A_0, A_1, B_1, B_2$. | $|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$; $A_0 = B_1 = Z, A_1 = X, B_1 = (X + Z)/\sqrt{2}, B_2 = (X - Z)/\sqrt{2}$. | 6.0% | 93.7% |
| (ii)        | Exhibits the Hardy paradox “maximally” \cite{22, 23} on $A_1, A_2, B_1, B_2$, with $A_0, B_0$ in the Schmidt basis. | Let $\alpha = \sqrt{(3 - \sqrt{5})/2}$. $|\psi\rangle = \alpha(10) + \sqrt{1 - 2\alpha^2}|11\rangle$; the 0 outcomes correspond to projectors onto $|a_0\rangle = |b_0\rangle \propto \sqrt{1 + 2\alpha^2 - \sqrt{1 - 2\alpha^2}} |0\rangle + 2\alpha |1\rangle$, $|a_1\rangle = |b_1\rangle = |1\rangle, |a_2\rangle = |b_2\rangle \propto |0\rangle + \sqrt{1 - 2\alpha^2}|1\rangle$. | 3.0% | 92.6% |
| (iii)       | Includes the Mayers-Yao self-test \cite{23} and the CHSH measurements. | $|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$; $A_0 = B_0 = Z, A_1 = B_1 = (X + Z)/\sqrt{2}, A_2 = B_2 = X, A_3 = B_3 = (X - Z)/\sqrt{2}$. | 6.8% | 92.7% |
| (iv)        | Maximally violates the elegant Bell inequality \cite{28, 29} with the measurements $A_1 - A_3, B_0 - B_3$. | $|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$; $A_0 = B_0 = - (X + Y + Z)/\sqrt{3}, A_1 = A_2 = Y, A_3 = X, B_1 = (X + Y - Z)/\sqrt{3}, B_2 = (X - Y + Z)/\sqrt{3}, B_3 = (X - Y + Z)/\sqrt{3}$. | 6.1% | 93.6% |
| (v)         | Maximally violates the elegant Bell inequality \cite{28, 29} with the measurements $A_0 - A_2, B_1 - B_4$. | $|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$; $A_0 = B_0 = Z, A_1 = Y, A_2 = X, B_1 = - (X + Y + Z)/\sqrt{3}, B_2 = (X + Y - Z)/\sqrt{3}, B_3 = (X - Y + Z)/\sqrt{3}, B_4 = (X + Y + Z)/\sqrt{3}$. | 4.3% | 95.5% |

III. 2-INPUT 2-OUTPUT PROTOCOLS

We observe that if $\rho_{E|00}, \rho_{E|11}$ were both assumed to be pure, then Eq. (11) could be replaced by a better relation, which saturates the other side of the Fuchs-van de Graaf inequality:

$$F(\rho_{E|00}, \rho_{E|11})^2 = 1 - d(\rho_{E|00}, \rho_{E|11})^2. \quad (13)$$

In the following theorem, we use this to show that for 2-input 2-output protocols, we can almost replace Eq. (11) with Eq. (13) after taking a particular concave envelope:

**Theorem 2.** Consider a DIQKD protocol as described in Sec. \cite{14} with $X = Y = 2$ and all measurements having binary outcomes. Suppose an explicit symmetrisation step is carried out, and the resulting probability distribution is $P_{\text{target}}$. Let $f$ be a concave function defined on the set of quantum distributions with $P_{\text{target}}(00) = P_{\text{target}}(11)$, such that $f$ is a DI upper bound on $(1 - \epsilon) d(\rho_{E|00}, \rho_{E|11})^2$. Then a sufficient condition for Eq. (1) to hold for large $n$ is

$$1 - \frac{f(P_{\text{target}})}{1 - \epsilon} > \frac{\epsilon}{1 - \epsilon}. \quad (14)$$

**Proof.** We use the fact that in a 2-input 2-output DI protocol, Jordan’s lemma can be used to argue \cite{11, 22} that without loss of generality, we can assume Eve’s strategy in each round consists of generating a random variable $X$ and storing it in a classical register, then implementing some corresponding qubit strategy, which is a strategy such that $\rho_{ABE|X} = 2 \times 2 \times 4$ pure state and all of Alice and Bob’s measurements are rank-1 projective measurements. For a qubit strategy, Eve’s state conditioned on the joint outcome of both measurements is pure. Hence if Alice and Bob measure $A_x, B_y$, and store their results in classical registers $A_x, B_y$, we can take the resulting single-round state $\rho_{A_x B_y E A}$ after symmetrisation to be in the form \cite{11}

$$\rho_{A_x B_y E A} = \sum_{a, b} \sum_{\lambda} \Pr(ab|xy\lambda) \Pr(\lambda) |a, b\rangle \langle a, b| \otimes \rho_{E|\lambda ab} \otimes |\lambda\rangle \langle \lambda|, \quad (15)$$

---

4 The notation $\rho_{E|\lambda ab}$ implicitly refers to different states for different measurement choices $(x, y)$, though it is not crucial to keep track of this in the proof, since it only uses the conditional states produced by the key-generating measurement. In any case, the notation is unambiguous if we choose to label the outcomes using disjoint sets for distinct measurements.
where the probabilities $\Pr(ab|xy\lambda)$ satisfy $\sum_\lambda \Pr(ab|xy\lambda)\Pr(\lambda) = \Pr(ab|xy)$, and for each $\lambda$ we have $\Pr(00|00\lambda) = \Pr(11|00\lambda) = (1-\epsilon_\lambda)/2$ for some $\epsilon_\lambda \in [0,1]$, due to the symmetrisation step. Using the reduction to qubit strategies, we can also take all $\rho_E|\lambda\eta$ to be pure states (see Appendix A). When Alice and Bob both perform the key-generating measurements and get the same outcome (which we shall denote here as $\gamma \in \{0,1\}$), Eve’s conditional states are

$$\rho_{E|\lambda\gamma} = \sum_\lambda \tilde{\Pr}(\lambda)\rho_{E|\lambda\gamma} \otimes |\lambda\rangle\langle\lambda|,$$

where $\tilde{\Pr}(\lambda) = \sum_j \tilde{\Pr}(\lambda_j)$ and $\rho_{E|\lambda\gamma} = \otimes_j \rho_{E|\lambda_j\gamma_j}$. Hence we have $H(C|E\Lambda)_\rho = \sum_\lambda \tilde{\Pr}(\lambda)H(C|E;\Lambda = \lambda)_\beta$. Applying the bound from [19] together with $h_2((1-p)/2) \leq 1 - p^2/\ln 4$ (this inequality follows from the Taylor expansion of $h_2$) then yields

$$H(C|E\Lambda)_\rho \geq \sum_\lambda \tilde{\Pr}(\lambda) \left(1 - h_2 \left(1 - \frac{1}{2} \frac{F(\rho_{E|\lambda\eta\eta}, \rho_{E|\lambda\eta\eta})}{F(\rho_{E|00\lambda\eta}, \rho_{E|11\lambda\eta})}\right)\right) \geq \frac{1}{\ln 4} \sum_\lambda \tilde{\Pr}(\lambda) F(\rho_{E|\lambda\eta\eta}, \rho_{E|\lambda\eta\eta})^2. \quad (18)$$

Using the IID structure,

$$\sum_\lambda \tilde{\Pr}(\lambda) F(\rho_{E|\lambda\eta\eta}, \rho_{E|\lambda\eta\eta})^2 = \sum_\lambda \prod_{j=1}^n \tilde{\Pr}(\lambda_j) F(\rho_{E|\lambda_j00}, \rho_{E|\lambda_j11})^2 = \left(\sum_\lambda \tilde{\Pr}(\lambda) F(\rho_{E|\lambda00}, \rho_{E|\lambda11})^2\right)^n. \quad (19)$$

Since the states $\rho_{E|00}, \rho_{E|11}$ are pure, they satisfy Eq. (13), and so

$$\sum_\lambda \tilde{\Pr}(\lambda) F(\rho_{E|\lambda00}, \rho_{E|\lambda11})^2 = 1 - \sum_\lambda \frac{1-\epsilon_\lambda}{1-\epsilon} \tilde{\Pr}(\lambda) d(\rho_{E|\lambda00}, \rho_{E|\lambda11})^2 \geq 1 - \epsilon \sum_\lambda \tilde{\Pr}(\lambda) f(Pr_{AB|XY\lambda}). \quad (20)$$

Finally, this is lower-bounded by $1 - f(Pr_{AB|XY})/(1-\epsilon)$ since $f$ is concave, so

$$\frac{H(C|E\Lambda M; \text{accept})}{H(C|C'; \text{accept})} \geq \left(\frac{1}{\ln 4} \left(1 - \frac{f(Pr_{AB|XY})}{1-\epsilon}\right)\right)^n \frac{\delta_n}{1+\delta_n} h_2\left(\frac{\delta_n}{1+\delta_n}\right) h_2(\delta_n)^{-1}, \quad (21)$$

and it can be shown (see Appendix A) that when Eq. (14) holds, then the right-hand side of Eq. (21) limits to $+\infty$ as $n \to \infty$. Therefore, Eq. (11) will hold for sufficiently large $n$.

If $f$ is the optimal concave upper bound on $(1-\epsilon)d(\rho_{E|00}, \rho_{E|11})^2$, in the sense that there always exists a mixture of qubit strategies such that $\sum_\lambda \Pr(\lambda)(1-\epsilon_\lambda)d(\rho_{E|\lambda00}, \rho_{E|\lambda11})^2 = f(Pr_{AB|XY})$, then the above analysis is essentially tight for large $n$. This is because in Eq. (18), the first inequality [19] is in fact saturated because $\rho_{E|\lambda\eta\eta}, \rho_{E|\lambda\eta\eta}$ are pure, and the second inequality is approximately saturated at large $n$ because $h_2((1-p)/2) = 1 - p^2/\ln 4 - O(p^4)$. \hfill \Box

At the moment, we do not have a concrete method for finding an optimal concave bound on $(1-\epsilon)d(\rho_{E|00}, \rho_{E|11})^2$. However, the following corollary introduces a tractable method for a more restrictive condition than Eq. (14):

**Corollary 1.** Consider a DIQKD protocol as described in Sec. 17 with $X = Y = 2$ and all measurements having binary outcomes. Suppose an explicit symmetrisation step is carried out, and the resulting probability distribution is $Pr_{AB|XY}$. Let $g$ be a DI upper bound on $d(\rho_{E|00}, \rho_{E|11})$ as a function of $Pr_{AB|XY}$. Then a sufficient condition for Eq. (11) to hold for large $n$ is

$$1 - g(Pr_{AB|XY}) > \frac{\epsilon}{1-\epsilon}. \quad (22)$$
Proof. Let \( \tilde{f} \) be the optimal upper bound on \((1 - \epsilon)d(\rho_{E|00}, \rho_{E|11})\), defined on the set of distributions such that \( \Pr(00|00) = \Pr(11|00) \). Since \( d(\rho_{E|00}, \rho_{E|11}) \leq 1 \), we have

\[
\tilde{f}(\Pr_{AB|XY}) \geq (1 - \epsilon)d(\rho_{E|00}, \rho_{E|11}) \geq (1 - \epsilon)d(\rho_{E|00}, \rho_{E|11})^2.
\]  

(23)

Also, \( \tilde{f} \) must be concave (see Appendix C), so choosing \( f = \tilde{f} \) satisfies the conditions of Theorem 2. Since \( \tilde{f} \) is the optimal bound on \((1 - \epsilon)d(\rho_{E|00}, \rho_{E|11})\), we must have \((1 - \epsilon)g(\Pr_{AB|XY}) \geq \tilde{f}(\Pr_{AB|XY})\). Hence when Eq. (22) holds, we have

\[
1 - \frac{\tilde{f}(\Pr_{AB|XY})}{1 - \epsilon} \geq 1 - g(\Pr_{AB|XY}) > \frac{\epsilon}{1 - \epsilon},
\]

and the claim follows by Theorem 2.

As in Section I, such a function \( g \) is provided by using the NPA hierarchy to bound the guessing probabilities. Effectively, Corollary 1 improves over the combination of Eq. (2) and Eq. (11) by replacing \((1 - \epsilon)g(\Pr_{AB|XY})^2 \) with \(1 - g(\Pr_{AB|XY})\). In Fig. 1, we show our results for \( \Pr_{\text{target}} \) given by the state \( |\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2} \) and measurements \( A_0 = Z, A_1 = X, B_0 = (X + Z)/\sqrt{2}, B_1 = (X - Z)/\sqrt{2} \), computing the bounds at NPA level 2 [23]. This is the same as scenario (i) in Table I (or the protocol in [1] with Alice and Bob interchanged), but omitting the measurement for Bob that is perfectly correlated to Alice for key generation. We find that this \( \Pr_{\text{target}} \) can tolerate depolarising noise of \( q_t \approx 7.7\% \) or detector efficiencies of \( \eta_t \approx 91.7\% \) while satisfying Eq. (22). This outperforms the thresholds of \( q_t \approx 7.1\%, \eta_t \approx 92.4\% \) for the one-way error correction protocol in [1], which is somewhat surprising given that the key-generating measurements for this \( \Pr_{\text{target}} \) are not perfectly correlated. In fact, if the proof in [1] for one-way error correction were applied to this \( \Pr_{\text{target}} \), we find it would only tolerate noise up to \( q_t \approx 3.1\% \).

For a limited-detector-efficiency model, the noise robustness can be further improved by optimising the state and measurements at each value of \( \eta_t \) to maximise the CHSH value [24]. With these states and measurements, applying the proof in [1] for one-way error correction shows that positive key rate is achievable down to \( \eta_t \approx 99.0\% \), while our approach shows that Eq. (22) is satisfied down to \( \eta_t \approx 89.1\% \). Hence the noise tolerance of the repetition-code protocol surpasses that of one-way error correction in this case as well.

Similarly, by numerically optimising the measurements for robustness against depolarising noise, we found that Eq. (22) can be satisfied up to \( q_t \approx 9.1\% \), using measurements in the \( x-z \) plane at angles \( \theta_{A_0} = 0.4187, \theta_{A_1} = 1.7900, \theta_{B_0} = 0.8636, \theta_{B_1} = 2.6340 \) on the state \( |\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2} \). This makes \( A_0B_0 \) more strongly correlated as compared to the previously described measurements, with only slightly poorer bounds on \( d(\rho_{E|00}, \rho_{E|11}) \). Against a limited-detector-efficiency noise model, this \( \Pr_{\text{target}} \) achieves \( \eta_t \approx 90.0\% \). Trying to choose \( \Pr_{\text{target}} \) such that \( A_0B_0 \) are perfectly correlated seems to reduce the CHSH value sufficiently to make the noise robustness worse than the cases considered above, though some optimisation reveals that \( \theta_{A_0} = \theta_{B_0} = 0, \theta_{A_1} = \pi/3, \theta_{B_1} = 2\pi/3 \) allows \( q_t \approx 7.1\%, \eta_t \approx 92.3\% \) (these measurement angles achieve a CHSH value of exactly 2.5). These threshold values were verified at NPA level 5, though we found no difference from the values at NPA level 2. The measurement angles were found using a fairly crude numerical optimisation, so they may not be the true optimal values.

Intuitively, it would seem that having the additional perfectly correlated key-generating measurement for Bob, as in scenario (i), can only improve on these results. However, we observed a surprising behaviour of the key-generating measurements \( A_0, B_0 \) in that scenario, namely that the bound on \( d(\rho_{E|00}, \rho_{E|11}) \) given by the SDP method becomes trivial even when there is still a CHSH violation by the measurements \( A_0, A_1, B_1, B_2 \). This contrasts with the 2-input 2-output protocol in Fig. 1, where the bound on \( d(\rho_{E|00}, \rho_{E|11}) \) is nontrivial whenever the CHSH violation is nonzero. A possible explanation is that since \( B_0 \) is not critical for Bell violation in scenario (i), it is constrained less strongly and so Eve can have more information about its outcome (see also Appendix D) where we outline a potential attack that achieves \( d(\rho_{E|00}, \rho_{E|11}) = 1 \) for \( q \geq 12.8\% \) in scenario (i), if the users only measure \( \epsilon \) and the CHSH value). Alternatively, it may be that the level of the NPA hierarchy which we used does not provide a tight bound in that scenario. Still, the noise threshold at which the \( d(\rho_{E|00}, \rho_{E|11}) \) bound becomes trivial may turn out to be not too closely related to the thresholds where Eq. (2) or Eq. (22) are satisfied, so this may not be a significant issue.

In any case, the proof technique in this section cannot be directly applied to any of the scenarios in Table I because Jordan's lemma cannot be applied simultaneously to more than two measurements. It would be of interest to find a way to work around this difficulty, perhaps by further study of when the analysis can essentially be reduced to states satisfying Eq. (13). For this purpose, it may be worth noting that pure states are not the only states satisfying Eq. (13) — for instance, if \( \rho_{E|00} \) and \( \rho_{E|11} \) are qubit states, the equality also holds if and only if they have the same eigenvalues (see Appendix E).

\[ \text{By simply replacing the error-correction term } H(A_0|B_0) = h_2(q) \text{ with } H(A_0|B_0) = h_2(\epsilon). \]
key-generating measurements way error correction, for instance in [3, 30, 31]. Specifically, in this proof we must consider the “security” of both output of the security of bounding the information leakage via the number of bits yields too crude a bound. The need to explicitly consider devices. For the repetition-code protocol, however, a large number of bits are publicly communicated, and hence to Eve is bounded simply by the length of the string, which is fixed based on the expected honest behaviour of the devices. For the repetition-code protocol, however, a large number of bits are publicly communicated, and hence bounding the information leakage via the number of bits yields too crude a bound. The need to explicitly consider the security of $B_0$ in this protocol can also be seen by considering an extreme example where Eve always knows the outcome of $B_0$, possibly at the cost of it being poorly correlated to $A_0$. In that case, regardless of how secure the output of $A_0$ is, Eve will always know the value of $C'$, making it impossible to distill key from the $(C, C')$ pairs. Hence any security proof for this protocol must involve some kind of security argument regarding $B_0$, even if only indirectly via measuring its correlations with $A_0$.

In Sec. II, the use of the Fuchs-van de Graaf inequality to bound the fidelity in Eq. (11) is likely not optimal, given the fairly large gap between the upper and lower bounds of the inequality. It might potentially be possible to directly bound $F(\rho_{E|00}, \rho_{E|11})$ by relating it to higher moments of the NPA hierarchy, in the vein of one approach for robust self-testing [22]. Alternatively, one could perhaps use the fact that for any pair of states $(\rho, \sigma)$, there exists a measurement such that the resulting outcome distributions $(P, Q)$ satisfy $F(P, Q) = F(\rho, \sigma)$ [33]. This effectively allows a reduction to classical side-information, in which case the minimum value of $F(P, Q)$ given the observed $Pr_{AB|XY}$ can be phrased as an optimisation problem analogous to [22]. The difficulty here is that it is unclear whether this optimisation has a useful dual that allows us to securely lower-bound the minimum value. Also, in device-independent protocols there is no a priori bound on the dimension of Eve’s states and hence the number of outcomes required for the measurement attaining $F(P, Q) = F(\rho, \sigma)$.

For the repetition-code protocol in device-dependent QKD, Eq. (2) was found to be not just a sufficient condition, but also necessary [14]. It may be of interest to investigate whether this is also the case for DIQKD, or construct attacks that upper-bound the noise tolerance of various scenarios (see for instance Appendix D). However, it may be more fruitful to instead consider other advantage-distillation protocols [15] [16], and find whether they lead to better noise tolerances. Also, in Section IV, we have not performed a very extensive search of possible choices for $Pr_{\text{target}}$, so there is still room for exploration of other target distributions.

Apart from noise tolerance, another important parameter in DIQKD is the key rate. The repetition-code protocol tends to produce very low key rates, since the fraction of accepted blocks is only $e^\epsilon + (1-e)^n$, and even in the accepted
blocks the asymptotic key rate (see Eq. (1)) is usually small. In this work we have mainly aimed to find only the conditions under which the key rate is positive, but there has been some work on improving key rates for advantage distillation in device-dependent QKD [15], and it would be useful to know whether this can be extended to DIQKD.

The analysis in this work was based on the assumption that the scenario is IID. However, there is a non-IID attack that would seem to be problematic if the noise is high enough for a large block size $n$ to be necessary, especially if the parameter estimation is imprecise. Namely, Eve can occasionally send a state where she perfectly knows the outcome of $A_0 B_0$, doing so often enough such that (with high probability) at least one such round is included in each block of the repetition-code protocol. In that case, Eve would know the values of $C$ and $C'$ in all blocks that include such a round, and the attack would disrupt the observed frequencies by only $O(1/n)$. It currently seems unclear how to provide a security proof of the repetition-code protocol in DIQKD against non-IID attacks, or whether the analysis could somehow be reduced to the IID scenario.

Finally, an open question in information theory is the existence of bound information, referring to correlations which require secret bits to be produced but from which there is no known method to distill a secret key, even with the use of advantage distillation [16,34]. There is a simple analogue to this in the context of DIQKD, namely whether there exist correlations which violate a Bell inequality but cannot be distilled into a secret key in a DI setting. Similar to the device-dependent case, our results here have a substantial gap between the threshold at which we can no longer prove the security of the protocol ($q \approx 9.1\%$) and the threshold at which the CHSH violation becomes zero ($q \approx 14.6\%$). It would be of interest to find whether this gap can be closed.

V. ACKNOWLEDGEMENTS

We thank Marco Tomamichel for highlighting the potential non-IID attack, as well as Srijita Kundu, Joseph Renes, Valerio Scarani and Le Phuc Thinh for helpful discussions. E. Y.-Z. Tan and R. Renner were funded by the Swiss National Science Foundation via the National Center for Competence in Research, QSIT, and by the Air Force Office of Scientific Research (AFOSR) via grant FA9550-16-1-0245. C.-W. Lim acknowledges support from the National Research Foundation (Singapore), the Ministry of Education (Singapore), the National University of Singapore, and the Asian Office of Aerospace Research and Development. Computations were performed with the NPAHierarchy function in QETLAB [35], using the CVX package [36,37] with solver SDPT3.

[1] S. Pironio, A. Acín, N. Brunner, N. Gisin, S. Massar, and V. Scarani, New J. Phys. 11, 045021 (2009)
[2] V. Scarani, Acta Physica Slovaca 62, 347 (2012)
[3] R. Arnon-Friedman, F. Dupuis, O. Fawzi, R. Renner, and T. Vidick, Nat. Commun. 9, 459 (2018)
[4] A. K. Ekert, Phys. Rev. Lett. 67, 661 (1991)
[5] J. Barrett, R. Colbeck, and A. Kent, Phys. Rev. Lett. 110, 010503 (2013)
[6] C. H. Bennett and G. Brassard, in Proceedings of International Conference on Computers, Systems and Signal Processing (1984) p. 175.
[7] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, Rev. Mod. Phys. 81, 1301 (2009)
[8] C.-H. F. Fung, B. Qi, K. Tamaki, and H.-K. Lo, Phys. Rev. A 75, 032314 (2007)
[9] I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Sláma, C. Kurtsiefer, and V. Makarov, Nat. Commun. 2, 349 (2011)
[10] N. Jain, B. Stiller, I. Khan, V. Makarov, C. Marquardt, and G. Leuchs, IEEE Journal of Selected Topics in Quantum Electronics 21, 168 (2015)
[11] U. M. Maurer, IEEE Transactions on Information Theory 39, 733 (1993)
[12] S. Wolf, Information-Theoretically and Computationally Secure Key Agreement in Cryptography, Ph.D. thesis ETH Zürich (1999).
[13] R. Renner, Security of Quantum Key Distribution, Ph.D. thesis ETH Zürich (2005).
[14] J. Bae and A. Acín, Phys. Rev. A 75, 012334 (2007)
[15] S. Watanabe, R. Matsumoto, T. Uyematsu, and Y. Kawano, Phys. Rev. A 76, 032312 (2007)
[16] S. Khatri and N. Lütkenhaus, Phys. Rev. A 95, 042320 (2017)
[17] I. Devetak and A. Winter, Proc. Roy. Soc. A: Math Phys. 461, 207 (2005).
[18] A. Winter, Commun. Math. Phys. 347, 291 (2016).
[19] W. Roga, M. Fannes, and K. Życzkowski, Phys. Rev. Lett. 105, 040505 (2010).
[20] J. Briët and P. Harremoës, Phys. Rev. A 79, 052311 (2009).
[21] C. A. Fuchs and J. van de Graaf, IEEE Transactions on Information Theory 45, 1216 (1999).
[22] L. P. Thinh, G. de la Torre, J.-D. Bancal, S. Pironio, and V. Scarani, New Journal of Physics 18, 035007 (2016).
[23] M. Navascués, S. Pironio, and A. Acín, New J. Phys. 10, 073013 (2008).
[24] P. H. Eberhard, Phys. Rev. A 47, R747 (1993).
[25] L. Hardy, Phys. Rev. Lett. 71, 1665 (1993).
[26] R. Rabelo, L. Y. Zhi, and V. Scarani, Phys. Rev. Lett. 109, 180401 (2012).
[27] D. Mayers and A. Yao, Quantum Inf. Comput. 4, 273 (2004).
[28] N. Gisin, arXiv:quant-ph/0702021v2 (2007).
[29] O. Andersson, P. Badziąg, I. Bengtsson, I. Dumitru, and A. Cabello, Phys. Rev. A 96, 032119 (2017).
[30] U. Vazirani and T. Vidick, Phys. Rev. Lett. 113, 140501 (2014).
[31] M. Tomamichel and A. Leverrier, Quantum 1, 14 (2017).
[32] T. H. Yang, T. Vértesi, J.-D. Bancal, V. Scarani, and M. Navascués, Phys. Rev. Lett. 113, 040401 (2014).
[33] C. A. Fuchs and C. M. Caves, Open Syst. Inf. Dyn. 3, 345 (1995).
[34] N. Gisin and S. Wolf, in Advances in Cryptology — CRYPTO 2000 edited by M. Bellare (Springer Berlin Heidelberg, Berlin, Heidelberg, 2000) pp. 482–500.
[35] N. Johnston, “QETLAB: A MATLAB toolbox for quantum entanglement, version 0.9,” http://qetlab.com (2016).
[36] M. Grant and S. Boyd, “CVX: Matlab Software for Disciplined Convex Programming, version 2.1,” http://cvxr.com/cvx (2014).
[37] M. Grant and S. Boyd, in Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, edited by V. Blondel, S. Boyd, and H. Kimura (Springer-Verlag Limited, 2008) pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.
[38] M. Hübner, Phys. Lett. A 163, 239 (1992).
Appendix A: Large-$n$ limits

We first note that letting $\phi(\delta) = h_2(\delta/(1 + \delta))$, we have $\phi(\delta), h_2(\delta) \to 0$ and $\phi'(\delta), h_2'(\delta) \to +\infty$ as $\delta \to 0^+$, so
\[
\lim_{\delta \to 0^+} \frac{\delta}{h_2(\delta)} = \lim_{\delta \to 0^+} \frac{1}{h_2'(\delta)} = 0,
\]
\[
\lim_{\delta \to 0^+} \frac{\phi(\delta)}{h_2(\delta)} = \lim_{\delta \to 0^+} \frac{\phi'(\delta)}{h_2'(\delta)} = \lim_{\delta \to 0^+} \frac{(1 + \delta - 2\delta \ln \delta)/(\delta(1 + \delta)^2)}{1/(\delta(1 - \delta))} = 1.
\]

Since $\delta_n \to 0^+$ as $n \to \infty$, the terms arising from the continuity bound hence have a finite limit,
\[
\lim_{n \to \infty} \left(-\delta_n - (1 + \delta_n)h_2\left(\frac{\delta_n}{1 + \delta_n}\right)\right) h_2(\delta_n)^{-1} = -1.
\]

Let $\beta = \epsilon/(1 - \epsilon) \in [0, 1)$. Then $\delta_n \leq \beta^n$, and for sufficiently large $n$ we have $\beta^n < 1/2$, so $h_2(\delta_n) \leq h_2(\beta^n) \leq -2\beta^n \log \beta^n$. Hence for any $\alpha > \beta$,
\[
\lim_{n \to \infty} \frac{\alpha^n}{h_2(\delta_n)} \geq \lim_{n \to \infty} \frac{\alpha^n}{2n\beta^n \log(1/\beta)} = +\infty.
\]

The claim in the proof of Theorem 1 then follows from the above results, by using $h_2((1 - p)/2) \leq 1 - p^2/\ln 4$ to get $1 - h_2\left((1 - F(\rho_{E|00}, \rho_{E|11})^n)/2\right) \geq F(\rho_{E|00}, \rho_{E|11})^{2n}/\ln 4$ and choosing $\alpha = F(\rho_{E|00}, \rho_{E|11})^2$ in the above limit. As for Theorem 2 we can just choose $\alpha = 1 - f(Pr_{AB|XY})/(1 - \epsilon)$ directly.

Appendix B: Effect of symmetrisation

The argument essentially follows the reduction to Bell-diagonal states in [1], with slight modifications to account for the fact that we are bounding fidelity rather than entropy. Denote the single-round state before symmetrisation as
\[
\rho_{\text{sym}} = \sum_{x,y} \Pr(x,y) |x\rangle\langle y| \otimes \sigma_{E|xy},
\]
where all $\sigma_{E|xy}$ are pure states, and $\Pr_{\text{sym}}$ refers to the initial probability distribution before symmetrisation. After the symmetrisation is carried out via a publicly communicated bit $T$, we have the state
\[
\rho_{\text{sym}} = \sum_{a,b} \Pr_{\text{sym}}(a,b) |a\rangle\langle b| \otimes \rho_{E|ab} \otimes |\lambda\rangle\langle \lambda|,
\]
where all $\rho_{E|ab}$ are pure states, and $\Pr_{\text{sym}}$ refers to the initial probability distribution before symmetrisation. After the symmetrisation is carried out via a publicly communicated bit $T$, we have the state
\[
\rho_{\text{sym}} = \sum_{a,b} \Pr_{\text{sym}}(a,b) |a\rangle\langle b| \otimes \rho_{E|ab} \otimes |\lambda\rangle\langle \lambda|,
\]
where $\Pr_{\text{sym}}(a,b) = \frac{1}{2} \Pr_{T}(a \oplus t,b \oplus t|xy\lambda)$, $
\rho_{E|ab} = \sum_{t} \Pr_{T}(a \oplus t,b \oplus t|xy\lambda) \sigma_{E|a \oplus t,b \oplus t} \otimes |t\rangle\langle t|.
\]

The probabilities $\Pr_{\text{sym}}(a,b)$ satisfy $\sum_{a,b} \Pr_{\text{sym}}(a,b)\Pr(\lambda) = \Pr(ab|xy\lambda)$, and for each value of $\lambda$, we have $\Pr(00|00\lambda) = \Pr(11|00\lambda)$. In the proof of Theorem 2, we should be considering the state $\rho$ with Eve having access to $E$ and $T$, and eventually need to bound $F(\rho_{E|00}, \rho_{E|11})$, but we would face the problem that the states $\rho_{E|ab}$ are not pure. However, we shall now argue that there exists a state $\rho'$ which produces the same probabilities $\Pr(ab|xy\lambda)$ and achieves $F(\rho_{E|00}, \rho_{E|11}) = F(\rho_{E|00}, \rho_{E|11})$ with pure conditional states $\rho'_{E|ab}$, essentially by replacing the classical register $T$ with an appropriate purification. Therefore, we can consider the state $\rho'$ instead of the state $\rho$ when bounding $F(\rho_{E|00}, \rho_{E|11})$.

Denoting the pure pre-measurement states $\sigma_{ABE|\lambda}$ as $|\psi\rangle\langle \psi|_{ABE|\lambda}$, consider the state
\[
\rho'_{ABETA} = \sum_{\lambda} \Pr(\lambda) |\psi'\rangle\langle \psi'|_{ABE|\lambda} \otimes |\lambda\rangle\langle \lambda|,
\]
where $|\psi'\rangle_{ABE|\lambda} = \frac{1}{\sqrt{2}} \left( |\psi\rangle_{ABE|\lambda} \otimes |0\rangle + (Y \otimes \text{id}) |\psi\rangle_{ABE|\lambda} \otimes |1\rangle \right)$. 


and let Alice and Bob’s devices perform the same measurements as they did on σ. Denote the post-measurement states after measuring A_x B_y on |ψ⟩_{ABE|Λ} and getting outcome ab as |ψ⟩_{E|ab}. Since Y ⊗ Y acting on Alice and Bob’s systems has the effect of flipping the outcomes of all Pauli measurements in the x-z plane, the post-measurement state can be found to be

\[ \rho'_{A_x B_y E \Lambda} = \sum_{a,b} \sum_{\Lambda} \Pr(ab|xy\lambda) \Pr(\lambda) |a,b\rangle\langle a,b| \otimes |\psi^\prime\rangle_{E|ab} \otimes |\lambda\rangle\langle \lambda|, \]

where \( \Pr(ab|xy\lambda) = \sum_t \frac{1}{2} \Pr_t(a \oplus t, b \oplus t|xy\lambda) \), \( |\psi^\prime\rangle_{ET|ab} = \sum_t \sqrt{\frac{\Pr_t(a \oplus t, b \oplus t|xy\lambda)}{\Pr(ab|xy\lambda)}} |\psi^\prime\rangle_{E|\lambda,ab} \otimes |t\rangle . \)

This state achieves the same probabilities \( \Pr(ab|xy\lambda) \) as \( \rho \) in Eq. (B2). In addition, we can verify that the pure states \( \rho'_{ET|ab} = |\psi^\prime\rangle_{ET|ab} \otimes |\lambda\rangle\langle \lambda| \) satisfy \( F(\rho'_{ET|00},\rho'_{ET|11}) = F(\rho_{ET|00},\rho_{ET|11}) \). Therefore, we can consider this \( \rho' \) in place of \( \rho \).

**Appendix C: Concavity of \( \hat{f} \)**

The argument is essentially the same as that for concavity of guessing-probability bounds, but modified to account for the “postselection” on the outcomes being 00 or 11. Let \( S \) be the set of quantum distributions \( \Pr_{AB|XY} \) such that \( \Pr(00|00) = \Pr(11|00) \). Consider any probability distribution \( \Pr(\lambda) \) and any family of distributions \( \Pr_{AB|XY\lambda} \in S \) indexed by \( \lambda \), and take an arbitrary \( \delta > 0 \). For each \( \lambda \), there exists a strategy for Eve that achieves the distribution \( \Pr_{AB|XY\lambda} \) and has \( (1 - \epsilon_\lambda)d(\rho_{E|00},\rho_{E|11}) \geq \hat{f}(\Pr_{AB|XY\lambda}) - \delta \), because \( \hat{f} \) is an optimal bound (i.e. there exist strategies arbitrarily close to saturating the bound). If Eve generates and stores a classical random variable \( \Lambda \) according to the distribution \( \Pr(\lambda) \), then implements the corresponding strategy, the resulting states take the same form as in Eq. (15), (16). This is a strategy that achieves probabilities \( \Pr(ab|xy) = \sum_\lambda \Pr(\lambda)\Pr(ab|xy\lambda) \), and therefore

\[ \hat{f}(\Pr_{AB|XY}) \geq (1 - \epsilon)d(\rho_{E|00},\rho_{E|11}) \]
\[ = (1 - \epsilon) \sum_\lambda \Pr(\lambda)\rho_{E|00} \otimes |\lambda\rangle\langle \lambda|, \sum_\lambda \Pr(\lambda)\rho_{E|11} \otimes |\lambda\rangle\langle \lambda| \]
\[ = (1 - \epsilon) \sum_\lambda \Pr(\lambda)d(\rho_{E|00},\rho_{E|11}) \]
\[ = \sum_\lambda \Pr(\lambda)(1 - \epsilon_\lambda)d(\rho_{E|00},\rho_{E|11}) \]
\[ \geq \sum_\lambda \Pr(\lambda)\hat{f}(\Pr_{AB|XY\lambda}) - \delta. \]

Since \( \delta \) was arbitrary, we conclude that \( \hat{f}(\Pr_{AB|XY}) \geq \sum_\lambda \Pr(\lambda)\hat{f}(\Pr_{AB|XY\lambda}) \), i.e. \( \hat{f} \) is concave on \( S \).

**Appendix D: Attack description**

For scenario (i) of Table 1, subject to depolarising noise \( q \), we have \( \epsilon = q \) and \( S = 2\sqrt{2}(1 - 2q) \), where \( S \) denotes the CHSH value of \( A_0, A_1, B_1, B_2 \) when the outcomes are labelled as \pm 1. We shall show that for \( q \geq 1/(5 + 2\sqrt{2}) \), there exists an adversarial distribution \( \Pr_{ad}(ab|xy) \) that attains these values of \( \epsilon \) and \( S \), but allows Eve to perfectly distinguish the outcome 00 from the outcome 11 on the measurement pair \( A_0B_0 \). Consider a scenario where Eve generates a classical random variable \( \Lambda \in \{0,1,2,3\} \), and implements one of the following strategies based on the value of \( \Lambda \):

- \( \Lambda = 0 \): The devices implement the ideal singlet state and CHSH measurements for \( A_0, A_1, B_1, B_2 \), and \( B_0 \) is in the basis perfectly correlated to \( A_0 \). This gives \( S = 2\sqrt{2} \) and \( \epsilon = 0 \).

- \( \Lambda = 1 \): The devices implement the ideal singlet state and CHSH measurements for \( A_0, A_1, B_1, B_2 \), and \( B_0 \) is in the basis perfectly anticorrelated to \( A_0 \). This gives \( S = 2\sqrt{2} \) and \( \epsilon = 1 \).

- \( \Lambda = 2 \): The devices implement a classical (completely insecure) distribution attaining \( S = 2 \) and \( \epsilon = 0 \).
\[ \Lambda = 3: \text{The devices implement a classical (completely insecure) distribution attaining } S = 2 \text{ and } \epsilon = 1. \]

For \( \Lambda \in \{0, 1\} \), Eve’s state is independent of the outcome of \( A_0 B_0 \), while for \( \Lambda \in \{2, 3\} \), Eve perfectly knows the outcome of \( A_0 B_0 \). The resulting distribution \( \Pr_{\text{ad}}(ab|xy) \) observed by Alice and Bob attains values of \( S \) and \( \epsilon \) that are related to the distribution of \( \Lambda \) by \( S = 2\sqrt{2}(\Pr_{\Lambda}(0) + \Pr_{\Lambda}(1)) + 2(\Pr_{\Lambda}(2) + \Pr_{\Lambda}(3)), \epsilon = \Pr_{\Lambda}(1) + \Pr_{\Lambda}(3) \). We now note that if \( \Pr_{\Lambda}(0) = 0 \), then whenever the outcome of \( A_0 B_0 \) is 00 or 11, it must have been the case that \( \Lambda \in \{2, 3\} \) (since the outcome of \( A_0 B_0 \) is always 01 or 10 when \( \Lambda = 1 \)), in which case the devices have implemented a completely insecure distribution and so Eve knows the outcome perfectly. Finally, we observe that it is indeed possible to have \( \Pr_{\Lambda}(0) = 0 \) with \( \epsilon = q \) and \( S = 2\sqrt{2}(1 - 2q) \), by choosing

\[
\Pr_{\Lambda}(1) = 1 - 2(2 + \sqrt{2})q, \quad \Pr_{\Lambda}(2) = 1 - q, \quad \Pr_{\Lambda}(3) = (5 + 2\sqrt{2})q - 1. \tag{D1}
\]

All these values are nonnegative when \( q \in [1/(5 + 2\sqrt{2}), 1/(2(2 + \sqrt{2}))] \approx [12.8\%, 14.6\%] \), and hence form a valid probability distribution. Therefore, Eve can distinguish the 00 and 11 outcomes perfectly for \( q \geq 12.8\% \) (the value \( q \approx 14.6\% \) is simply the noise value at which \( S = 2 \), so for higher noise values there is already no CHSH violation and no certification of security).

However, this attack focuses only on producing appropriate values of \( \epsilon \) and \( S \). It might already be possible to rule it out by considering the full distribution \( \Pr_{\Lambda|B|XY} \) instead. In addition, perfectly distinguishing the 00 and 11 outcomes is not precisely equivalent to a complete attack on the repetition-code protocol, since it is essentially ignoring what happens in the case where \( A_0 \neq B_0 \) in the accepted block, which occurs with small but nonzero probability \( \delta_n \). Intuitively, this should not affect the efficacy of the attack since \( \delta_n \) is small; also, \( C \neq C' \) for such a block so Bob’s guess \( C' \) will be wrong in any case. However, we do not further pursue this line of inquiry since this attack seems to be of limited scope. (Notice that this attack cannot be straightforwardly applied to the distributions in Sec. III because for those protocols the CHSH value includes \( \langle A_0 B_0 \rangle \) directly, and it is impossible to have \( S = 2\sqrt{2} \) and \( \epsilon = 0 \) simultaneously in that case.)

**Appendix E: Saturating the Fuchs-van de Graaf inequality**

For qubit states \( \rho, \sigma \) with Bloch vectors \( \vec{u}, \vec{v} \), we have the following formulas for root-fidelity \( \text{[38]} \) and trace distance:

\[
F(\rho, \sigma)^2 = \left( 1 + \vec{u} \cdot \vec{v} + \sqrt{1 - \vec{u}^2} \sqrt{1 - \vec{v}^2} \right) / 2, \quad d(\rho, \sigma)^2 = |\vec{u} - \vec{v}|^2 / 4 = (u^2 + v^2 - 2\vec{u} \cdot \vec{v})/4, \tag{E1}
\]

where \( u = |\vec{u}|, v = |\vec{v}| \). Therefore,

\[
F(\rho, \sigma)^2 + d(\rho, \sigma)^2 = \left( 2 + 2\sqrt{1 - u^2} \sqrt{1 - v^2} + u^2 + v^2 \right) / 4, \tag{E2}
\]

and solving for \( F(\rho, \sigma)^2 + d(\rho, \sigma)^2 = 1 \) yields the unique solution \( u^2 = v^2 \). Hence we conclude that for qubit states, we have \( F(\rho, \sigma)^2 + d(\rho, \sigma)^2 = 1 \) if and only if \( \rho \) and \( \sigma \) have Bloch vectors of the same length, which is equivalent to \( \rho \) and \( \sigma \) having the same eigenvalues.

This property does not seem to extend to non-qubit states: for instance, the family of qutrit states \( \rho = p\ket{0}\bra{0} + (1 - p)\ket{1}\bra{1}, \sigma = p\ket{0}\bra{0} + (1 - p)\ket{2}\bra{2} \) with \( p \in [0, 1] \) instead satisfies \( F(\rho, \sigma) = p \) and \( d(\rho, \sigma) = 1 - p \), which saturates the opposite bound in the Fuchs-van de Graaf inequality.