Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 47 (2014)

An Interactive Information Odometer with Applications

Mark Braverman* Omri Weinstein†

March 10, 2017

Abstract

We introduce a novel technique which enables two players to maintain an estimate of the internal information cost of their conversation in an online fashion without revealing much extra information. We use this construction to obtain new results about communication complexity and information-theoretic privacy.

As a first corollary, we prove a strong direct product theorem for communication complexity in terms of information complexity: If I bits of information are required for solving a single copy of f under μ with probability $2/3$, then any protocol attempting to solve n independent copies of f under μ^n using $o(n \cdot I)$ communication, will succeed with probability $2^{-\Omega(n)}$. This is tight, as Braverman and Rao [BR11] previously showed that $O(n \cdot I)$ communication suffice to succeed with probability $\sim (2/3)^n$.

We then show how the information odometer can be used to achieve the best possible information-theoretic privacy between two untrusted parties: If the players' goal is to compute a function $f(x, y)$, and f admits a protocol with information cost is I and communication cost C, then our odometer can be used to produce a "robust" protocol which: (i) Assuming both players are honest, computes f with high probability, and (ii) Even if one party is malicious, then for any $k \in \mathbb{N}$, the probability that the honest player reveals more than $O(k \cdot (I + \log C))$ bits of information to the other player is at most $2^{-\Omega(k)}$.

Finally, we outline an approach which uses the odometer as a proxy for breaking state of the art interactive compression results: We show that our odometer allows to reduce interactive compression to the regime where $I = O(\log C)$, thereby opening a potential avenue for improving the compression result of [BBCR10] and to new direct sum and product theorems in communication complexity.

1 Introduction

Overview of the main technical construction

In this paper we consider the following problem. Alice and Bob are given inputs $x, y \sim \mu$, and are executing a communication protocol π. During the course of the execution of π they wish to maintain an information odometer — an *online* estimate (say, within a factor of 2) of the amount of information they have revealed to each other about their inputs. In more technical terms, they

*Department of Computer Science, Princeton University, mbraverm@cs.princeton.edu, Research supported in part by an Alfred P. Sloan Fellowship, NSF CCF-0816797, an NSF CAREER award (CCF-1149888), a Turing Centenary Fellowship, and a Packard Fellowship in Science and Engineering.

†Department of Computer Science, Princeton University, oweinste@cs.princeton.edu. Research supported by a Simons Fellowship in Theoretical Computer Science.

ISSN 1433-8092

wish to maintain an estimate on the internal information cost of the protocol π so far. Moreover, since applications of this primitive involve limiting the amount of information revealed, they wish to implement it without revealing too much additional information about their inputs in the process. Ideally, the information overhead of implementing it up to any point in time should scale with the information cost of π so far. In this paper, we introduce a technique that enables such an implementation.

Before discussing applications, let us discuss the challenges in implementing such an information odometer. Firstly, we note that even if the original protocol π does not involve interaction, estimating information revealed requires interaction[1].

The fact that interaction is required means that no "simple" unilateral solution (where Alice and Bob keep some counters separately) is possible, and makes a generic information odometer more difficult to construct. Luckily, while the protocol π can be quite complex, we can always break it down into the individual bits that are being transmitted. Therefore, we can focus on estimating the amount of information transmitted in a single bit sent, say, from Alice to Bob. The distribution of Alice's message M in this case is described by one number $p = \Pr[M = 1|\text{history}, X = x] \in (0, 1)$, such that her message is given by the Bernoulli distribution B_p: 1 with probability p and 0 with probability $1 - p$. For technical convenience, we will only focus on the case when $p \in (1/3, 2/3)$ — this can be done essentially without loss of generality, since the sending of a highly-biased bit can be simulated by the majority of several, slightly-biased bits, without increasing the information cost of the protocol (see Lemma 23). Note that the value of the probability p depends on Alice's input x, as well as on the transcript so far. The actual sampling of $M \sim B_p$ is done using Alice's private randomness.

What does Bob learn about x from a message $M \sim B_p$? Not surprisingly, the answer depends on what Bob already knows. More specifically, it is given by the Kullback-Leibler divergence between the actual distribution of M, and Bob's belief about this distribution. Note that since $M \in \{0, 1\}$ is a binary message, Bob's belief is given by a Bernoulli variable B_q (where $q = \Pr[M = 1|\text{history}, Y = y]$). Since $p \in (1/3, 2/3)$, we must also have $q \in (1/3, 2/3)$. The amount of information learned by Bob is given by $\mathsf{D}\left(\frac{B_p}{B_q}\right)$. For $p, q \in (1/3, 2/3)$ it is the case that $\mathsf{D}\left(\frac{B_p}{B_q}\right) = \Theta((p - q)^2)$. In particular, Bob learns nothing if $q = p$ (i.e. if he already knows p). Therefore, the odometer problem reduces to the task of estimating $I := (p - q)^2$, while revealing not much more than I bits of information to the players in the process. More specifically, we show how to sample a Bernoulli random variable $B_{(p-q)^2}$, while revealing at most $O(H((p - q)^2)) = O((p - q)^2 \log 1/(p - q)^2)$ bits of information. While this quantity is more than $(p - q)^2$ by a $\log 1/(p - q)$ factor, this will be sufficient for most applications. Our test produces an (essentially) unbiased estimator on the amount of information revealed in a given round. By running this estimator on a *subsample* of the rounds, rather than on all the rounds of π, we can keep the overhead below the information cost of π itself, while maintaining a good unbiased estimate of the amount of information revealed so far.

We have therefore reduced the odometer problem to the following scenario. Alice and Bob

[1]Consider the following simple scenario. Alice is given a sequence of blocks $X_1, X_2, \ldots, X_k$ and a subset $S \subset \{1, \ldots, k\}$. Bob is also given a sequence of blocks $Y_1, \ldots, Y_k$ and a subset $T \subset \{1, \ldots, k\}$ for $i \in T$, $X_i = Y_i$, and for $i \notin T$, X_i and Y_i are statistically independent. In the protocol π, Alice performs the following action: For each $i \in [k]$ she sends the block X_i if $i \in S$, and sends a random block R_i otherwise. Thus π is a one-round protocol. The amount of information revealed by π is proportional to $|S \setminus T|$, and the amount of information revealed by the first t blocks is proportional to $a_t := |(S \setminus T) \cap \{1, \ldots, t\}|$. Note that maintaining an estimate on a_t requires the parties to compute $S \setminus T$, which would require Alice and Bob to interact.

are given numbers $p \in (1/3, 2/3)$ and $q \in (1/3, 2/3)$, respectively. Their goal is to sample $B_{(p-q)^2}$, while revealing at most $O(H(B_{(p-q)^2}))$ information to each other. The simplest strategy that clearly doesn't work is to have Alice send Bob p and have Bob sample $B_{(p-q)^2}$ (or vice versa). This does not work since p may reveal many bits of information about x (and q may reveal many bits of information about y). A slightly less naïve approach is based on the idea of correlated sampling of [Hol07]. We can sample a number $Z \in_U [0, 1]$ uniformly at random. Alice and Bob then exchange information on whether $p < Z$ and $q < Z$, respectively. If the answers do not match, they output 1, otherwise they output 0. It is not hard to see that this procedure produces a sample from the distribution $B_{|p-q|}$. By repeating it twice and outputing the conjunction of the two answers, we can get a sample from $B_{|p-q|^2}$. Unfortunately, it is not hard to see that this procedure may reveal as much as $\Omega(H(B_{|p-q|})) = \Omega(|p-q| \log 1/|p-q|)$ to the parties, which is prohibitively high.

Our approach is based on the correlated sampling above. Instead of Z being chosen using public randomness, Z is chosen by Alice from a distribution Z_p which depends on the value of p. Alice then sends Z_p to Bob. The distribution Z_p is designed to meet the following two conditions: (1) a sample $Z \sim Z_p$ reveals at most $O(H(B_{(p-q)^2}))$ bits of information about p (and thus about x) to someone who knows q; (2) the probability that Z falls between p and q is $\sim (p-q)^2$ (note that for $Z \in_U [0, 1]$ this probability was $\sim |p-q|$). Satisfying these two conditions allow us to sample from $B_{(p-q)^2}$ by seeing whether Z falls between p and q (using condition (2)). Condition (1) ensures that the value of Z does not reveal too much information to Bob about x in the process. Our choice of the distribution of Z draws inspiration from Raz's counter-example to the strong parallel repetition conjecture [Raz11].

As discussed above, we primarily apply this basic primitive as follows. At each step i of π we execute the protocol above with some probability α, obtain a sample $S_i \sim B_{(p-q)^2}$, and maintain the sum Σ_i of the S_i's so far. This way, if I_i^π is the amount of information revealed by π by round i, we have that Σ_i is an unbiased estimator of $\alpha \cdot I_i^\pi$. Therefore Σ_i implements an information odometer for π. While Σ_i is stochastic, by choosing $\alpha < 1$ that is not too small, we can also ensure that Σ_i has sufficiently nice concentration properties for our applications we discuss below.

Applications

Conditional abort and applications to direct product theorems in communication complexity Our first application is to proving a direct product theorem for communication complexity in terms of information complexity. Direct sum and product theorems have had a long history in the area of communication complexity [Kla10, Sha03, LSS08, She11, JPY12, MWY13, PRW97, BBCR10]. For a broader overview of the problem and its importance in computational complexity we refer the reader to [JPY12, BRWY12] and references therein. Both direct sum and direct product theorems assert a lower bound on the communication complexity of solving n copies of f in terms of the cost of a single copy (or, potentially, in terms of another quantity related to f). A direct sum theorem aims to give a lower bound (ideally one linear in n) on computing n copies of f with error at most $\varepsilon > 0$ in terms of the cost of computing a single copy of f with error ε. A direct product theorem further asserts that unless sufficient communication resources are provided, the probability of successfully computing all n copies of f will be exponentially small, potentially as low as $(1-\varepsilon)^{\Omega(n)}$.

In the context of randomized communication, there is a tight connection between the direct sum question about the communication complexity of a function f and its information complexity. Specifically, it was shown in [BR11] that the communication cost of computing n copies of f with

error at most ε per copy scales as n times the information complexity $\mathsf{IC}(f,\varepsilon)$ of computing f with error at most ε — that is, the amount of information Alice and Bob must reveal to each other to compute f with error at most ε. While understanding the gap between the information complexity of f and its communication complexity remains open, at least in terms of information complexity, the question is settled. To get a lower bound on the amortized communication complexity of f^n one starts with a communication-C_n protocol for f^n, which makes an error of at most ε *on each coordinate*, and shows how to convert it into a protocol for a single copy of f with error of at most ε and *information cost* of at most C_n/n.

The upper bound in [BR11] not only shows that the cost of computing n copies of f in parallel with error $\leq \varepsilon$ on each copy is $\sim n \cdot \mathsf{IC}(f,\varepsilon)$, but since it does so by executing independent copies in parallel, its probability of success on all copies simultaneously is $\approx (1-\varepsilon)^n$. Therefore, the best direct product theorem we could hope for is in terms of the information complexity of an individual copy of f: "a protocol which uses $\ll n \cdot \mathsf{IC}(f,\varepsilon)$ communication to solve n copies of f cannot succeed with probability more than $(1-\varepsilon)^{\Omega(n)}$".

Several prior works (e.g [Jai11, JPY12, BRWY12]) aim to get a generic direct product theorem for communication complexity. Other works prove a direct product theorem in terms of weaker complexity measures of the underlying function, such as the discrepancy $\mathsf{disc}_\mu(f)$ of the function ([LSS08]) or the (stronger) smooth rectangle bound [JK09]. More precisely, Jain and Yao [JY12] show that any protocol attempting to compute f^n under μ^n using $\ll n \cdot \mathsf{srect}_\mu(f)$ communication, will succeed with probability only $2^{-\Omega(n)}$, where $\mathsf{srect}_\mu(f)$ denotes the smooth rectangle bound of f under μ. Our direct product theorem implies all previous results in this category, since it has been shown that $\mathsf{IC}_\mu(f) \geq \mathsf{srect}_\mu(f) \geq \mathsf{disc}_\mu(f)$ (see [KLL$^+$12]). Moreover, the discussion in the previous paragraph asserts that our direct product result (Theorem 2) is asymptotically tight (as communication and information are asymptotically equal), while such guarantee is not known to hold for the previous measures.

The most directly relevant effort on the direct product problem was carried out by [JPY12] (for the bounded round case) and [BRWY12] (for general protocols), which aim to give a direct product theorem in terms of the information/ communication complexity of f. In terms of their logical flow, these papers follow earlier ideas in parallel repetition theorems [Raz98, Hol07, Rao08] and proceed as follows: starting with a low-communication protocol for f^n, and assuming its success probability is high — at least $(1-\varepsilon)^{o(n)}$, one simultaneously applies the same method as in the direct sum theorem, *and* conditions on the event that the n-copy protocol is successful. This latter conditioning leads to the resulting object not being a communication protocol any longer. However, with some additional work, one gets *a protocol that computes f and is statistically close to a low-information protocol*. To get the full direct product theorem for information complexity one would need to get *a protocol that computes f and is a low-information protocol*.

A protocol π that is statistically close to a low-information needs not be a low-information protocol itself. Consider, for example, a protocol π where with probability δ Alice sends her input $X \in \{0,1\}^n$ to Bob, and with probability $1-\delta$ she sends a random string. Then π is δ-close to a 0-information protocol, but has information complexity of $\approx \delta \cdot n$, which could be arbitrarily high. [BRWY12] showed that the previous protocol compression techniques that work for compressing low-information protocols [BBCR10] also work on protocols that are statistically close to a low-information protocol. Therefore, all direct sum results for communication complexity from [BBCR10] can be upgraded to direct product results. This, however, is weaker than a full direct product theorem in terms of information complexity.

To turn the [BRWY12] construction into a direct product theorem in terms of information complexity, one needs a generic way of turning a protocol that is statistically close to a low-information one into a low information protocol. Prior to the present paper, no such way was known. Note that the information odometer is precisely the primitive one can use for this purpose: if π is statistically close to a protocol that only reveals I bits of information, we can use the odometer and abort π after it reveals at most $100I$ bits of information. The fact that π is statistically close to an information-I protocol guarantees that we do not abort too frequently, and that this simulation succeeds. Note that we need the additional information complexity of the odometer to be $O(I)$, to make sure that indeed the information complexity of (truncated π + odometer) is $O(I)$. Putting this together with [BRWY12], we obtain an essentially optimal direct product theorem for communication complexity in terms of information complexity (Theorem 2 below).

Interactive computation between two untrusting parties: from honest-but-curious to malicious Next, we consider an application to the setting where Alice and Bob do not trust each other and wish to compute a function $f(X, Y)$ of their inputs while revealing as little information to each other as possible. This setting has been extensively studied in the theoretical cryptography literature. In the the case of 3+ parties with private channels (and honest majority), [BOGW88] showed that secure multiparty computation is possible, that is, it is possible to compute any function of the player's inputs while revealing nothing beyond the value of the function to the players. It is known that no such protocol can exist for two parties, even in the case of *honest-but-curious* participants. In this model, Chor and Kushilevitz [CK91] characterized the family of two-party Boolean functions computable with perfect privacy. This characterization was extended by Kushilevitz [Kus92] and Beaver [Bea89] to general-valued functions, asserting that most function are not privately computable. Subsequent papers studied the privacy loss of specific functions, and explored communication tradeoffs required to achieve perfect or approximate privacy in the honest model (Bar Yehuda et al [BYCKO93] [FJS10] [ACC$^+$12]).

In the *malicious* model, where one of the parties is assumed to be adversarial, much less was known. When the malicious party is assumed to be computationally bounded, and thus one can use cryptographic primitives, [GMW87] ensure the "best possible" privacy can be preserved, assuming the existence of so called "trapdoor permutations" [2]. Other works define a weaker notion of privacy and obtain privacy-preserving schemes for specific functions under these notions ([Pin03, MNPS]). None of these works has a pure statistical security guarantee against general, unrestricted adversaries.

As information-theoretically secure two-party computation is impossible for most functions, several approaches for quantifying privacy loss have been proposed over the years in the security and privacy literature [Kla02, FJS10, MMP$^+$10, KLX13]. In fact, one way to view the information complexity $\mathsf{IC}(f, \varepsilon)$ is as the smallest (average) amount of information Alice and Bob must reveal to each other to compute f with error ε (here the information revealed by the value of $f(X, Y)$ is included in the information complexity). Thus, information complexity gives the precise answer to the two-party private computation in the information-theoretic *honest-but-curious* model: Alice and Bob will try to learn about Y and X respectively from the protocol, while adhering to its prescribed execution.

Therefore, in the honest-but-curious case, a protocol π whose information cost is close to the

[2] The authors show that a malicious player cannot learn anything more that the value of $f(X', Y)$ for any X' of her choice.

information complexity of f will achieve a near-optimal performance in terms of privacy, revealing only $\approx I := \mathsf{IC}(f, \varepsilon)$ information to Alice and Bob. That is, assuming Alice and Bob adhere to the execution of π[3].

What happens when either Alice or Bob is malicious? There are easy examples where a cheating Bob can extract $\omega(I)$ bits of information on Alice's input, if the protocol is executed naively (an instructive example is the standard hashing protocol for the Equality function). Is there a way to compile π into a protocol π' such that (1) if Alice and Bob are honest is close to π in terms of computing f; (2) even if Alice or Bob are dishonest, reveals at most $O(I)$ information to the dishonest party (that is, a dishonest Bob cannot "phish" more than $O(I)$ bits of information out of Alice)? If information complexity was known to be equal to communication complexity, we could just compress π into a protocol π' with $O(I)$ bits of communication. Even if Alice or Bob are dishonest, they cannot cause the protocol π' to run for more than $O(I)$ rounds, and thus they cannot make it reveal more than $O(I)$ bits of information. Unfortunately, the recent result of [GKR14] asserts that there are I-bit information protocols which cannot be simulated by less than than $2^{\Omega(I)}$ bits of communication, and therefore this approach does not work.

We adapt our odometer construction to get a generic (black-box) conversion from a low-information protocol in the honest-but-curious model to a low-information protocol for the *adversarial* model. The basic premise is simple: we would like to maintain an estimate on the amount of information revealed so far, and abort if this number exceeds, say, $10I$. This plan is complicated by the fact that the dishonest party (say Bob) may try to attack this process in various ways. Firstly, he can try to fool the odometer into thinking that he learns less information than he actually does. Secondly, and perhaps more importantly, Bob can try to use the odometer itself to learn additional information about X. In particular, if it is Bob's turn to select the variable Z discussed above, Bob may cheat and select Z adversarially to elicit information from Alice. We modify the odometer protocol so that such cheating *can only hasten the termination of the simulation* (and cause Bob learn less information). We note that in our simulation Alice does not try to enforce Bob's compliance; rather, we just guarantee that the odometer has a proper estimate on what Bob learned so far, and thus it allows us to terminate once too much information has been revealed. Our conversion result postulates that Alice and Bob share the knowledge of a prior distribution μ of their inputs (information-theoretic quantities are meaningless without an underlying prior). We believe that these results can be generalized to the *prior-free* setting using techniques similar to the ones used to define prior-free information complexity in [Bra12].

A better understanding of the hardness of interactive protocol compression Finally, the odometer construction sheds some new light on the problem of compressing interactive communication. The interactive compression problem [BBCR10, BR11, Bra12] asks whether any protocol π whose information cost is I (i.e. which reveals an average of I bits of information to the players) can be simulated by a protocol π' that actually only uses $\sim I$ bits of communication. The question of interactive compression is known to be equivalent to the direct sum problem for randomized communication complexity [BR11], and better compression schemes correspond to stronger direct

[3] In the secure computation literature information loss is typically measured as the difference between what the parties learn (the information cost) and what they were supposed to learn (the mutual information between the output and the other partie's input). To keep notation simple, we ignore the latter term here, since it does not substantially affect any of the result. To be specific, one may assume that Alice and Bob are trying to compute only a few bits of output, and thus this term is negligible.

sum theorems. Classical results in information theory show that a similar statement holds when π is non-interactive (i.e. only consists of a single message).

In [Bra12], it was shown that such protocol π can be compressed in to a protocol that uses $2^{O(I)}$ communication (but nevertheless independent of the original communication C!). A recent breakthrough result of Ganor, Kol and Raz [GKR14] asserts that this compression is tight, by exhibiting a function f whose information complexity is I, yet requires $\geq 2^{\Omega(I)}$ communication to solve. This negative result rules out only the strongest possible direct sum theorem (namely, that there is a function f for which $\mathsf{CC}(f^n) \lesssim \frac{n}{\log n}\mathsf{CC}(f)$), but does not rule out somewhat weaker, yet nontrivial direct sums, which can be obtained by weaker compression schemes that are allowed to depend (mildly) on the original communication C as well as on I.

Indeed, the state of-the-art interactive compression result of [BBCR10] shows how to compress a protocol π which uses C bits of communication and I bits of information into a protocol that uses $g(I, C) = \tilde{O}(\sqrt{I \cdot C})$ bits of communication, leading to a partial direct sum result ($\mathsf{CC}(f^n) \geq \tilde{\Omega}(\sqrt{n}) \cdot \mathsf{CC}(f) \ \forall\ f$). In this language, [GKR14]'s result only rules out compression to $g(I, C) \leq I \cdot o(\log C)$ communication, so it is still hopeful to reduce the dependence on C in the compression result of [BBCR10]. In particular, if one could compress π into $g(I, C) \approx I \cdot C^{o(1)}$ bits of communication, this would yield a near-optimal direct sum (and direct product) result in communication complexity.

Our odometer construction suggests a "meta-approach" for pursuing such improved compression, as it allows one to break any protocol π into smaller pieces and compress each piece separately: For example, we can pause the protocol after ~ 1 bit has been revealed, and then continue, thus breaking π into $\sim I$ pieces, each revealing $\Theta(1)$ information (a similar approach led to near-optimal compression for the *external information* measure). Unfortunately, in this case the additive $\log C$ overhead of our odometer scheme matters — it implies that each piece will reveal $O(\log C)$ information while using $\leq C$ communication. Still, if we were able to compress a protocol whose information cost is $O(\log C)$ and whose communication cost is $\leq C$ into a protocol which uses $g(C, \log C)$ communication, then we could compress π into a protocol which uses $O(I \cdot g(C, \log C))$ communication. Thus, our odometer reduces the interactive compression problem to the arguably simpler regime where $I = O(\log C)$. In particular, if we could compress into $g(C, \log C) = C^{o(1)}$, then we could compress any π into $I \cdot C^{o(1)}$ communication — in turn implying a near-linear direct sum theorem. Note that both compression schemes from [BBCR10] and [Bra12] yield an upper bound of $g(C, \log C) = C^{O(1)}$ in this case. We discussed this approach further in Section 8.

2 Our Results

We begin by showing how to construct a single-round information odometer. The following lemma serves as the main building block in subsequent applications and constructions in this paper.

Theorem 1 (One round information odometer)**.** *Let $(p, q) \sim \mathcal{D}$ be two numbers $\in (1/3, 2/3)$, such that $\forall\ q\ \mathbb{E}_{p|q}[p] = q$. Suppose that Alice is given p (not known to Bob), and Bob is given q (not known to Alice). Then there is a a (2-round) protocol τ such that:*

- *At the end of execution, the players output "1" with probability exactly $2(p-q)^2$.*
- *The expected information cost of τ is small: If $T = T(p, q)$ denotes the transcript of τ, then*

there is some global constant $c_\tau > 0$ such that

$$\mathbb{E}_{p,q}\left[\mathsf{D}\left(\frac{(T|pq)}{(T|q)}\right) + \mathsf{D}\left(\frac{(T|pq)}{(T|p)}\right)\right] \leq c_\tau \cdot \mathbb{E}_{p,q}\left[H((p-q)^2)\right].$$

We then use our odometer construction together with previous techniques from [BRWY12] to prove a strong direct product theorem for communication in terms of information complexity: Let $\mathsf{suc}(\mu, f, C)$ denote the maximum success probability of a protocol with communication complexity (at most) C in computing a function $f(x, y)$ when the inputs are drawn from the distribution μ. Similarly, let $\mathsf{suc}^{\mathsf{i}}(\mu, f, I)$ denote the maximum success probability of a protocol with *information* complexity (at most) I in computing f under μ. Let $f^n(x_1, \ldots, x_n, y_1, \ldots, y_n)$ denote the function that maps its inputs to the n bits $(f(x_1, y_1), f(x_2, y_2), \ldots, f(x_n, y_n))$ and μ^n denote the product distribution on n pairs of inputs, where each pair is sampled independently according to μ. We prove the following result:

Theorem 2. *Let f be a 2-party Boolean function. There are universal constants $\alpha, \beta > 0$ such that if $\gamma = \beta(1 - \mathsf{suc}^{\mathsf{i}}(\mu, f, I))/2$, $T \geq 2$, and $T \log(T \log(T/\gamma)) < \alpha\gamma^2 n \cdot I$, it holds that $\mathsf{suc}(\mu^n, f^n, T) \leq \exp\left(-\gamma^2 n\right)$.*

Towards the above result, we prove the following technical theorem, showing how our odometer can be used to convert a protocol which is statistically close to having low (internal) information cost, to a protocol which actually has low information cost.

Theorem 3 (Conditional abort theorem)**.** *Let θ be an alternating, smooth[4] protocol with inputs x, y, public randomness r, and messages m, and suppose q is another distribution on these variables such that $\theta(xyrm) \overset{\epsilon}{\approx} q(xyrm)$. Denote $I_q := I_q(X; M|YR) + I_q(Y; M|XR)$. Then, there exists a protocol π that 15ϵ-simulates θ with $\|\pi\| \leq O(\|\theta\| \log(\|\theta\|))$ and*

$$\mathsf{IC}(\pi) \leq O\left(\frac{I_q + \log(\|\theta\| + 1)}{\epsilon^2}\right).$$

We then turn to the setting of communication between two untrusted parties. We show how our information odometer can be used, in a black-box fashion, to achieve (the best possible) information-theoretic privacy against an *arbitrary* adversary. More specifically, we prove:

Theorem 4 (Private simulation, informally stated)**.** *Let θ be a two-party communication protocol such that $\mathsf{IC}(\theta) = I$. Then for any $\delta > 0$, there is a communication protocol $\tilde{\pi}$ using "live" randomness, with the following properties:*

- *If both parties are honest, then $\tilde{\pi}$ 2δ-simulates θ.*
- $\mathsf{IC}(\tilde{\pi}) \leq O(I + \log(\|\theta\|))$.
- *There is a global constant $\lambda > 0$ such that for any protocol $\tilde{\pi}'$ where at least one party is honest (follows $\tilde{\pi}$), the following holds: $\forall\ k \in \mathbb{N}$,*

$$\Pr[\textit{ Honest party reveals more than } \lambda k(I/\delta + \log(\|\theta\| + 1)) \textit{ bits of information}] \leq 2^{-\Omega(k)}.$$

[4] See definition 22.

That is, an honest player never reveals to the other player much more than the essential amount of information required to solve f. The protocol does not assume any prior knowledge about the honesty of any player.

Finally, in Section 8, we discuss the implications of our odometer construction to the interactive compression problem, in light of the recent (exponential) separation result of [GKR14]. We outline a potential strategy for improving state of the art compression results, which uses the odometer to "break" the underlying protocol into low-information pieces ($\sim \log C$), and compress each one separately. In general, we obtain the following claim:

Claim 5. *Suppose there is a compression protocol that takes as an input a protocol π_1 with communication cost C_1 and* worst case [5] *information cost I_1, and compresses it into a protocol π_1' of communication complexity $g(C_1, I_1)$. Then any protocol π with communication cost C and information cost I can be compressed into a protocol with communication cost $\tilde{O}(I \cdot g(C, \log C))$.*

Organization The paper is organized as follows. We first show how to construct the (single-round) information odometer (in Section 4 below). Subsequent sections are applications of this construction. In Section 5 we prove Theorem 3. Next, in Section 6, we apply this result to prove the strong direct product theorem (Theorem 2). In Section 7 we prove the secure simulation result (Theorem 31). Finally, in Section 8 we outline the connection between the odometer and the one-shot interactive-compression problem.

3 Preliminaries

3.1 Notation

Unless otherwise stated, logarithms in this text are computed in base two. Random variables are denoted by capital letters and values they attain are denoted by lower-case letters. For example, A may be a random variable and then a denotes a value A may attain and we may consider the event $A = a$. Given $a = a_1, a_2, \ldots, a_n$, we write $a_{\leq i}$ to denote $a_1, \ldots, a_i$. We define $a_{>i}$ and $a_{\leq i}$ similarly.

We use the notation $p(a)$ to denote both the distribution on the variable a, and the number $\Pr_p[A = a]$. The meaning will usually be clear from context, but in cases where there may be confusion we shall be more explicit about which meaning is being used. We write $p(a|b)$ to denote either the distribution of A conditioned on the event $B = b$, or the number $\Pr[A = a|B = b]$. Again, the meaning will usually be clear from context. Given a distribution $p(a, b, c, d)$, we write $p(a, b, c)$ to denote the marginal distribution on the variables a, b, c (or the corresponding probability). We often write $p(ab)$ instead of $p(a, b)$ for conciseness of notation. If W is an event, we write $p(W)$ to denote its probability according to p. We denote by $\mathbb{E}_{p(a)}[g(a)]$ the expected value of $g(a)$ with respect to a distributed according to p.

For two distributions p, q, we write $|p(a) - q(a)|$ to denote the ℓ_1 distance between the distributions p and q. We write $p \overset{\epsilon}{\approx} q$ if $|p - q| \leq \epsilon$.
The *divergence* between two distributions p, q is defined to be

$$\mathsf{D}\left(\frac{p(a)}{q(a)}\right) = \sum_a p(a) \log \frac{p(a)}{q(a)}.$$

[5] This notion is formally defined in Section 8 of [BW14].

By slight abuse of notation, when p and q are numbers $\in [0,1]$,we define the binary divergence as

$$\mathsf{D}\left(\frac{p}{q}\right) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

Similarly, the *binary entropy* function of a number $p \in [0,1]$ is defined as

$$H(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}.$$

For three random variables A, B, C with underlying probability distribution $p(a,b,c)$, the *mutual information* between A, B conditioned on C is defined as

$$I(A;B|C) = \mathop{\mathbb{E}}_{p(cb)}\left[\mathsf{D}\left(\frac{p(a|bc)}{p(a|c)}\right)\right] = \mathop{\mathbb{E}}_{p(ca)}\left[\mathsf{D}\left(\frac{p(b|ac)}{p(b|c)}\right)\right] = \sum_{a,b,c} p(abc) \log \frac{p(a|bc)}{p(a|c)}.$$

Proposition 6. *Let A, B, C, D be random variables such that $I(B;D|AC) = 0$. Then $I(A;B|C) \geq I(A;B|CD)$.*

Proof. We apply the chain rule twice:

$$\begin{aligned} I(A;B|CD) &= I(AD;B|C) - I(D;B|C) = I(A;B|C) + I(D;B|AC) - I(D;B|C) = \\ &= I(A;B|C) - I(D;B|C) \leq I(A;B|C). \end{aligned}$$

□

Proposition 7. *Let A, B, C, D be four random variables in the same probability space. If $I(A;D|C) = 0$, then it holds that $I(A;B|C) \leq I(A;B|CD)$.*

Proof. Again we apply the chain rule in two different orders. On one hand, we have

$$I(A;BD|C) = I(A;B|C) + I(A;D|CB) \geq I(A;B|C)$$

since mutual information is nonnegative. On the other hand,

$$I(A;BD|C) = I(A;D|C) + I(A;B|CD) = I(A;B|CD)$$

since $I(A;D|C) = 0$ by the independence assumption on A and D. Combining both equations completes the proof. □

3.2 Communication and Information Complexity

Given a protocol π that operates over inputs $x, y \sim \mu$, and uses public randomness[6] r and messages m, we write $\pi(xymr)$ to denote the joint distribution of these variables. We write $\|\pi\|$ to denote

[6] In our paper we define protocols where the public randomness is sampled from a continuous (i.e. non-discrete) set. Nevertheless, we often treat the randomness as if it were supported on a discrete set, for example by taking the sum over the set rather than the integral. This simplifies notation throughout our proofs, and does not affect correctness in any way, since all of our public randomness can be approximated to arbitrary accuracy by sufficiently dense finite sets..

the *communication complexity* of π, namely the maximum number of bits that may be exchanged by the protocol.

A central notion in our work is the information complexity of a protocol (see [BBCR10, Bra12] and references within for a more detailed overview). The *(internal) information cost* of π is defined to be $\mathsf{IC}(\pi) = I_\pi(X; M|YR) + I_\pi(Y; M|XR)$.

Proposition 8 ([BR11])**.** $\forall\ \pi\ \ \mathsf{IC}(\pi) \leq \|\pi\|$.

Let $q(x, y, a)$ be an arbitrary distribution. We say that a protocol π *δ-simulates* q, if there is a function g and a function h such that

$$\pi(x, y, g(x, r, m), h(y, r, m)) \overset{\delta}{\approx} q(x, y, a, a), \tag{1}$$

where $q(x, y, a, a)$ is the distribution on 4-tuples (x, y, a, a) where (x, y, a) are distributed according to q. Thus if π δ-simulates q, the protocol allows the parties to sample a according to $q(a|xy)$.

If λ is a protocol with inputs x, y, public randomness r' and messages m', we say that π δ-simulates λ if π δ-simulates $\lambda(x, y, (r', m'))$. We say that π computes f with success probability $1 - \delta$, if π δ-simulates $\pi(x, y, f(x, y))$.

Remark 9. *A central lemma from [BRWY12], which is used in the proof of Theorem 3 in this paper, requires a stronger notion of simulation, namely, that outcome of the simulation is apparent even to an* external *observer who does not know x or y. More precisely, we say that π* strongly *δ-simulates q if in* (1)*, the function $g(x, r, m)$ does not depend on x. However, for information purposes, we note that these two notions are equivalent, as as one party can always write the final output of the protocol at the end of execution, thereby making the simulation strong. This message, call it M, will reveal no extra internal information to the receiving party, as π is assumed to (weakly) simulate q, and so $I(M; X|YR) \leq H(M|YR) = 0$. Therefore, in this paper we use the standard notion of simulation to mean strong simulation.*

3.3 Useful inequalities

Proofs for the following simple facts can be found in [CT91].

Fact 10 (Divergence is Non-negative)**.** $\mathsf{D}\left(\frac{p(a)}{q(a)}\right) \geq 0$.

Fact 11 (Chain Rule)**.** *If $a = a_1, \ldots, a_s$, then*

$$\mathsf{D}\left(\frac{p(a)}{q(a)}\right) = \sum_{i=1}^{s} \mathop{\mathbb{E}}_{p(a_{<i})} \left[\mathsf{D}\left(\frac{p(a_i|a_{<i})}{q(a_i|a_{<i})}\right)\right].$$

Fact 12 (Projection minimizes divergence)**.** *Let $T, X, Y \sim p(txy)$ be (correlated) random variables in the same probability space. Then for any random variable $Z = Z(y) \sim q$, it holds that*

$$\forall y \quad \mathbb{E}_{x|y}\left[\mathsf{D}\left(\frac{T|xy}{T|y}\right)\right] \leq \mathbb{E}_{x|y}\left[\mathsf{D}\left(\frac{T|xy}{Z}\right)\right].$$

Proof. Fix any y and denote $T' := T|y$, $T'|x := T|xy$ and $p'(tx) := p(tx|y)$. Then

$$\mathbb{E}_{x|y}\left[\mathsf{D}\left(\frac{T|xy}{T|y}\right)\right] - \mathbb{E}_{x|y}\left[\mathsf{D}\left(\frac{T|xy}{Z}\right)\right] = \mathbb{E}\left[\mathsf{D}\left(\frac{T'|x}{T'}\right)\right] - \mathbb{E}\left[\mathsf{D}\left(\frac{T'|x}{Z}\right)\right]$$
$$= \sum_{xt} p'(xt)\left[\log\frac{p'(tx)}{p'(t)} - \log\frac{p'(tx)}{q(t)}\right] = \sum_{xt} p'(xt)\log\frac{q(t)}{p'(t)} = -\mathsf{D}\left(\frac{p'(t)}{q(t)}\right) \le 0$$

where the last transition is by Fact 10. Rearranging completes the proof. □

Proposition 13 (Properties of binary entropy)**.** *For any $x \in [0,1]$, the binary entropy function $H(x)$ satisfies the following properties:*

(i) $H(x) \le 2\sqrt{x(1-x)}$.

(ii) For any $y \in [0,1]$, $y \cdot H(x) \le H(yx)$.

(iii) For any $y \ge 1$, $y \cdot H(x) \ge H(yx)$.

(iv) If $|x-y| \le \epsilon$, $|H(x) - H(y)| \le H(\epsilon)$.

All the above facts essentially follow from concavity of entropy ($H(x/2+y/2) \ge H(x)/2 + H(y)/2$). For detailed proofs see [CT91].

Proposition 14 (ℓ_2^2 approximates divergence)**.** *For any $p, q \in [1/3, 2/3]$, it holds that*

$$2(p-q)^2 \le \mathsf{D}\left(\frac{p}{q}\right) \le \frac{9}{2}\cdot(p-q)^2.$$

Proof. The left hand side is Pinsker's inequality. To prove the right hand side, we have:

$$\mathsf{D}\left(\frac{p}{q}\right) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} = p\log\frac{q-(q-p)}{q} + (1-p)\log\frac{1-q+(q-p)}{1-q}$$
$$= p\log\left(1+\frac{p-q}{q}\right) + (1-p)\log\left(1+\frac{q-p}{1-q}\right) \le p\cdot\frac{p-q}{q} + (1-p)\cdot\frac{q-p}{1-q}$$
(since $\log(1+x) \le x$)
$$= (p-q)\left(\frac{p}{q} - \frac{1-p}{1-q}\right) = (p-q)\left(\frac{p-pq-q+pq}{q(1-q)}\right) = \frac{(p-q)^2}{q(1-q)} \le \frac{9}{2}\cdot(p-q)^2$$

where the last inequality follows from the assumption that $q \in [1/3, 2/3]$, which implies that $q(1-q) \ge 2/9$. □

Lemma 15 (Multiplicative Chernoff Bound)**.** *Let $X = \sum_{i=1}^n X_i$ be the sum of n independent random variables, where $\mathbb{E}[X_i] = p_i$. Denote $\eta := \sum_i p_i$. Then*

- *For all $1 \ge \delta \ge 0$, $\Pr[X < (1-\delta)\eta] \le e^{-\frac{\delta^2\eta}{2}}$.*
- *For all $\delta \ge 0$, $\Pr[X > (1+\delta)\eta] \le e^{-\frac{\delta^2\eta}{2+\delta}}$.*

The first proposition is the standard Chernoff bound (e.g, [AS92]). The second proposition follows from the same proof as in [AS92], by observing that $\ln(1+\delta) > 2\delta/(2+\delta)$ for all $\delta > 0$, and so $\delta - (1+\delta)\ln(1+\delta) \leq -\delta^2/(2+\delta)$.
We will need need the following variant of the Chernoff bound:

Corollary 16. *Let $X = \sum_{i=1}^{n} X_i$ be the sum of n independent random variables, where $\mathbb{E}[X_i] = p_i$. Denote $\eta := \sum_i p_i$. Then for every $\beta \geq \eta$ and all $\delta \geq 2$, it holds that*

$$\Pr[X > (1+\delta)\beta] \leq e^{-\frac{\delta\beta}{2}}.$$

Proof.

$$\begin{aligned}
&\Pr[X > (1+\delta)\beta] = \Pr[X > (1+\delta)(\beta/\eta)\eta] \qquad (\text{Define } \delta' := (1+\delta)(\beta/\eta) - 1)\\
&= \Pr[X > (1+\delta')\eta] \leq e^{-\frac{\delta'^2\eta}{2+\delta'}} \qquad (\text{By Lemma 15})\\
&\leq e^{-\frac{\delta'\eta}{2}} \qquad (\text{ Since } \beta \geq \eta \text{ and } \delta \geq 2 \Rightarrow 2\delta' > 2+\delta')\\
&= e^{-\frac{(1+\delta)\beta-\eta}{2}} = e^{-\frac{(\beta-\eta)+\delta\beta}{2}} \leq e^{-\frac{\delta\beta}{2}} \qquad (\text{Since } \beta \geq \eta\ )
\end{aligned}$$

□

Theorem 17 (Azuma's inequality)**.** *Let $\{X_k\}_{k=0}^{\infty}$ be a sub-martingale such that $|X_i - X_{i-1}| \leq c_i$ almost surely. Then for any $N \in \mathbb{N}$ and any $k \in \mathbb{R}^+$,*

$$\Pr[X_N - X_0 \leq -k] \leq \exp\left(-\frac{k^2}{2\sum_{i=0}^{N} c_i^2}\right).$$

4 A single round information odometer

In this section we prove Theorem 1, the main building block of the information odometer.

Proof of Theorem 1. The players run the protocol τ from Figure 1.

Analysis: Throughout the entire analysis, we assume that $p \leq 1/2$, as it is straightforward to verify that $\mu_p(z) = \mu_{1-p}(1-z)$. First, let us analyze the probability with which the players output "1". Note that the assumption that $p, q \in [1/3, 2/3]$ implies that either $q \in [0, p]$ or $q \in [p, p+1/2]$. If $q \in [p, p+1/2]$, then by construction we have

$$\begin{aligned}
\Pr[\text{players output "1"}] = \Pr[I^p \neq I^q] = \Pr_{\mu_p}[Z \in [p,q]] = \int_p^q \mu(z)dz = \int_p^q 4(z-p)dz =\\
= \left[2z^2 - 4pz)\right]_p^q = 2q^2 - 4pq - 2p^2 + 4p^2 = 2(p-q)^2. \qquad (2)
\end{aligned}$$

Similarly, if $q \in [0, p]$, then

$$\Pr[\text{players output "1"}] = \int_q^p \mu(z)dz = \int_p^q 4(p-z)dz = 2(p-q)^2,$$

The protocol τ

1. Given her number p, Alice samples a number $Z_p \in [0,1]$, according to the following probability density function:

$$\mu_p(z) = \begin{cases} 4(p-z) & \text{if } 0 \le z < p \\ 4(z-p) & \text{if } p \le z \le p+1/2 \\ 2-4(z-p-1/2) & \text{if } p+1/2 < z \le 1 \end{cases}$$

   If $p > 1/2$, Alice samples from $\mu_{1-p}(1-z)$.

2. Alice sends Z_p to Bob.
3. Alice sends Bob a bit I^p indicating whether "$Z_p > p$".
4. Bob responds by sending a bit I^q indicating whether "$Z_p > q$".
5. The players output "1" iff $I^p \neq I^q$.

Figure 1: A protocol for estimating internal information cost. The probability that the protocol outputs "1" is $2(p-q)^2$.

as claimed in the first proposition of the Theorem.

We turn to analyze the information cost of τ. By the chain rule,the information cost of the protocol can be written as

$$\mathsf{IC}(\tau) = I(Z_p; p|q) + I(I^p; p|q, Z_p) + I(I^q; q|p, Z_p, I^p).$$

We analyze step 2 of the protocol and steps 3,4 separately. **Step 2:** The heart of the proof is showing that the information Z_p conveys to Bob (with input q) about Alice's input p, is in fact comparable to the divergence between p and q:

Lemma 18. *There is some global constant $c > 0$ such that: $I(Z_p; p|q) \le c \cdot \mathbb{E}_{p,q}\left[H((p-q)^2)\right]$.*

The key step is the following technical lemma which asserts that the divergence between the distribution of Z_p and a "shift" of it $Z_{p'}$ is proportional to $H((p-p')^2)$ (see Figure 2):

Lemma 19. *There is some global constant $c < 200$, such that for any $p, p' \in (1/3, 2/3)$, it holds that $\mathsf{D}\left(\frac{Z_p}{Z_{p'}}\right) \le c \cdot H((p-p')^2)$.*

The lemma is proved by a direct calculation of the divergence, so we defer it to the appendix. We now show how Lemma 19 implies Lemma 18:

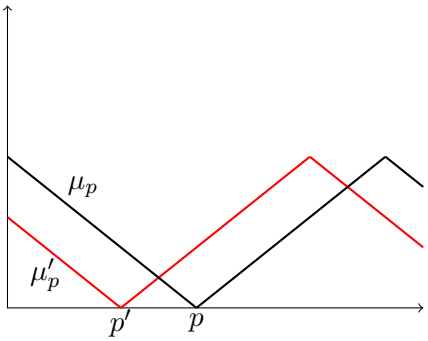

Figure 2: The distribution μ_p for $p = 0.5, p' = 0.3$. The divergence between μ_p and $\mu_{p'}$ is proportional to $H((p-p')^2)$. The structure of the density function μ_p ensures that the log-ratio between the distributions mostly cancels out, up to second order terms.

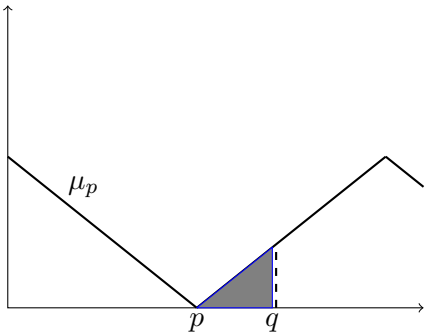

Figure 3: The distribution μ_p for $p = 0.5$. For any q, the probability that $p < Z_p < q$ is equal to the area of the triangle enclosing $p, q, \mu_p(q)$.

Proof of Lemma 18.

$$I(Z_p; p|q) = \mathbb{E}_{p,q}\left[\mathsf{D}\left(\frac{Z_p}{\mathbb{E}_{p|q}[Z_p]}\right)\right] \leq \mathbb{E}_{p,q}\left[\mathsf{D}\left(\frac{Z_p}{Z_q}\right)\right]$$
$$\leq c \cdot \mathbb{E}_{p,q}\left[H((p-q)^2)\right]$$

where the second transition is by Fact 12 (taken with $T = Z, X = p, Y = q, Z(q) = Z_q$), and the last transition is by Lemma 19. □

We continue to bound the information of the remaining steps of the protocol τ. **Steps 3 and 4:** For any fixed p, q, let W denote the indicator random variable of the event "$Z_p \in [p, q]$". Note that at this point, both players already know Z_p, and conditioned on I^p and Z_p, W determines I^q (and vice versa for I^p). The data processing inequality implies that the information cost of the above steps can be upper bounded as follows

$$\begin{aligned} &I(I^p; p|q, Z_p) + I(I^q; q|p, Z_p, I^p) = I(I^p; p|q, I^q, Z_p) + I(I^q; q|p, I^p, Z_p) \\ &\leq H(I^p|I^q Z_p) + H(I^q|I^p Z_p) \leq H(W|I^q Z_p) + H(W|I^p Z_p) \\ &\leq 2H(W) = 2\mathbb{E}_{p,q}\left[H(2(p-q)^2)\right], \end{aligned} \tag{3}$$

where the first transition follows since q and Z_p together determine I^q, the second transition follows since conditioning does not increase entropy, and the last transition is by (2). By Lemma 18 and

(3), we conclude that

$$\begin{aligned}
&\mathbb{E}_{p,q}\left[\mathsf{D}\left(\frac{(T|pq)}{(T|q)}\right)+\mathsf{D}\left(\frac{(T|pq)}{(T|p)}\right)\right]\\
&=I(Z_p;p|q)+I(I^p;p|q,Z_p)+I(I^q;q|p,Z_p,I^p)\\
&\leq I(Z_p;p|q)+\mathbb{E}_{pq}\left[H(I^p|I^qZ_p)+H(I^q|I^pZ_p)\right]\\
&\leq c\cdot\mathbb{E}_{p,q}\left[H((p-q)^2)\right]+2\mathbb{E}_{p,q}\left[H(2(p-q)^2)\right]\\
&\leq c\cdot\mathbb{E}_{p,q}\left[H((p-q)^2)\right]+4\mathbb{E}_{p,q}\left[H((p-q)^2)\right]\\
&\text{(By the third proposition of Fact 13)}\\
&\leq (c+4)\cdot\mathbb{E}_{p,q}\left[H((p-q)^2)\right]
\end{aligned}$$

which concludes the second proposition and thus the whole proof of Theorem 1.

□

Corollary 20. *Consider the same setup of Theorem 1, and suppose the players execute the protocol τ, except that Alice sends only the first $\ell-2$ bits of (the binary representation of) Z_p. Denote this protocol τ_ℓ (and its transcript by T). Then there is a global constant c_τ such that:*

- *At the end of execution, the players output "1" with probability $2(p-q)^2\pm O(2^{-\ell})$.*
- $\mathsf{IC}(\tau_\ell)=\mathbb{E}_{p,q}\left[\mathsf{D}\left(\frac{(T|pq)}{(T|q)}\right)+\mathsf{D}\left(\frac{(T|pq)}{(T|p)}\right)\right]\leq c_\tau\cdot\mathbb{E}_{p,q}\left[H((p-q)^2)\right].$
- $\|\tau_\ell\|=\ell$.

Proof Sketch. First proposition: Denote $\ell':=\ell-2$. Since $Z_p^{\ell'}-Z_p\leq 2^{-\ell'}$, $\Pr(I^p\neq I^p_{\ell'})\leq\Pr(Z_p^{\ell'}<p\wedge Z_p>p)\leq\Pr(Z_p\in[p,p+2^{-\ell'}])=\int_p^{p+2^{-\ell'}}\mu_p(z)dz=\int_p^{p+2^{-\ell'}}4(z-p)dz=4p2^{-\ell'}+4p2^{-2\ell'}-4p2^{-\ell'}=4p2^{-2\ell'}<2^{-(\ell+1)}$. the same goes for $\Pr(I^q\neq I^q_{\ell'}$, and the statement directly follows. The second proposition follows immediately from the data processing inequality, since $Z^{\ell'}$ (the first ℓ' bits of Z_p) is a deterministic function of Z_p, and so $I_{p\sim\mathcal{D}|q}(Z^{\ell'};p)\leq I_{p\sim\mathcal{D}|q}(Z;p)=\mathsf{D}\left(\frac{Z_p}{Z_q}\right)$, and the rest of the analysis follows from Theorem 1. □

5 Proof of the conditional abort theorem

In this section we prove Theorem 3 − we show how the information odometer from Section 4 can be used to modify any protocol statistically close to having low information, to a protocol that actually has low information. We will first need to set up some definitions. The following definition will be central to our analysis.

Definition 21. *For fixed inputs x,y and public randomness r, and a fixed (partial) path $m=m_1m_2\ldots m_j$ in θ, define*

$$\mathbb{D}^\theta_x(m_{\leq j})=\sum_{t=1}^{j}\left[\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}yr)}\right)\right]\quad,\quad\mathbb{D}^\theta_y(m_{\leq j})=\sum_{t=1}^{j}\left[\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}xr)}\right)\right].$$

The divergence cost of a path $m_{\leq j}$ under x, y, r is

$$\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) := \mathbb{D}^{\theta}_{x}(m_{\leq j}) + \mathbb{D}^{\theta}_{y}(m_{\leq j})$$

When we refer to the divergence cost as a random variable we use the notation $\mathbb{D}^{\theta}_{XYR}(M_{\leq j})$. A straightforward application of the chain rule for mutual information shows that

$$\mathbb{E}\left[\mathbb{D}^{\theta}_{xyr}(m)\right] = I_{\theta}(M; X|YR) + I_{\theta}(M; Y|XR) = \mathsf{IC}(\theta). \tag{4}$$

For completeness, we provide a formal proof of this fact in the appendix.

5.1 Smooth simulation

To prove Theorem 3, we will also need to assume our protocols are such that each bit m_t sent is not too biased, in every possible transcript[7]. This is formalized by the following definition:

Definition 22 (Smooth protocols [BBCR10])**.** *A protocol π is δ-smooth if $\forall\, x, y, r, t, m_{<t}$:*

$$(i)\ \pi(M_t = 1|xrm_{<t}) \in \{1/2 - \delta, 1/2 + \delta\} \quad \textit{and} \quad (ii)\ \pi(M_t = 1|yrm_{<t}) \in \{1/2 - \delta, 1/2 + \delta\}.$$

We say that a protocol is smooth if it is $\frac{1}{3}$-smooth.

The following lemma asserts that any protocol can be simulated by a smooth protocol, with a small overhead in the communication. The proof is adapted from [BBCR10], with slight modifications. For completeness, we present a short proof below.

Lemma 23 (Smooth Simulation)**.** *There exists a constant $s > 0$ such that for every protocol π and distribution μ on inputs x, y and all $0 < \epsilon < 1$ there exists a smooth protocol τ that $\epsilon-$simulates π, $\|\tau\| \leq s\|\pi\| \log(\|\pi\|/\epsilon)$, and $\mathsf{IC}(\tau) \leq \mathsf{IC}(\pi) + 2$.*

Proof. Set $\delta = 1/3$. Every time Alice wants to send a bit $M = M(X)$ in π, she instead sends $k = s \log(\|\pi\|/\epsilon)/\delta^2$ bits $W_1, \ldots, W_k$ which are each independently and privately chosen to be the correct value with probability $1/2 + \delta$. For odd messages sent by Alice, this ensures that condition (i) in definition 22 is satisfied for every $x, r, m_{<t}$ in τ. But since $\tau(M_t = 1|yrm_{<t}) = \mathbb{E}_{x|y}[\tau(M_t = 1|xrm_{<t})]$ and $\tau(M_t = 1|xrm_{<t}) \in \{1/2 - \delta, 1/2 + \delta\}$ *for every* x, the same guarantee applies to $\tau(M_t = 1|yrm_{<t})$. An analogues argument holds for even rounds when Bob is the sender of the message. After this, the receiving player takes the majority of the W_j's to reconstruct the intended transmission. The players then proceed assuming that the majority of the bits was the real sampled transmission. By the Chernoff bound, we can set s to be large enough so that the probability that *any* transmission is received incorrectly is, say, at most $(\epsilon/\|\pi\|^2)$. By the union bound applied to each of the $\|\pi\|$ transmissions, we have that except with probability $(\epsilon/\|\pi\|)$, all transmissions are correctly received. In particular, the distribution of the above transcript ϵ-simulates the correct distribution.

We proceed to bound the information cost of the alternate protocol. Let us denote by M_t the t'th message of the original protocol π, and by $M'_t := W^t_1, \ldots, W^t_k$ the t'th message of the smooth

[7]The reason we want this property is that the divergence between two such distributions can be well approximated by their ℓ_2 distance, which is a more convenient measure to work with. See Proposition 14.

protocol τ. Let $\mathcal{G}$ the event that all transmissions of the intended messages were received correctly. We have

$$
\begin{aligned}
&I(\tau; X|Y, \mathcal{G}) = \sum_t I(M'_t; X|Y M'_{<t}, \mathcal{G}) \\
&\leq \sum_t I(M'_t; X|Y M'_{<t} M_{<t}, \mathcal{G}) \quad \text{(Since conditioned on } \mathcal{G},\ M'_{<t} \text{ determines } M_{<t}) \\
&= \sum_t I(M'_t; X|Y M_{<t}, \mathcal{G}) \quad \text{(Since by definition, } M'_t \text{ depends only on the decoded messages } M_{<t}) \\
&= \sum_t I(W_1^t, \ldots, W_k^t; X|Y M_{<t}, \mathcal{G})
\end{aligned}
$$

Consider each term above. We have

$$
\begin{aligned}
&I(W_1^t, \ldots, W_k^t; X|Y M_{<t}, \mathcal{G}) \leq I(M_t, W_1^t, \ldots, W_k^t; X|Y M_{<t}, \mathcal{G}) \\
&= I(M_t; X|Y M_{<t}, \mathcal{G}) + I(W_1^t, \ldots, W_k^t; X|Y M_{\leq t}, \mathcal{G}) \\
&\leq I(M_t; X|Y M_{<t}, \mathcal{G}) + I(W_1^t, \ldots, W_k^t; XY|M_{\leq t}, \mathcal{G}) \\
&= I(M_t; X|Y M_{<t}, \mathcal{G}) + I(W_1^t, \ldots, W_k^t; X|M_{\leq t}, \mathcal{G}) + I(W_1^t, \ldots, W_k^t; Y|M_{\leq t} X, \mathcal{G}) \\
&= I(M_t; X|Y M_{<t}, \mathcal{G}),
\end{aligned}
$$

where in the last transition we used the fact that, conditioned on $M_{<t}$ and $\mathcal{G}$, M_t determines the distribution of the W_j^t's in the t'th round (hence $I(W_1, \ldots, W_k; X|M_{\leq t}, \mathcal{G}) = 0$) and the fact $I(W_1^t, \ldots, W_k^t; Y|M_{\leq t} X, \mathcal{G}) = 0$ by definition of a protocol (hence $I(W_1, \ldots, W_k; Y|M_{\leq t} X, \mathcal{G}) = 0$). An analogues argument for the messages sent by Bob implies that

$$I(\tau; X|Y, \mathcal{G}) + I(\tau; Y|X, \mathcal{G}) \leq \mathsf{IC}(\pi).$$

Finally, since $\Pr[\overline{\mathcal{G}}] \leq \epsilon/\|\pi\|$, and $I(\tau; X|Y, \overline{\mathcal{G}}) + I(\tau; Y|X, \overline{\mathcal{G}}) \leq \|\pi\|$, we conclude that

$$
\begin{aligned}
&\mathsf{IC}(\tau) \leq H(\epsilon/\|\pi\|) + I(\tau; X|Y, \mathcal{G}) + I(\tau; Y|X, \mathcal{G}) + (\epsilon/\|\pi\|) \cdot \left(I(\tau; X|Y, \overline{\mathcal{G}}) + I(\tau; Y|X, \overline{\mathcal{G}})\right) \\
&\leq H(\epsilon/\|\pi\|) + \mathsf{IC}(\pi) + \epsilon \leq \mathsf{IC}(\pi) + 2.
\end{aligned}
$$

□

Finally, to simplify our analysis we will require that at each round a player sends only one bit:

Definition 24 (Alternating protocols)**.** *A protocol is alternating if each party sends exactly one bit of communication at each round.*

Proposition 25. *Any smooth protocol π can be (exactly) simulated by an alternating smooth protocol τ such that $|\tau| \leq 2\|\pi\|$, and $\mathsf{IC}(\tau) = \mathsf{IC}(\pi)$.*

Proof sketch. If in some round t in π, a party wishes to send more than one consecutive bit, we "split" her message in τ by having the other party send a uniformly random bit in between the original messages. Clearly, communication grows by a factor of at most 2, and the information cost remains the same as in π, as a uniform bit conveys no information at all. A uniform bit is clearly smooth, so smoothness is preserved.

□

5.2 Proof of Theorem 3

The high-level idea is for the players to run the protocol θ, while keeping an estimate of the internal information cost leaked so far, in an online fashion, using the information odometer protocol τ_ℓ from Section 4. If their estimate indicates that the information "leaked" is too high, they abort the protocol. Since θ is ϵ-close to a low-information distribution q, we can show that most paths in θ cannot reveal $\gg I_q/\epsilon^2$ information, and thus setting the abort threshold appropriately ensures that π aborts with small probability. The crucial point is that the additional information incurred by running the odometer for any partial path of θ is comparable to the information which θ itself revealed so far, and thus if we only run the odometer with probability $\propto 1/I_q$, we can keep the information overhead of π below I_q. The simulating protocol π is described in Figure 4.

Protocol π for simulating θ with conditional abort

$\ell \longleftarrow 2\log(\|\theta\|+1)$.
$\rho \longleftarrow \frac{2I_q+4/(e\ln 2)+3\log(\|\theta\|+1)}{\epsilon^2} + \frac{2\log(1/\epsilon)}{\epsilon}$.
$\alpha \longleftarrow \frac{\ln(1/\epsilon)}{\rho}$.
Count $\longleftarrow 0$.

For each round $t \in [\|\theta\|]$, Do:

1. The speaker in round t privately samples his message m_t as prescribed in θ ($m_t \sim \theta(m_t|m_{<t}xr)$ for odd t, and $m_t \sim \theta(m_t|m_{<t}yr)$ for even t).
2. With probability α (using independent public randomness at each round), the players run the protocol τ_ℓ from Corollary 20, setting $p_t := \theta(M_t = 1|m_{<t}xr)$, $q_t := \theta(M_t = 1|m_{<t}yr)$ for odd t, and $p_t := \theta(M_t = 1|m_{<t}yr)$, $q_t := \theta(M_t = 1|m_{<t}xr)$ for even t (Note that the smoothness of π ensures $p_t, q_t \in (1/3, 2/3)$ so the premises are satisfied).
3. If τ_ℓ outputs "1", both players set[a] Count $\longleftarrow$ Count $+1$.
4. If Count $> \lceil 4\alpha\rho \rceil$, the players abort the protocol[b].
5. Otherwise, the current speaker sends m_t as prescribed by θ.

[a] Since both players know the value of Count at each point, this line is well defined.
[b] for convenience, we achieve the aborts by having the players send 0's for the rest of the protocol, until $\|\theta\|$ bits have been communicated.

Figure 4: A low-information simulation of θ.

We turn to formalize the intuition above. For fixed x, y, r and a (partial) path $m_{\le j}$ in θ, let $Count(x, y, r, m_{\le j})$ denote the value of the random variable $Count$ which the players maintain throughout the protocol π. The following claim asserts that $Count(x, y, r, m_{\le j})$ provides the players a very sharp estimate on the divergence cost of the sampled path $m_{\le j}$:

Claim 26. *For any $x, y, r, m_{\le j}$, it holds that:*

- $\alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - 5\|\theta\|/2^{\ell})/5 \leq \mathbb{E}[Count(x,y,r,m_{\leq j})] \leq \alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) + (\|\theta\|/2^{\ell}))$,
- $\forall \delta \geq 2$:
  $\Pr_{R_\tau}[Count(x,y,r,m_{\leq j}) > (1+\delta)\alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) + (\|\theta\|/2^{\ell}))] \leq e^{-\frac{\delta\alpha\mathbb{D}^{\theta}_{xyr}(m_{\leq j})}{2}}$,
- $\forall 1 > \delta > 0$:
  $\Pr_{R_\tau}[Count(x,y,r,m_{\leq j}) < \frac{(1-\delta)\alpha}{5}(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - (5\|\theta\|/2^{\ell}))] \leq e^{-\frac{\alpha\delta^2}{10}(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - (5\|\theta\|/2^{\ell}))}$,

where R_τ denotes the randomness of π used in step 2 of the protocol.

Proof. For $t \in [j]$, let $\mathbf{1}^t_{\tau_\ell}$ be the indicator variable denoting the output of protocol τ_ℓ at step 2 of round t of π. Note that $\mathbb{E}[Count(x,y,r,m_{\leq j})] = \sum_{t=1}^{j} \alpha\mathbb{E}[\mathbf{1}^t_{\tau_\ell}]$. Furthermore, by the first proposition of Corollary 20, $\mathbb{E}[\mathbf{1}^t_{\tau_\ell}] \in 2(p_t - q_t)^2 \pm O(2^{-\ell})$. therefore we have

$$\mathbb{E}_{R_\tau}[Count(x,y,r,m_{\leq j})] = \sum_{t=1}^{j} \alpha\mathbb{E}[\mathbf{1}^t_{\tau_\ell}] \leq \alpha\left(\sum_{t=1}^{j}(2(p_t-q_t)^2 + O(2^{-\ell}))\right)$$

$$\leq \alpha\left(\sum_{t=1}^{j} \mathsf{D}\left(\frac{p_t}{q_t}\right) + O(j/2^{\ell})\right) \quad \text{(by Proposition 14)}$$

$$= \alpha\left(\sum_{t\ odd} \mathsf{D}\left(\frac{\theta(M_t=1|m_{<t}xyr)}{\theta(M_t=1|m_{<t}yr)}\right) + \sum_{t\ even} \mathsf{D}\left(\frac{\theta(M_t=1|m_{<t}xyr)}{\theta(M_t=1|m_{<t}xr)}\right) + O(j/2^{\ell})\right) \tag{5}$$

$$= \alpha\left(\sum_{t\ odd} \mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}yr)}\right) + \sum_{t\ even} \mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}xr)}\right) + O(j/2^{\ell})\right) \tag{6}$$

$$= \alpha\left(\sum_{t=1}^{j}\left[\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}xr)}\right) + \mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}yr)}\right)\right] + O(j/2^{\ell})\right)$$

$$= \alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) + O(j/2^{\ell})), \tag{7}$$

where in (5) we used the fact that for odd t, $\theta(M_t = 1|m_{<t}xr) = \theta(M_t = 1|m_{<t}xyr)$ by definition of a protocol (analogously for even t), and in (6) we used the fact that $\mathsf{D}\left(\frac{p}{q}\right) = \mathsf{D}\left(\frac{1-p}{1-q}\right)$. Applying the same argument with the "$\geq$" direction of Proposition 14, we get

$$\mathbb{E}_{R_\tau}[Count(x,y,r,m)] \geq \alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - (5j/2^{\ell}))/5. \tag{8}$$

Thus, given $x, y, r, m_{\leq j}$, $Count(x,y,r,m_{\leq j})$ is the sum of j ($\leq \|\theta\|$) independent random variables with expectation η, such that

$$\alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - (5\|\theta\|/2^{\ell}))/5 \leq \eta \leq \alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) + (\|\theta\|/2^{\ell})), \tag{9}$$

as claimed in the first proposition of the lemma.

Since $\eta \leq \alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) + (\|\theta\|/2^{\ell}))$, Corollary 16 implies that for all $\delta > 2$

$$\Pr_{R_\tau}[Count(x,y,r,m_{\leq j}) > (1+\delta)\alpha(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) + O(j/2^{\ell}))] \leq e^{-\frac{\delta\alpha\mathbb{D}^{\theta}_{xyr}(m_{\leq j})}{2}}$$

which concludes the second proposition of the lemma. For the other direction, the standard Chernoff bound (Lemma 15) implies that for $0 < \delta < 1$

$$\Pr_{R_\tau}[Count(x,y,r,m_{\leq j}) < \frac{(1-\delta)\alpha}{5}(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - 5j/2^{\ell})] \leq \Pr_{R_\tau}[Count(x,y,r,m_{\leq j}) < (1-\delta)\eta] \leq$$
$$\leq e^{-\frac{\delta^2\eta}{2}} \leq e^{-\frac{\alpha\delta^2}{10}(\mathbb{D}^{\theta}_{xyr}(m_{\leq j}) - (5\|\theta\|/2^{\ell}))} \quad , \quad \text{concluding the third proposition of the lemma.}$$

□

Claim 26 above will allow the players to control the information cost of π and to abort the execution only when the information of a path vastly exceeds the typical divergence cost (as they posses a sharp estimate on the information revealed at any step of the protocol). Most of the remaining proof of Theorem 3 is devoted to showing that the information cost of π is comparable to the threshold ρ at which the protocol aborts. This analysis is facilitated by viewing the information that π reveals as a stochastic process, with the "aborting" index C serving as a stopping rule for that process. We then invoke the optional stopping theorem ([Doo75]) to bound the expected information cost at the final step of the protocol. We defer the rest of this (technical) proof to the appendix (Section A).

6 Direct product in terms of information complexity

Let π be a (deterministic) protocol for computing f^n over inputs $\overline{x} = x_1, \ldots, x_n$ and $\overline{y} = y_1, \ldots, y_n$ drawn from μ^n, and let $\mathsf{suc}^{\mathsf{i}}(\mu, f, I)$ be the largest success probability of a protocol whose information cost is at most I, in computing a single copy of f under μ. To prove Theorem 2, we follow the approach of [BRWY12]. Let W be the event that π correctly computes f^n. For $i \in [n]$, let W_i denote the event that the protocol π correctly computes the i'th copy $f(x_i, y_i)$. Let $\pi(W)$ denote the probability of W, and $\pi(W_i|W)$ denote the conditional probability of the event W_i given W (clearly, $\pi(W_i|W) = 1$). We shall prove that if $\pi(W)$ is not very small and $\|\pi\| \ll I \cdot n$, then $(1/n)\sum_{i=1}^{n} \pi(W_i|W) < 1$, which is a contradiction. In fact, the proof holds for an arbitrary event W, as long as it occurs with large enough probability:

Lemma 27 (Main Lemma)**.** *Let f be a 2-party Boolean function. There are universal constants $\alpha, \beta > 0$ so that the following holds. For every $\gamma > 0$, and event W such that $\pi(W) \geq 2^{-\gamma^2 n}$, if $\|\pi\| \geq 2$, and $\|\pi\| \log(\|\pi\| \log(\|\pi\|/\gamma)) < \alpha n \gamma^2 \cdot I$, then $\frac{1}{n}\sum_{i\in[n]} \pi(W_i|W) \leq \mathsf{suc}^{\mathsf{i}}(\mu, f, I) + \gamma/\beta$.*

Let us first see how Lemma 27 easily implies Theorem 2.

Proof of Theorem 2. As outlined above, let W denote the event that π computes f correctly in all n coordinates. So, $(1/n)\sum_{i\in[n]} \pi(W_i|W) = 1$. Set $\gamma = \beta(1 - \mathsf{suc}^{\mathsf{i}}(\mu, f, I))/2$ so that $\mathsf{suc}^{\mathsf{i}}(\mu, f, I) + \gamma/\beta < 1$. Then by Lemma 27, either $\|\pi\| < 2$, $\|\pi\| \log(\|\pi\| \log(\|\pi\|/\gamma)) \geq \alpha n \gamma^2 I$, or $\pi(W) < 2^{-\gamma^2 n}$. □

It therefore remains to prove Lemma 27. The overall idea is to use the n-fold protocol π to produce a single-copy protocol with information cost $< I$ that computes f correctly with probability at least $(1/n)\sum_{i=1}^{n} \pi(W_i|W) - O(\gamma)$. This would imply that $(1/n)\sum_{i\in[n]} \pi(W_i|W) \leq \mathsf{suc}^{\mathsf{i}}(\mu, f, I) + O(\gamma)$, as desired. To this end, we wish to show that there exists a good simulating protocol for a random coordinate of $\pi|W$, whose average information cost is low (roughly $\|\pi\|/n$) and still

computes f on this coordinate with good probability. The existence of such protocol was proven in [BRWY12], except their protocol is not guaranteed to actually have low information cost, but to merely be statistically close to a low-information protocol:

Lemma 28 (Claims 26 and 27 from [BRWY12], restated)**.** *There is a protocol σ taking inputs $x, y \sim \mu$ so that the following holds:*

- *σ publicly chooses a uniform $i \in [n]$ independent of x, y, and R_i which is part of the input to π (intuitively, R_i determines the "missing" inputs x_{-i}, y_{-i} of π).*
- *$\mathbb{E}_{x,y,m,i,r_i} |\sigma(xyr_i m) - \pi(x_i y_i r_i m|W)| \leq 2\gamma$ (that is, σ is close to the distribution $(\pi|W)_i$ for average i).*
- *$\mathbb{E}_i \left[I_{\pi|W}(X_i; M|Y_i R_i i) + I_{\pi|W}(Y_i; M|X_i R_i i) \right] \leq 4\|\pi\|/n$.*

Note that the last proposition only guarantees that the information cost of the transcript under the distribution $(\pi|W)$ is low (on an average coordinate), while we need this property to hold for the simulating protocol σ. Unfortunately, $(\pi|W)$ is no longer a protocol[8] ! Nevertheless, since the second and third propositions of Lemma 28 ensure that σ is 2γ-close to a low-information distribution $q = \pi(x_i y_i r_i m|W)$, Theorem 3 can be applied so as to modify it to actually have low information (roughly $\|\pi\|/n$). For appropriately chosen parameters, this information cost will be $< I$, leading to our anticipated contradiction. A formal proof follows.

Proof of Lemma 27. Let $\beta = 1/48$. Let α be a sufficiently small constant to be determined shortly, and suppose that π is so that

$$\|\pi\| \log(\|\pi\| \log(\|\pi\|/\gamma)) < \alpha n \gamma^2 I. \tag{10}$$

As usual, let m denote the messages of π. Let σ be the protocol given by Lemma 28. By Lemma 23 and Proposition 25, σ can be made into an alternating, smooth protocol σ', such that $\sigma'(xyrm')$ γ-simulates $\sigma(xyrm)$ and $\|\sigma'\| \leq 2s\|\sigma\| \log(\|\sigma\|/\gamma) = 2s\|\pi\| \log(\|\pi\|/\gamma)$, for a sufficiently large constant s. Therefore, if we denote the distribution $q := \pi(i x_i y_i r_i m|W)$, the second proposition of Lemma 28 and the triangle inequality imply that σ' 3γ-simulates q. In particular, there there is a function $g_{out}(xyrm') \to \{0,1\}$ mapping transcripts m' of σ' to output bits of transcripts m (of π), such that $\sigma'(g_{out}(xyrm')) \overset{3\gamma}{\approx} q(m_{out})$. Since $\mathsf{suc}(\mu, f, q) = \Pr_q[m_{out} = f] = \frac{1}{n} \sum_i \pi(W_i|W)$, This fact ensures that

$$\mathsf{suc}(\mu, f, \sigma') \geq \sum_{i \in [n]} \pi(W_i|W) - 3\gamma. \tag{11}$$

We may now apply Theorem 3 (setting $\epsilon = 3\gamma$, $I_q = 4\|\pi\|/n$) to conclude that there exists a protocol τ that 45γ-simulates σ', and a large enough constant κ such that

$$\begin{aligned} \mathsf{IC}(\tau) &\leq \kappa \cdot \frac{I_q + \log(\|\sigma'\| + 1)}{\epsilon^2} = \kappa \cdot \frac{4\|\pi\|/n + \log(2s\|\pi\| \log(\|\pi\|/\gamma) + 1)}{9\gamma^2} \\ &\leq \kappa \cdot \frac{4s\|\pi\| \log(\|\pi\| \log(\|\pi\|/\gamma))}{9\gamma^2 n} < \frac{4\kappa s \cdot \alpha n \gamma^2 I}{9\gamma^2 n} \leq I \end{aligned} \tag{12}$$

[8] And this is the main challenge overcame in [BRWY12], namely, that this non-markov distribution can be approximated by an *actual* protocol σ.

by (10) and setting $\alpha = \frac{9}{4\kappa s}$. Finally, since τ 45γ-simulates σ', (12) implies that

$$\begin{aligned}
&\mathsf{suc}^{\mathsf{i}}(\mu, f, I) \geq \mathsf{suc}(\mu, f, \tau) \geq \mathsf{suc}(\mu, f, \sigma') - 45\gamma \\
&\geq \frac{1}{n} \sum_{i \in [n]} \pi(W_i|W) - 48\gamma \qquad \text{(By (11))} \\
&= \frac{1}{n} \sum_{i \in [n]} \pi(W_i|W) - \gamma/\beta \quad \text{(by choice of } \beta = 1/48).
\end{aligned}$$

□

Remark 29. *We note that the proof of Theorem 2 never used the fact that f was Boolean. Indeed, the theorem holds for arbitrary functions $\mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, by modifying the output function q_{out} in the proof above so that $g_{out}(xyrm') \to |\mathcal{Z}|$.*

7 Information-theoretically secure communication

In this section we prove Theorem 31, which shows how the information odometer can be used to "transform" any communication protocol into an information-theoretically secure one, even against malicious players, at the price of a mild overhead in the information (and communication) cost. We now make this precise. The following definition is crucial to our proof.

Definition 30 (Live communication protocols)**.** *A live communication protocol is a protocol π using independent public randomness R_v at pre-specified nodes v of the protocol tree of π. The realization of R_v on the public tape is revealed only upon reaching that node.*

We are now ready to state Theorem 31 at its full version.

Theorem 31. *Let θ be a two-party communication protocol such that $\mathsf{IC}(\theta) = I$. Then for any $\delta > 0$, there is a live communication protocol $\tilde{\pi}$ with the following properties:*

- *If both parties are honest, then $\tilde{\pi}$ 2δ-simulates θ.*
- $\mathsf{IC}(\tilde{\pi}) \leq \tilde{I} = O(I/\delta + \log(\|\theta\|))$.
- *There is a global constant $\lambda > 0$ such that for any protocol $\tilde{\pi}'$ where at least one party is honest (follows $\tilde{\pi}$), the following holds:*

$$\forall\ k \in \mathbb{N} \qquad \Pr[\mathbb{D}_X^{\tilde{\pi}'}(\tilde{\Pi}') > \lambda k(I/\delta + \log(\|\theta\| + 1))] \leq 2^{-\Omega(k)} \quad \textit{if Alice is honest.}$$

$$\forall\ k \in \mathbb{N} \qquad \Pr[\mathbb{D}_Y^{\tilde{\pi}'}(\tilde{\Pi}') > \lambda k(I/\delta + \log(\|\theta\| + 1))] \leq 2^{-\Omega(k)} \quad \textit{if Bob is honest.}$$

The protocol $\tilde{\pi}$ does not assume any prior knowledge about the honesty of any player.

Before we present the proof, a few remarks are in order:

1. The second proposition concerns only the information revealed by the honest player. Indeed, it is impossible (and nonsensical) to guarantee anything about the information revealed by a dishonest player, as there is nothing preventing him from sending bits from his input as long as the protocol is running.

2. Our secure protocol makes a crucial use of *live* public randomness. This resource seems inevitable for security purposes, as if the public random string is realized in the beginning of the protocol, the protocol is essentially a distribution on deterministic protocols, and from this point on, the malicious party may choose paths of the protocol allowing it to "cheat" the honest party. This will become clearer throughout the proof.

Proof. Without loss of generality, we may assume that θ is alternating and smooth since Lemma 23 asserts that smoothing can only increase the information cost of the protocol by a factor of 2 (say). We emphasize that we will only assume smoothness for honest players messages (while a dishonest player is free to send arbitrary messages).

Recall the (single-round) odometer protocol τ from Figure 1. The parties will use a similar protocol $\tilde{\tau}$ as a proxy for estimating the internal information revealed up to each step of the simulation. $\tilde{\tau}$ is a variation of the protocol τ with the minor (yet crucial) twist that in each round, the order of speakers is chosen at random using *live* randomness. The protocol $\tilde{\tau}$ is defined in Figure 5. Let $\tilde{\tau}_\ell$ be the ℓ-bit simulation of $\tilde{\tau}$, analogues to τ_ℓ from Corollary 20.

The protocol $\tilde{\tau}$

1. Given p, Alice samples a number $Z \in [0,1]$, according to the following probability density function:
$$\mu_p(z) = \begin{cases} 4(p-z) & \text{if } 0 \le z < p \\ 4(z-p) & \text{if } p \le z \le p+1/2 \\ 2-4(z-p-1/2) & \text{if } p+1/2 < z \le 1 \end{cases}$$
2. Alice sends Z_p to Bob.
3. The players use *live* public randomness to flip an unbiased coin G.
4. If $G = 0$,
   (a) Alice sends Bob a bit I^p indicating whether "$Z_p > p$".
   (b) Bob sends a bit I^q indicating whether "$Z_p > q$".

   If $G = 1$,
   (a) Bob sends a bit I^q indicating whether "$Z_p > q$".
   (b) Alice sends Bob a bit I^p indicating whether "$Z_p > p$".
5. The players output "1" iff $I^p \neq I^q$.

Figure 5: A protocol for estimating the internal information cost of a single round of communication.

To simulate θ, the parties run is the simulation protocol π from Section 5.2, except that in Step 2, whenever τ_ℓ needs to be invoked, the parties instead invoke $\tilde{\tau}_\ell$. The overall protocol $\tilde{\pi}$ is defined in Figure 6.

Notice that τ_ℓ and $\tilde{\tau}_\ell$ have the exact same outcome for honest players (since the order of speakers does not change the transcript), so in this case $\tilde{\pi}$ is equivalent to the protocol π from Figure 4 (with

The protocol $\tilde{\pi}$

The parties run the protocol π from Figure 4, with the following two changes:

- $\rho \longleftarrow I/\delta$, $\alpha \longleftarrow \ln(1/\delta)/\rho$.
- At each round t the parties replace Step 2 of π with Step 2′: With probability α, invoke the protocol $\tilde{\tau}_\ell$ (setting p_t, q_t as before). At each round the parties use *live* public randomness.

Figure 6: A secure simulation of θ.

the appropriate parameters α, ρ). Therefore, The first proposition of the theorem follows from the proof of Claim 42, noting that now $\theta(\mathbb{D}^\theta_{xyr}(m) > \rho) = \theta(\mathbb{D}^\theta_{xyr}(m) > I/\delta) \leq \delta$ by Markov's inequality and non-negativity of divergence-cost, and $e^{-\alpha\rho} = \delta$.

Note that the second proposition of the theorem would follow from the third proposition, recalling (4) and the fact $\mathbb{E}[Y] = \sum_{i=1}^{\infty} \Pr[Y > i]$, and therefore the rest of our analysis (and the main effort of the proof) is devoted to the third proposition, where one party is assumed to be dishonest. Suppose without loss of generality that Alice is honest and Bob is not (note that this assumption is only for the sake of analysis). Formally, this means that the players are executing a protocol $\tilde{\pi}'$ of the following form (over inputs $x, y \sim \mu$ and using (live) public randomness r):

$$\begin{aligned} \tilde{\pi}'(m_t | xym_{<t}r_{<t}) &= \tilde{\pi}(m_t | xm_{<t}r_{<t}) \quad \text{if Alice is the sender of } m_t \\ \tilde{\pi}'(m_t | xym_{<t}r_{<t}) &= g_t(m_t | ym_{<t}r_{<t}) \quad \text{if Bob is the sender of } m_t, \end{aligned}$$

where $g_t : Y \times M_{<t} \times R_{<t} \longrightarrow \{0,1\}$ is an arbitrary boolean function which Bob may use.

The key step of our analysis of $\tilde{\pi}'$ is showing that the protocol $\tilde{\tau}$ cannot be manipulated by a dishonest player, in the sense that the probability of $\tilde{\tau}$ outputting 1 in a certain round remains comparable to the information learnt by the dishonest player (Bob) in this round. This ensures that the random variable *Count* still provides Alice an indication of the information she revealed so far to Bob. We now turn to formalize the above.

Let $\mathbf{1}^t_{\tilde{\tau}_\ell}$ be the indicator variable denoting the output of protocol $\tilde{\tau}_\ell$ in round t (providing that it was invoked). Let G_t denote the (live) random coin used in Step 4 of $\tilde{\tau}$. Define $\gamma^t_1 := \Pr(\mathbf{1}^t_{\tilde{\tau}_\ell} \mid G_t = 1)$, and $\gamma^t_0 := \Pr(\mathbf{1}^t_{\tilde{\tau}_\ell} \mid G_t = 0)$. Note that

$$\Pr(\mathbf{1}^t_{\tilde{\tau}_\ell}) = \gamma^t_0/2 + \gamma^t_1/2. \tag{13}$$

Recall that the players are using live public randomness at each round of $\tilde{\pi}$ – let us denote the randomness up to round t by $r_{<t}$. Let $h_{<t} := y\tilde{\pi}_{<t}r_{<t}$ denote the entire knowledge Bob has before round t of $\tilde{\pi}'$ was executed, and as usual let m denote the messages of θ sent in $\tilde{\pi}'$. For a round t where Alice speaks, define

$$q_t := \tilde{\pi}(M_t = 1 \mid h_{<t}) \ , \ p_t := \tilde{\pi}(M_t = 1 \mid xh_{<t}) \ , \ \text{and} \ \ \mu(p_t) := \tilde{\pi}(p_t \mid h_{<t}). \tag{14}$$

when Bob speaks define the analogues quantities with x replaced by y. Then

$$q_t = \int_x \tilde{\pi}(M_t = 1 \mid xh_{<t}) \cdot \tilde{\pi}(x \mid h_{<t}) \, dx = \int_{p_t} p_t \mu(p_t) \, dp_t \tag{15}$$

by changing the variable of integration. The following Lemma is the heart of the proof.

Lemma 32. *Suppose $\tilde{\tau}_\ell$ is invoked at round t of $\tilde{\pi}$, and denote by T the transcript of $\tilde{\tau}_\ell$. Then*

- *If the receiver of m_t is the dishonest player, then $\gamma_1^t \geq \mathbb{E}_{p_t,p_q}(p_t - q_t)^2 - 2^{-\ell}$. Furthermore, the information learnt by the (dishonest) receiver is*

$$\mathbb{E}_{xy\tilde{\pi}_{<t}r_{<t}}\left[\mathsf{D}\left(\frac{(T_t|xyr_{<t}\tilde{\pi}_{<t})}{\!\!=\!\!(T_t|yr_{<t}\tilde{\pi}_{<t})}\right)\right] \leq c_\tau \cdot \mathbb{E}_{p_t,q_t}\left[H((p_t-q_t)^2)\right],$$

  where c_τ is the constant from Corollary 20.

- *If the sender of m_t is the dishonest player, then*

$$\mathbb{E}_{xy\tilde{\pi}_{<t}r_{<t}}\left[\mathsf{D}\left(\frac{(T_t|xyr_{<t}\tilde{\pi}_{<t})}{\!\!=\!\!(T_t|yr_{<t}\tilde{\pi}_{<t})}\right)\right] \leq H(\gamma_0^t).$$

Proof. Let us begin with the (easier) case where the dishonest player (Bob) is the *sender* in $\tilde{\tau}_\ell$. For simplicity of notation, denote throughout the proof $p = p_t, q = q_t$. Suppose that Bob sends messages Z', B in $\tilde{\tau}$ (an honest Bob will send Z_p, I^p). Recall that $\gamma_0^t = \Pr(\mathbf{1}^t_{\tilde{\tau}_\ell}|\text{“}G_t = 0\text{”})$. Note that when $G_t = 0$, the random variable $\mathbf{1}^t_{\tilde{\tau}_\ell}$ determines I^q (more precisely, the sender can determine I^q given $\mathbf{1}^t_{\tilde{\tau}_\ell}$ and Z', B which he knows). Therefore, the information that $\tilde{\tau}_\ell$ conveys to Bob in this case can be upper bounded by

$$\begin{aligned}
&\mathbb{E}_{xy\tilde{\pi}_{<t}r_{<t}}\left[\mathsf{D}\left(\frac{(T_t|xyr_{<t}\tilde{\pi}_{<t})}{\!\!=\!\!(T_t|yr_{<t}\tilde{\pi}_{<t})}\right)\right] = I(T_t; X|YR_{<t}\tilde{\Pi}_{<t}\text{“}G_t = 0\text{”}) \\
&= I(I^q; X|YR_{<t}\tilde{\Pi}_{<t}\text{“}G_t = 0\text{”}Z'B) \leq H(I^q|\text{“}G_t = 0\text{”}Z'B) \\
&\leq H(\mathbf{1}^t_{\tilde{\tau}_\ell}|G_t = 0) = H(\gamma_0^t)
\end{aligned}$$

by definition of γ_0^t and since conditioning reduces entropy. We turn to the case $G_t = 1$. We claim that in fact

$$I(T_t; X|YR_{<t}\tilde{\Pi}_{<t}\text{“}G_t = 0\text{”}) = I(T_t; X|YR_{<t}\tilde{\Pi}_{<t}\text{“}G_t = 1\text{”}).$$

This is true since, by construction of $\tilde{\tau}$, Alice's message I^q only depends on Z' which is sent *before* G_t is determined, hence the distributions $\tilde{\tau}_\ell(I^q \mid \text{“}G_t = 0\text{”})$ and $\tilde{\tau}_\ell(I^q \mid \text{“}G_t = 1\text{”})$ are equal. Therefore, we conclude that the internal information cost of the message I^q is at most $H(\gamma_0^t)$, as desired. Note that, as promised, the above argument did not use any smoothness assumption on the messages of the malicious player. This proof shows that if the dishonest sender tries to “cheat” by lowering the success rate of $\tilde{\tau}_\ell$ (thus preventing *Count* from increasing), then he will also learn very little by this attempt. This is because he needs to “commit” for Z' before he knows the order of the steps.

We turn to analyze the case where (dishonest) Bob is the *receiver* in $\tilde{\tau}_\ell$. Throughout the proof we analyze the original (non-truncated) protocol $\tilde{\tau}$, and then sketch how to obtain the desirable guarantees for $\tilde{\tau}_\ell$. First, note that in this case the messages Z_p, I^p of Alice are not affected by the receiver's message (regardless of whether $G = 0$ or 1). Therefore, the amount of information learnt by the dishonest receiver follows from the exact same analysis as in the honest case, which by the second proposition of Theorem 1 is at most

$$c_\tau \cdot \mathbb{E}_{p_t,q_t}\left[H((p_t-q_t)^2)\right]. \tag{16}$$

Note that the the application of Theorem 1 only requires the *sender's* message (which in this case is the *honest* player) to be smooth (this is enough since $p_t \in \{1/3, 2/3\}$ implies the same for q_t, as $q_t = \mathbb{E}[p_t]$. For a more detailed explanation, see the proof of Lemma 23). We now argue about the "success" probability of $\tilde{\tau}$ when the receiver is dishonest. Note that the receiver only sends a single bit in $\tilde{\tau}$ (the honest receiver will send I^q). If $G_t = 0$ he goes second, so he can always report $b = I^p$ and make the test fail so $\gamma_0^t = 0$ in this case. However, if $G_t = 1$, then his message b is only a function of $h_{<t}$ and z. We can therefore model the receiver's message as a function $B : Z_p \times H_{<t} \longrightarrow \{0, 1\}$. For any such strategy B, let

$$\gamma_1^t(B) := \Pr(\mathbf{1}_{\tilde{\tau}}^t = 1 \mid G_t = 1, \text{receiver sends } B(Z_p, H_{<t})).$$

Clearly, $\gamma_1^t \geq \inf_B\{\gamma_1^t(B)\}$.
Let B^* be the "honest player" strategy $B^* = \mathsf{1}_{q<z}$. We shall prove the following claim:

Claim 33. *For any receiver strategy B, $\gamma_1^t(B) \geq \frac{1}{2} \cdot \gamma_1^t(B^*)$.*

Before we prove the this claim, let us see how it finishes the proof of the Lemma. Note that if the receiver uses B^*, the outcome of the protocol $\tilde{\tau}$ is the same as that of τ. Therefore, it follows from the second proposition of Theorem 1 that $\gamma_1^t(B^*) = 2(p-q)^2$. Finally, Claim 33 implies that

$$\gamma_1^t \geq \frac{1}{2} \cdot \gamma_1^t(B^*) = \frac{1}{2} \cdot 2(p-q)^2 = (p-q)^2. \tag{17}$$

To obtain the promised guarantee on $\tilde{\tau}_\ell$, note that if the receiver has a strategy B' in $\tilde{\tau}_\ell$ such that $\gamma_1^t(B') < (p-q)^2 - 2^{-\ell}$, then the success probability of B' under $\tilde{\tau}$ is $< (p-q)^2$, contradicting (17). This is because $\Pr(Z_p < p < Z') \leq 2^{-\ell}$, and otherwise the output of $\tilde{\tau}$ and $\tilde{\tau}_\ell$ are the same (See the proof of Corollary 20). Therefore, it must hold that $\gamma_1^t \geq (p-q)^2 - 2^{-\ell}$, as claimed. Moreover, the data processing inequality ensures that the information bound in (16) continues to hold if we replace $\tilde{\tau}$ by $\tilde{\tau}_\ell$, as Z_ℓ is determined by Z_p.

It thus remains to prove Claim 33:

Proof of Claim 33. Fix $z, h_{<t}$. Recall the definition of $\mu_p(z)$ from Figure 5, and let $\mu_z(p) := \tilde{\pi}(p \mid zh_{<t})$ be the "inverse" distribution of $\mu_p(z)$ (the distribution of p given z). If the receiver reports $B(z, h_{<t}) = 0$ ("$q < z$"), then his answer will be inconsistent with the (honest) sender's bit I^p exactly when $p > z$, which happens with probability

$$s_0 := \int_p \mu_z(p) \cdot \mathsf{1}_{p>z}\, dp = \frac{1}{K(z)} \int_p \mu_p(z) \cdot \mu(p) \cdot \mathsf{1}_{p>z}\, dp, \tag{18}$$

where we used the (continuous version of) Bayes rule and $K(z) = \int_p \mu_p(z)\mu(p)\, dp$. Similarly, if the receiver reports $B(z, h_{<t}) = 1$ ("$q > z$"), then his answer will be inconsistent with I^p with probability

$$s_1 := \int_p \mu_z(p) \cdot \mathsf{1}_{p<z}\, dp = \frac{1}{K(z)} \int_p \mu_p(z) \cdot \mu(p) \cdot \mathsf{1}_{p<z}\, dp \tag{19}$$

Thus,

$$\gamma_1^t(B(z, h_{<t})) = \min\{s_0, s_1\}. \tag{20}$$

Define

$$\tilde{s}_0 := \frac{4}{K(z)} \int_p |z-p| \cdot \mu(p) \cdot \mathsf{1}_{p>z}\, dp \quad , \quad \tilde{s}_1 := \frac{4}{K(z)} \int_p |z-p| \cdot \mu(p) \cdot \mathsf{1}_{p<z}\, dp.$$

Proposition 34. *$s_0 \le \tilde{s}_0 \le 2s_0$ and $s_1 \le \tilde{s}_1 \le 2s_1$.*

Proof. We prove the statement for s_0 (an analogues argument applies for s_1). Note that $\frac{\tilde{s_0}}{s_0} = \frac{4|z-p|}{\mu_p(z)}$, so it suffices to prove that $2|z-p| \le \mu_p(z) < 4|z-p|$. Indeed, suppose w.l.o.g that $p \le 1/2$. By definition, $\mu_p(z) = 4|z-p|$ in the region $z \le p + 1/2$ (see Figure 5), so it only remains to handle the region $z > p + 1/2$. In this case,

$$\mu_p(z) - 4|z-p| = 2 - 4(z-p-1/2) - 4(z-p) = 8(p+1/2-z) < 0,$$

since $z > p$ in this region and therefore $|z-p| = z-p$. Thus $\mu_p(z) \le 4|z-p|$. On the other hand, since $p \ge 1/3$ by the smoothness assumption, we have

$$\mu_p(z) - 2|z-p| = 2 - 4(z-p-1/2) - 2(z-p) = 4 + 6p - 6z \ge 4 + 6 \cdot \frac{1}{3} - 6 = 0.$$

Rearranging sides concludes the proof. □

Combining (20) with Proposition 34, we get that

$$\gamma_1^t(B(z, h_{<t})) \ge \frac{1}{2} \min\{\tilde{s}_0, \tilde{s}_1\}.$$

To finish the proof, it remains to show that the strategy choosing $\min\{\tilde{s}_0, s_1\}$ is equivalent to $B^* = 1_{q<z}$. Indeed,

$$\begin{aligned}
\tilde{s}_1 - \tilde{s}_0 &= \frac{4}{K(z)} \left(\int_p |z-p| \cdot \mu(p) \cdot 1_{p<z} \, dp - \int_p |z-p| \cdot \mu(p) \cdot 1_{p>z} \, dp \right) = \\
&= \frac{4}{K(z)} \int_p (z-p) \cdot sgn(z-p) \cdot \mu(p) \, dp = \frac{4}{K(z)} \int_p (z-p) \cdot \mu(p) \, dp \\
&= \frac{4}{K(z)} \left[z - \left(\int_p p\mu(p) \, dp \right) \right] = \frac{4}{K(z)} \cdot (z-q), \qquad (21)
\end{aligned}$$

where the last transition follows from (15). Finally, note that the expression in (21) is positive iff $\tilde{q} > z$, and thus $\gamma_1^t(B^*(z, h_{<t})) = \min\{\tilde{s}_0, s_1\}$. Since the argument holds for any $z, h_{<t}$, we conclude that for any strategy B it holds that $\gamma_1^t(B) \ge \frac{1}{2}\gamma_1^t(B^*)$, which finishes the proof of Claim 33.

□

□

With Lemma 32 in hand, we are now ready to argue about the amount of information revealed by the honest player in $\tilde{\pi}'$. As before, let us denote the transcript of $\tilde{\pi}'$ by $\tilde{\Pi}' = M_{<C} T_{\le C}$, where M is the transcripts of θ and T is the concatenation of transcripts of $\tilde{\tau}$ in all rounds where it was executed. Define

$$\mathbb{D}_X^{\tilde{\pi}'}(T_{\le j}) := \alpha \cdot \sum_{t=1}^{j} \left[\mathsf{D} \left(\frac{T_t | XYR_{<t} \tilde{\Pi}'_{<t}}{T_t | YR_{<t} \tilde{\Pi}'_{<t}} \right) \right] \qquad (22)$$

Note that, unlike the analysis of Theorem 3, conditioning the above information on the entire history $\tilde{\Pi}'_{<t} = M_{<t} T_{<t}$ (and not just the messages $M_{<t}$ of θ) is now mandatory, since, as discussed

above, in $\tilde{\pi}'$ the dishonest player may choose his messages based on previous outcomes of $\tilde{\tau}_\ell$, and thus the information he learns is measured respectively. In analogy with (22), for the divergence cost of messages of θ sent in $\tilde{\pi}'$, we slightly abuse the notation to redefine

$$\mathbb{D}_X^{\tilde{\pi}'}(M_{\leq j}) := \sum_{t=1}^{j} \mathsf{D}\left(\frac{M_t|XYR_{<t}\tilde{\Pi}'_{<t}}{M_t|YR_{<t}\tilde{\Pi}'_{<t}}\right).$$

Note that by the the chain rule we have

$$\mathbb{D}_X^{\tilde{\pi}'}(\tilde{\Pi}'_{\leq j}) = \mathbb{D}_X^{\tilde{\pi}'}(M_{\leq j}) + \mathbb{D}_X^{\tilde{\pi}'}(T_{\leq j}). \tag{23}$$

We shall therefore bound the value $\mathbb{E}[\mathbb{D}_X^{\tilde{\pi}'}(M_{\leq j})] + \mathbb{E}[\mathbb{D}_X^{\tilde{\pi}'}(T_{\leq j})]$ in terms of the value of $Count$, in the same manner as the proof of Theorem 3. To this end, define for any $j \in [\|\theta\|]$

$$L_j := \sum_{t=0}^{j} \mathbf{1}^t_{\tau_\ell} - \frac{\alpha}{18} \cdot \mathbb{D}_X^{\tilde{\pi}'}(M_{\leq j}) + j \cdot \alpha \cdot 2^{-\ell} - \frac{\mathbb{D}_X^{\tilde{\pi}'}(T_{\leq j}) - c_\tau \alpha j \cdot H(2^{-\ell}) - j/(\|\theta\|+1)}{8 \cdot 4c_\tau \cdot \log(2 \cdot 4c_\tau(\|\theta\|+1))}. \tag{24}$$

The next claim asserts that $\mathcal{L} := \{L_j\}_{j=0}^{\|\theta\|+1}$ is a sub-martingale:

Claim 35. $\mathbb{E}[L_j \mid L_{j-1}] \geq L_{j-1}$.

Proof. This claim is an analogue of Claim 46 from Section 5.2. Set:

- $K = \|\theta\| + 1$.
- $X_t = \mathbf{1}^t_{\tilde{\tau}_\ell}$.
- $Y_t = \mathsf{D}\left(\frac{T_t|XYR_{<t}\tilde{\Pi}'_{<t}}{T_t|YR_{<t}\tilde{\Pi}'_{<t}}\right) - c_\tau\alpha \cdot H(2^{-\ell})$.
- $Z_t = \mathsf{D}\left(\frac{M_t|XYR_{<t}\tilde{\Pi}'_{<t}}{M_t|YR_{<t}\tilde{\Pi}'_{<t}}\right) - 18 \cdot 2^{-\ell}$.
- $\lambda = \alpha/18$.
- $\beta = 4c_\tau$.

Using the chain rule for divergence and the fact that $\tilde{\tau}_\ell$ is executed with probability α at each round, (24) can be rewritten as

$$L_j := \sum_{t=1}^{j} X_t - \lambda \cdot \sum_{t=1}^{j} Z_t - \frac{\sum_{t=1}^{j} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)}. \tag{25}$$

We are thus in the setup of Lemma 44. To apply the Lemma, we must show that:

1. $\mathbb{E}[Y_t|L_{t-1}] \leq \beta \cdot H(\mathbb{E}[X_j|L_{t-1}])$.
2. $\lambda \cdot \mathbb{E}[Z_t|L_{t-1}] \leq \frac{1}{2} \cdot \mathbb{E}[X_j|L_{t-1}]$.

First proposition : Recall that $\tilde{\tau}_\ell$ is executed w.p α at each round, and otherwise $Y_t \mid L_{t-1} = 0$. If, in round t, the dishonest player is the receiver, the first proposition of Lemma 32 implies that

$$\begin{aligned}
&\mathbb{E}[Y_t \mid L_{t-1}] = \alpha \cdot \mathbb{E}\left[\mathsf{D}\left(\frac{T_t|xyr_{<t}\tilde{\pi}'_{<t}}{T_t|yr_{<t}\tilde{\pi}'_{<t}}\right)\right] - c_\tau\alpha \cdot H(2^{-\ell}) \\
&\leq \alpha c_\tau \cdot \mathbb{E}_{p_t,q_t}\left[H((p_t-q_t)^2)\right] - c_\tau\alpha \cdot H(2^{-\ell}) \\
&\leq 2c_\tau\alpha \cdot (H(\mathbb{E}(p_t-q_t)^2) - H(2^{-\ell})) \quad \text{(by the same calculation as in (42))} \\
&\leq 2c_\tau\alpha \cdot H(\gamma_1^t), \qquad (26)
\end{aligned}$$

where the second before last transition follows from Proposition $13/(iv)$ taken with $x = (p_t - q_t)^2$,$y = \gamma_1^t$ and $\epsilon = 2^{-\ell}$, and the first proposition of Lemma 32 which implies $|\gamma_1^t - \mathbb{E}(p_t - q_t)^2| \leq 2^{-\ell}$. The last transition follows from Proposition $13/(iii)$.
Otherwise, if the dishonest player is the sender in round t, the second proposition of Lemma 32 implies that

$$\mathbb{E}[Y_t \mid L_{t-1}] = \alpha \cdot \mathbb{E}\left[\mathsf{D}\left(\frac{T_t|xyr_{<t}\tilde{\pi}'_{<t}}{T_t|yr_{<t}\tilde{\pi}'_{<t}}\right)\right] \leq \alpha H(\gamma_0^t). \qquad (27)$$

Recall by (13) that $\mathbb{E}[\mathbf{1}^t_{\tilde{\tau}_\ell}] = \gamma_0^t/2 + \gamma_1^t/2$. Since entropy is nonnegative and concave, we have

$$\begin{aligned}
&\max\{H(\gamma_0^t)\ ,\ H(\gamma_1^t)\} \leq H(\gamma_0^t) + H(\gamma_1^t) = 2(H(\gamma_0^t)/2 + H(\gamma_1^t)/2) \\
&\leq 2H(\gamma_0^t/2 + \gamma_1^t/2) = 2H(\mathbb{E}[\mathbf{1}^t_{\tilde{\tau}_\ell}]). \qquad (28)
\end{aligned}$$

Thus, combining (26),(27) and (28), we conclude that in both cases

$$\mathbb{E}[Y_t \mid L_{t-1}] \leq 4c_\tau\alpha \cdot H(\mathbb{E}[\mathbf{1}^t_{\tilde{\tau}_\ell}]) \leq 4c_\tau H(\alpha \cdot \mathbb{E}[\mathbf{1}^t_{\tilde{\tau}_\ell}]) = 4c_\tau \cdot H(\mathbb{E}[X_t|L_{t-1}]) = \beta \cdot H(\mathbb{E}[X_t|L_{t-1}])$$

where in the second transition we used Proposition $13/(ii)$ with $y = \alpha$.
Second proposition : Note that when the dishonest player is the sender of m_t, $\mathbb{E}[Z_t \mid L_{t-1}] = I(M_t; X|YR_{<t}\tilde{\Pi}'_{<t}) = 0$ (by definition of a protocol), so in this case the statement trivially follows as $X_t \geq 0$. Otherwise, when the dishonest player is the receiver of m_t, the first proposition of Lemma 32 guarantees that $\mathbb{E}[\mathbf{1}^t_{\tau_\ell}] \geq \frac{\gamma_1^t}{2} \geq \frac{1}{2} \cdot ((p_t - q_t)^2 \pm 2^{-\ell})$. We therefore have

$$\begin{aligned}
&\lambda \cdot \mathbb{E}[Z_t \mid L_{t-1}] = \frac{\alpha}{18}\left(\mathbb{E}\left[\mathsf{D}\left(\frac{M_t|xyr_{<t}\tilde{\pi}'_{<t}}{M_t|yr_{<t}\tilde{\pi}'_{<t}}\right)\right] - 18 \cdot 2^{-\ell}\right) \\
&= \frac{\alpha}{18}\left(\mathsf{D}\left(\frac{p_t}{q_t}\right) - 18 \cdot 2^{-\ell}\right) \quad \text{(By definition of } p_t, q_t \text{ in (14))} \\
&\leq \frac{\alpha}{18}\left(\frac{9}{2} \cdot (p_t - q_t)^2 - 18 \cdot 2^{-\ell}\right) \\
&\text{(By Proposition 14 and smoothness of the honest player's messages)} \\
&= \frac{\alpha \cdot \left(\frac{1}{2}(p_t - q_t)^2 - 2^{-\ell}\right)}{2} \leq \frac{\alpha \cdot \frac{\gamma_1^t}{2}}{2} \leq \frac{\alpha \cdot \mathbb{E}[\mathbf{1}^t_{\tau_\ell}]}{2} = \frac{1}{2} \cdot \mathbb{E}[X_t \mid L_{t-1}]. \qquad (29)
\end{aligned}$$

We may now apply Lemma 44 to conclude that $\mathcal{L}$ is a sub-martingale, which completes the proof. □

To prove the third proposition of Theorem 31, we shall show that the value of the sub-martingale $\mathcal{L}$ at the stopping time C is sharply concentrated. This is the content of the next claim:

Claim 36. *For all $k \geq 1$, it holds that* $\Pr[L_C < -k] \leq 2^{-\Omega(k)}$.

To see how Claim 36 finishes the proof, note that $L_C \geq -k$ implies

$$\sum_{t=0}^{C} \mathbf{1}^t_{\tilde{\tau}_\ell} - \frac{\alpha}{18} \cdot \mathbb{D}^{\tilde{\pi}'}_X(M_{\leq C}) + C \cdot \alpha \cdot 2^{-\ell} - \frac{\mathbb{D}^{\tilde{\pi}'}_X(T_{\leq C}) - c_\tau \alpha C \cdot H(2^{-\ell}) - C/(\|\theta\|+1)}{8 \cdot 4c_\tau \cdot \log(2 \cdot 4c_\tau(\|\theta\|+1))} > -k. \quad (30)$$

Applying the same calculation as in (36) and recalling that the value of *Count* is at most $\lceil 4\alpha\rho \rceil$ at the aborting index $C \leq (\|\theta\|+1)$, we can rearrange (30) to obtain

$$\frac{\alpha}{18} \cdot \mathbb{D}^{\tilde{\pi}'}_X(M_{\leq C}) + \frac{\mathbb{D}^{\tilde{\pi}'}_X(T_{\leq C})}{8 \cdot 4c_\tau \cdot \log(2 \cdot 4c_\tau(\|\theta\|+1))} \leq \lceil 4\alpha\rho \rceil + \alpha + 1 + k \leq 5\alpha\rho + k.$$

Since both terms in the LHS are non-negative, each of them must be at most $5\alpha\rho + k$. Recalling that $\alpha = \ln(1/\delta)/\rho$, we get:

- $\frac{\alpha}{18} \cdot \mathbb{D}^{\tilde{\pi}'}_X(M_{\leq C}) \leq 5\alpha\rho + k \implies \mathbb{D}^{\tilde{\pi}'}_X(M_{\leq C}) \leq 5\alpha\rho \cdot \frac{18}{\alpha} + \frac{18 \cdot k}{\alpha} = 18\rho(k+5)$.
- $\mathbb{D}^{\tilde{\pi}'}_X(T_{\leq C}) \leq (8 \cdot 4c_\tau \cdot \log(2 \cdot 4c_\tau(\|\theta\|+1))) \cdot 5\alpha\rho + k \leq O(k + \ln(1/\delta)) \log(\|\theta\|+1)$.

Thus, whenever $L_C \geq -k$, it holds that $\mathbb{D}^{\tilde{\pi}'}_X(M_{\leq C}) + \mathbb{D}^{\tilde{\pi}'}_X(T_{\leq C}) \leq O(k(I/\delta + \log(\|\theta\|+1)))$, (since $\rho = I/\delta$). Finally, by Claim 36 and (23) we conclude that for a sufficiently large constant $\lambda > 0$,

$$\Pr[\mathbb{D}^{\tilde{\pi}'}_X(\tilde{\Pi}') > \lambda k(I/\delta + \log(\|\theta\|+1))] \leq 2^{-\Omega(k)},$$

completing the entire proof of Theorem 31. It therefore remains to prove Claim 36.

Proof of Claim 36. For simplicity of notation, let us rewrite the sub-martingale $\mathcal{L}$ as in (25):

$$L_j = \sum_{t=1}^{j} X_t - \left(\lambda \cdot \sum_{t=1}^{j} Z_t + \frac{\sum_{t=1}^{j} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)} \right).$$

We would like to apply Azuma's inequality (Theorem 17) to $\mathcal{L}$. Unfortunately, it is easy to see that $|L_j - L_{j-1}|$ may be as large as ≈ 1 (e.g, if $X_j = 1$ and $Z_j \approx 0$, i.e, $\mathbf{1}^t_{\tilde{\tau}_\ell} = 0$ but $p_j \approx q_j$). With this bound Azuma's inequality will give a very weak concentration bound. To circumvent this problem, let us define the sub-process $\mathcal{L}' = \{L_{j_1}, L_{j_2}, L_{j_3}, \ldots\}$ such that

$$j_{i+1} = \min(C \ , \ \{j : j > j_i \ \ |L_j - L_{j_i}| > 1\}). \quad (31)$$

Claim 37. *$\mathcal{L}'$ is also a sub-martingale.*

Proof. For every j, the index j_{i+1} is a stopping rule with respect to the sub-martingale $\{L_{j+1}, L_{j+2}, \ldots\}$. Hence by the optional stopping theorem we have $\mathbb{E}[L_{j_{i+1}} \mid L_{j_i}] \geq L_{j_i}$. □

Now, observe that (31) implies that for every i,

$$c_i := |L_{j_{i+1}} - L_{j_i}| \leq 2 \quad \text{almost surely.} \quad (32)$$

Define a stopping rule for $\mathcal{L}'$ as follows: $C' = \min\{j_i = C \ , \ L_{j_i} < -k\}$.

Claim 38. $C' \leq \lceil 4\alpha\rho \rceil + k$.

Proof. If C' is such that $j_i = C$, then $\sum_{t=1}^{C'} X_t \leq \lceil 4\alpha\rho \rceil$ (this is the definition of the index C). Otherwise, $-\left(\lambda \cdot \sum_{t=1}^{C'} Z_t + \frac{\sum_{t=1}^{C'} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)}\right) < -k$ (since the contribution of X_t is nonnegative). But both expressions $\sum_{t=1}^{j} X_t$ and $-\left(\lambda \cdot \sum_{t=1}^{j} Z_t + \frac{\sum_{t=1}^{j} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)}\right)$ are *monotone* in t (see the definitions in Claim 35 and recall that divergence is non-negative). Therefore, $|L_{j_{i+1}} - L_{j_i}| > 1$ implies that in each step of $\mathcal{L}'$, either $\sum_{t=1}^{j_{i+1}} X_t$ increased by ≥ 1, or $-\left(\lambda \cdot \sum_{t=1}^{j_{i+1}} Z_t + \frac{\sum_{t=1}^{j_{i+1}} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)}\right)$ decreased by ≥ 1. The claim follows. □

We now apply Azuma's inequality (Theorem 17) to the sub-martingale $\mathcal{L}'$, setting $X_i = L_{j_i}$, $N = C'$ and $c_i = |L_{j_{i+1}} - L_{j_i}| \leq 2$. The theorem implies that for any $k \in \mathbb{R}^+$,

$$\Pr[L_C \leq -k] \leq \Pr[L_{C'} \leq -k] \leq \exp\left(-\frac{k^2}{2\sum_{i=0}^{C'} c_i^2}\right) \leq \exp\left(-\frac{k^2}{2 \cdot (\lceil 4\alpha\rho \rceil + k) \cdot 4}\right)$$
$$\leq \exp\left(-\frac{k^2}{8\ln(1/\delta) \cdot k}\right) \leq \exp\left(-\Omega(k)\right)$$

where the first transition follows from the fact that $L_{C'} < L_C$, and the third transition follows from Claim 38, and the definition of $\alpha = \ln(1/\delta)/\rho$. This concludes the proof of of the claim and the entire proof of Theorem 31. □

□

8 Towards better interactive compression?

In this section we discuss the implications of our construction for the interactive compression question. While we do not prove new compression results, our construction helps clarify the main challenges involved in improving the current state-of-the-art in compressing interactive communication, and suggests a "meta"-approach for making progress on this fascinating problem.

As mentioned in the introduction, the problem of compressing interactive communication can be summarized as follows: "Given a protocol π whose information cost $\mathsf{IC}(\pi, \mu)$ is I and whose communication cost is C, is there an equivalent — compressed — protocol π' that only uses $O(I)$ communication?". Note that if π is non-interactive then the answer to this question is positive [Huf52]. A more modest goal would be to compress π into a protocol π' that uses some function $g(I, C)$ of communication, such as $g(I, C) = poly(I) \cdot polylog(C)$. The compression question is closely related to the direct sum problem for randomized communication complexity. In fact, these questions are essentially equivalent to each other [BR11] — the better we can compress, the stronger direct sum holds for communication complexity

The two best general compression results to date are incomparable to each other. The first one, due to [BBCR10], gives $g(I, C) = \tilde{O}(C^{1/2} \cdot I^{1/2})$. The second one, due to [Bra12], gives $g(I, C) = 2^{O(I)}$, and was recently shown to be tight in a breakthrough result of Ganor et al. [GKR14]. Note that the second bound becomes non-trivial once $I \ll \log C$. More precisely, the compression scheme of [Bra12] starts with an information-I protocol, and produces a $2^{O(I/\varepsilon)}$-

communication protocol while failing with probability ε. Failing with probability ε is inevitable, since I is an average-case quantity, and thus with a small probability ε the information cost of π will be very high (potentially as high as I/ε) making it impossible to compress in less than $2^{O(I/\varepsilon)}$-communication with existing techniques. However, one can easily extend the compression result of [Bra12] to show that if we are given a π whose information cost is *uniformly* bounded by I (i.e. with high probability over paths taken by the protocol, the divergence cost is bounded by I), then one can compress π into $2^{O(I)}$ communication while only introducing a negligible amount of additional error:

Claim 39 (Adapted from [Bra12])**.** *Let $\rho, \epsilon > 0$ be error parameters, and let π an ϵ-error protocol for f, such that $\Pr_\mu[\mathbb{D}^\pi_{xyr}(m) > I] \leq \rho$. Then for any distribution μ, $CC^\mu_{\rho+\epsilon}(f) \leq 2^{O(I)}$.*

Claim 39 gives rise to the following strategy for compressing a protocol π: (1) partition π into pieces $\pi_1, \pi_2, \ldots$, such that each piece reveals only I_1 bits of information (thus the total number of pieces is $\sim I/I_1$); (2) compress each piece using $2^{O(I_1)}$ communication. Such a plan, if successful, would yield a total communication cost of $O(2^{O(I_1)} \cdot I/I_1)$. If one can make I_1 as small as $O(1)$, this would give a method for interactive compression.

Indeed, this strategy has been successfully carried out in [BBCR10] for compressing to *external information cost* of π. The external information cost $\mathsf{IC}^{ext}(\pi)$ of π measures the amount of information $\pi(X, Y)$ reveals about X, Y to an external observer. It is always the case that $\mathsf{IC}^{ext}(\pi) \geq \mathsf{IC}(\pi) = I$, and thus compressing a protocol to $\mathsf{IC}^{ext}(\pi) := I^{ext}$ is easier than compressing it to I. Since step (2) of the strategy is guaranteed by Claim 39, the main challenge is executing step (1). "Partitioning" means terminating π after $\sim I_1$ information has been revealed. This produces the first piece π_1. Then terminating after another $\sim I_1$ information is revealed produces π_2 etc. In the case of external information Alice can privately estimate the amount of information learnt by an external observed from her messages (since she has the ability to take the external observer's point of view). A similar statement holds about Bob. This allows Alice and Bob to partition the protocol into pieces of low ($O(1)$) *external* information cost, thus enabling compression of π to $O(I^{ext} \cdot polylog(C))$ communication.

Is a similar partitioning possible for internal information instead of external? This question is essentially equivalent to the odometer problem studied in this paper. In particular, we can use our odometer construction to pause the protocol π after $O(1)$ bits of information have been communication. Unfortunately, in the process we reveal an additive overhead of $O(\log C)$ bits of information, and thus the resulting information complexity of each part $\pi_1, \pi_2, \ldots$ is $O(\log C)$ rather than $O(1)$. Thus after applying step (2) of the compression plan we get a total communication cost of $I \cdot 2^{O(\log C)}$, which is not better than the original cost C. Unfortunately, [GKR14] asserts it is hopeless to improve the exponential dependence on I in the result of [Bra12], so this the latter approach will not work. Nevertheless, it is still hopeful to use the above approach to improve [BBCR10]'s compression result (see the following subsection). This statement can be generalized as follows: Each chunk π_i has information complexity $I_1 = O(\log C)$, and communication complexity $C_1 \leq C$. Therefore, if we could compress π_i into a protocol π'_i of communication complexity $g(I_1, C_1)$, the odometer will imply that any π can be compressed to $O(I \cdot g(I_1, C_1))$ communication. We thus obtain the following claim[9]:

[9]Since, at this point, this is a qualitative statement, we leave errors out of the statement to avoid complicating the notation.

Claim 40. *Suppose there is a compression protocol that takes as an input a protocol π_1 with communication cost C_1 and* worst case *information cost I_1, and compresses it into a protocol π_1' of communication complexity $g(C_1, I_1)$. Then a protocol π with communication cost C and information cost I can be compressed into a protocol with communication cost $\tilde{O}(I \cdot g(C, \log C))$.*

Claim 40 implies that it is sufficient to compress protocols whose information cost is logarithmic in their communication cost. In particular, if one could compress a protocol with communication cost C and information cost $\log C$ to a protocol with communication cost $g(C, \log C) = C^{o(1)}$, it would imply that *any* protocol with communication cost C and information cost I can be compressed to communication $I \cdot C^{o(1)}$. Note that both the scheme from [BBCR10] and [Bra12] yield only an upper bound of $g(C, \log C) = C^{O(1)}$ in this case.

Acknowledgement

We would like to thank Thomas Watson for his helpful comments and in particular for bringing to our attention the technical error in a previous version of this article, affecting the global constant c_τ in Theorem 1 (and the analysis in Lemma 19).

References

[ACC+12] Anil Ada, Arkadev Chattopadhyay, Stephen A. Cook, Lila Fontes, Michal Kouck?, and Toniann Pitassi. The hardness of being private. In *IEEE Conference on Computational Complexity*, pages 192–202. IEEE, 2012.

[AS92] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley–Interscience Series, John Wiley & Sons, Inc., New York, 1992.

[BBCR10] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *Proceedings of the 2010 ACM International Symposium on Theory of Computing*, pages 67–76, 2010.

[Bea89] Donald Beaver. Perfect privacy for two-party protocols. In J. Feigenbaum and M. Merritt, editors, *Proceedings of DIMACS Workshop on Distributed Computing and Cryptology*, volume 2, pages 65–77. American Mathematical Society, 1989.

[BOGW88] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation (extended abstract). In *STOC*, pages 1–10, 1988.

[BR11] Mark Braverman and Anup Rao. Information equals amortized communication. In Rafail Ostrovsky, editor, *FOCS*, pages 748–757. IEEE, 2011.

[Bra12] Mark Braverman. Interactive information complexity. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 505–524, New York, NY, USA, 2012. ACM.

[BRWY12] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct products in communication complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:143, 2012.

[BW14] Mark Braverman and Omri Weinstein. An interactive information odometer with applications. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:47, 2014.

[BYCKO93] Reuven Bar-Yehuda, Benny Chor, Eyal Kushilevitz, and Alon Orlitsky. Privacy, additional information, and communication. *IEEE Transactions on Information Theory*, 39:55–65, 1993.

[CK91] Benny Chor and Eyal Kushilevitz. A zero-one law for boolean privacy. *STOC 89 and SIAM J. Disc. Math*, 4:36–47, 1991.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. J. Wiley and Sons, New York, 1991.

[Doo75] J. L. Doob. Stochastic process measurability conditions. *Annales de l'institut Fourier*, 25(3-4):163–176, 1975.

[FJS10] Joan Feigenbaum, Aaron D. Jaggard, and Michael Schapira. Approximate privacy: Foundations and quantification (extended abstract). In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, pages 167–178, New York, NY, USA, 2010. ACM.

[GKR14] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication for boolean functions. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:113, 2014.

[GMW87] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game or a completeness theorem for protocols with honest majority. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 218–229, New York, NY, USA, 1987. ACM.

[Hol07] Thomas Holenstein. Parallel repetition: Simplifications and the no-signaling case. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, 2007.

[Huf52] D.A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[Jai11] Rahul Jain. New strong direct product results in communication complexity. 2011.

[JK09] Rahul Jain and Hartmut Klauck. The partition bound for classical communication complexity and query complexity. *CoRR*, abs/0910.4266, 2009.

[JPY12] Rahul Jain, Attila Pereszlenyi, and Penghui Yao. A direct product theorem for the two-party bounded-round public-coin communication complexity. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 167–176. IEEE, 2012.

[JY12] Rahul Jain and Penghui Yao. A strong direct product theorem in terms of the smooth rectangle bound. *CoRR*, abs/1209.0263, 2012.

[Kla02] Hartmut Klauck. On quantum and approximate privacy. In *STACS*, pages 335–346, 2002.

[Kla10] Hartmut Klauck. A strong direct product theorem for disjointness. In *STOC*, pages 77–86, 2010.

[KLL+12] Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao. Lower bounds on information complexity via zero-communication protocols and applications. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:38, 2012.

[KLX13] Iordanis Kerenidis, Mathieu Laurière, and David Xiao. New lower bounds for privacy in communication protocols. In *ICITS*, pages 69–89, 2013.

[Kus92] Eyal Kushilevitz. Privacy and communication complexity. *SIAM J. Discrete Math.*, 5(2):273–284, 1992.

[LSS08] Troy Lee, Adi Shraibman, and Robert Spalek. A direct product theorem for discrepancy. In *CCC*, pages 71–80, 2008.

[MMP+10] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil P. Vadhan. The limits of two-party differential privacy. In *FOCS*, pages 81–90, 2010.

[MNPS] D. Malkhi, N. Nisan, B. Pinkas, and Y. Sella. Fairplay - a secure two-party computation system. In *Proceedings of the 13th USENIX Security Symposium*, pages 287–302, Berkeley, CA, USA. USENIX Association.

[MWY13] Marco Molinaro, David Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *SODA*, page to appear, 2013.

[Pin03] Benny Pinkas. Fair secure two-party computation. In *EUROCRYPT*, pages 87–105, 2003.

[PRW97] Itzhak Parnafes, Ran Raz, and Avi Wigderson. Direct product results and the GCD problem, in old and new communication models. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing (STOC '97)*, pages 363–372, New York, May 1997. Association for Computing Machinery.

[Rao08] Anup Rao. Parallel repetition in projection games and a concentration bound. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, 2008.

[Raz98] Ran Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, June 1998. Prelim version in STOC '95.

[Raz11] Ran Raz. A counterexample to strong parallel repetition. *SIAM J. Comput.*, 40(3):771–777, 2011.

[Sha03] Ronen Shaltiel. Towards proving strong direct product theorems. *Computational Complexity*, 12(1-2):1–22, 2003. Prelim version CCC 2001.

[She11] Alexander A. Sherstov. Strong direct product theorems for quantum communication and query complexity. In *STOC*, pages 41–50, 2011.

A Remaining Proof of Theorem 3

We proceed to show that π aborts with with small probability ($\leq 15\epsilon$), and that is has low information cost ($O(\rho + \log \|\theta\|)$). We begin with the following lemma, which asserts the intuition that, since θ is ϵ-close a low-information distribution q, most paths in θ cannot reveal too much information. We defer its proof to Section D:

Lemma 41. $\theta\left(\mathbb{D}^{\theta}_{xyr}(m) > \rho\right) < 14\epsilon$.

For fixed x, y, r, m, define C to be the index at which the protocol aborted (if exists). That is, C is the smallest index $j \in [\|\theta\|]$ s.t

$$Count(x, y, r, m_{\leq j}) > 4\alpha\rho.$$

If no such j exists in m, define $C := \|\theta\| + 1$. Note that C is a random variable even after having fixed x, y, r, m.

We first argue that π is a good simulation of θ:

Claim 42. *π 15ϵ-simulates θ.*

Proof. It suffices to show that the probability that π aborts is 15ϵ.

$$\begin{aligned}&\Pr[\pi \text{ aborts}] \leq \Pr[Count(x, y, r, m) > 4\alpha\rho] \\ &\leq \theta(\mathbb{D}^{\theta}_{xyr}(m) > \rho) + \Pr[Count(x, y, r, m) > 4\alpha\rho \mid \mathbb{D}^{\theta}_{xyr}(m) \leq \rho]\end{aligned}$$

The first term is $\leq 14\epsilon$ by Lemma 41. The second term can be bounded as follows. Let $\epsilon' := \|\theta\|/2^{\ell}$ ($\epsilon' = 1/\|\theta\| < 1$ by choice of ℓ in π). Then whenever $\mathbb{D}^{\theta}_{xyr}(m) \leq \rho$,

$$4\alpha\rho \geq 3\alpha(\rho + \epsilon') \geq 3\alpha(\mathbb{D}^{\theta}_{xyr}(m) + \epsilon') \geq \mathbb{E}[Count(x, y, r, m)],$$

where the last inequality is by Claim 26. Since conditioned on m, $Count(x, y, r, m)$ is the sum of independent random variables, applying Corollary 16 with $\beta := \alpha(\rho + \epsilon'), \delta := 2$, we get

$$\begin{aligned}&\Pr\left[Count(x, y, r, m) > 4\alpha\rho \mid \mathbb{D}^{\theta}_{xyr}(m) \leq \rho\right] \\ &\leq \Pr\left[Count(x, y, r, m) > 3\alpha(\rho + \epsilon')\right] = e^{-\frac{2\alpha(\rho+\epsilon')}{2}} \leq e^{-\alpha\rho} = \epsilon\end{aligned}$$

by choice of $\alpha := \ln(1/\epsilon)/\rho$. In conclusion, $|\pi - \theta| \leq 14\epsilon + \epsilon = 15\epsilon$, as desired. □

We now turn to argue about the information cost of π. To this end, let us denote by $M_{<C}$ the transcript corresponding to all messages of θ sent in π, and by $T_{\leq C}$ the transcript corresponding to the "odometer" messages sent in step 2 of π (i.e, the concatenated transcripts of τ_{ℓ} in all rounds where it was executed). In this notation, we have that for all $j \in [\|\pi\|]$,

$$\Pi_{\leq j} = M_{\leq j} T_{\leq j}.$$

In particular, $\Pi = M_{<C}T_{\leq C}$. Define the divergence cost of each transcript as follows:

$$\mathbb{D}^{\theta}_{XYR}(M_{\leq j}) := \sum_{t=1}^{j}\left[\mathsf{D}\left(\frac{M_t|XYRM_{<t}}{\!\!\!\!=\!\!\!\!}\right) + \mathsf{D}\left(\frac{M_t|XYRM_{<t}}{\!\!\!\!=\!\!\!\!}\right)\right],$$
$$\mathbb{D}^{\theta}_{XYRM}(T_{\leq j}) := \sum_{t=1}^{j}\left[\mathsf{D}\left(\frac{T_t|XYR\Pi_{<t}}{\!\!\!\!=\!\!\!\!}\right) + \mathsf{D}\left(\frac{T_t|XYR\Pi_{<t}}{\!\!\!\!=\!\!\!\!}\right)\right]. \tag{33}$$

The proof of the following claim is deferred to Section E.

Claim 43. $\mathsf{IC}(\pi) \leq \mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] + \mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})] + 4\log(\|\theta\|+1)$.

To complete the proof, it remains to bound $\mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})]$ and $\mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})]$. We shall show that the first term is $O(\rho)$, and the second one is $O(\log\|\theta\|)$.

The intuition for the proof is as follows. Lemma 26 shows that the variable $Count(x,y,r,m_{\leq j})$ (normalized by α) is an unbiased estimator for the divergence cost $\mathbb{D}^{\theta}_{xyr}(m_{\leq j})$ of the path $m_{\leq j}$. Therefore by the end of the protocol, $\mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})]$ is essentially upper bounded by the (normalized) value of $Count$, which by construction is $O(\rho)$. As for $\mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})]$, the second proposition of Corollary 20 ensures that the information cost of τ_ℓ at round t is is comparable to that leaked by m_t in round t of θ (the latter is $\Theta((p_t-q_t)^2)$, and the former is $O(H((p_t-q_t)^2))$). Since τ_ℓ is only executed w.p $\alpha \approx 1/\rho$ at each round, this fact (roughly) implies $\mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})] \leq \alpha \cdot \mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] \cdot \log(\|\theta\|) = O(\alpha\rho \cdot \log(\|\theta\|)) = O(\log(\|\theta\|))$.

It turns out that the right way to formalize this intuition is using (very basic) martingale theory, by viewing the information cost of π as a stochastic process, with the "aborting" index C serving as a stopping rule for that process. The primary reason for doing so is the optional stopping theorem [Doo75], which allows one to relate the expected value of a (sub)martingale at stopping time to its initial value. We begin with the following general lemma, which is used several times in this paper:

Lemma 44. *Let $\{X_t\}_{t=0}^{K}$, $\{Y_t\}_{t=0}^{K}$, $\{Z_t\}_{t=0}^{K}$ be three stochastic processes in the same probability space, $X_t, Y_t, Z_t \geq 0$. For every $j \in [K]$, let*

$$L_j := \sum_{t=1}^{j} X_t - \lambda \cdot \sum_{t=1}^{j} Z_t - \frac{\sum_{t=1}^{j} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)}.$$

If for every $t \in [K]$ it holds that

(i) $\mathbb{E}[Y_t|L_{t-1}] \leq \beta \cdot H(\mathbb{E}[X_j|L_{t-1}])$

(ii) $\lambda \cdot \mathbb{E}[Z_t|L_{t-1}] \leq \frac{1}{2} \cdot \mathbb{E}[X_j|L_{t-1}]$

then $\mathcal{L} := \{L_j\}_{j=0}^{K}$ is a submatringale (i.e, $\mathbb{E}[L_j \mid L_{j-1}] \geq L_{j-1}\ \forall\ j$).

The proof is short but technical, so we defer it to Section F. We are now ready to prove our claim:

Claim 45. *There is a global constant $\lambda > 0$ such that*

$$\mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] + \mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})] \leq 30\rho + \lambda\log(1/\epsilon)\log(\|\theta\|+1).$$

Proof. Recall that in each round t the protocol τ_ℓ is executed with probability α, and that T_t denotes the transcript of τ_ℓ in round t ($T_t = \emptyset$ if τ_ℓ was not executed in this round). Then $T_{\leq C} = T_1 T_2 \dots T_C$. Furthermore, note that for any round t, T_t is independent from $T_{<t}$ given X, Y, R, M (since τ_ℓ is executed with independent randomness in each round, and we may assume, for analysis purpose, that $M \sim \theta$ was sampled in the beginning of π since the order doesn't affect the distribution of T_t). This means that $\pi(T_t|xyr\pi_{<t}) = \pi(T_t|xyrm_{<t})$. Now, for any $j \in [\|\theta\|]$, define

$$L_j := \sum_{t=0}^{j} \mathbf{1}^t_{\tau_\ell} - \frac{\alpha}{6} \cdot \mathbb{D}^\theta_{XYR}(M_{\leq j}) + j \cdot \alpha \cdot 2^{-\ell} - \frac{\mathbb{D}^\theta_{XYRM}(T_{\leq j}) - 2c_\tau \alpha j \cdot H(2^{-\ell}) - j/(\|\theta\|+1)}{8 \cdot 2c_\tau \cdot \log(2 \cdot 2c_\tau(\|\theta\|+1))}, \tag{34}$$

where c_τ is the constant from Corollary 20. Note that $\mathcal{L} := \{L_j\}_{j=0}^{(\|\theta\|+1)}$ is a stochastic process. Our main effort will be to show that $\mathcal{L}$ is in fact a *sub-martingale*:

Claim 46 ($\mathcal{L}$ is a sub-martingale)**.** $\mathbb{E}[L_j \mid L_{j-1}] \geq L_{j-1}$.

Let us first see why this finishes the proof of Claim 45. The index C at which the protocol π aborts is a well defined stopping rule for $\mathcal{L}$ ($C = \min\{j > 0 \ : \ \sum_{t=1}^{j} \mathbf{1}^t_{\tau_\ell} = \lceil 4\alpha\rho \rceil\}$), hence the optional stopping theorem [Doo75] together with Claim 46 implies that $E[L_C] \geq \mathbb{E}[L_0] = 0$, i.e,

$$\begin{aligned} \lceil 4\alpha\rho \rceil &\geq \mathbb{E}\left[\sum_{t=1}^{C} \mathbf{1}^t_{\tau_\ell}\right] \geq \\ &\geq \frac{\alpha}{6} \cdot \mathbb{E}[\mathbb{D}^\theta_{XYR}(M_{<C})] - (\|\theta\|+1) \cdot \alpha \cdot 2^{-\ell} + \frac{\mathbb{E}[\mathbb{D}^\theta_{XYRM}(T_{\leq C})] - 2c_\tau(\|\theta\|+1) \cdot H(2^{-\ell}) - 1}{8 \cdot 2c_\tau \log(2 \cdot 2c_\tau(\|\theta\|+1))} \end{aligned} \tag{35}$$

since by definition the value of $Count$ is at most $\lceil 4\alpha\rho \rceil$ at the aborting index C, and $j \leq (\|\theta\|+1)$. Note that

$$2c_\tau \alpha C \cdot H(2^{-\ell}) \leq 2c_\tau(\|\theta\|+1) \cdot H(2^{-\ell}) \leq 2c_\tau(\|\theta\|+1) \cdot 2\sqrt{2^{-\ell}} \leq 2c_\tau(\|\theta\|+1) \cdot 2/(\|\theta\|+1) = 4c_\tau \tag{36}$$

by Proposition 13/(i) and choice of $\ell = 2\log(\|\theta\|+1)$. Thus by (35) and the fact that $2^{-\ell} < 1/(\|\theta\|+1)$, we have

$$\begin{aligned} \lceil 4\alpha\rho \rceil &\geq \frac{\alpha}{6} \cdot \mathbb{E}[\mathbb{D}^\theta_{XYR}(M_{<C})] - \alpha + \frac{\mathbb{E}[\mathbb{D}^\theta_{XYRM}(T_{\leq C})] - 4c_\tau - 1}{8 \cdot 2c_\tau \log(2 \cdot 2c_\tau(\|\theta\|+1))} \\ &\geq \frac{\alpha}{6} \cdot \mathbb{E}[\mathbb{D}^\theta_{XYR}(M_{<C})] - \alpha + \frac{\mathbb{E}[\mathbb{D}^\theta_{XYRM}(T_{\leq C})]}{8 \cdot 2c_\tau \log(2 \cdot 2c_\tau(\|\theta\|+1))} - 1 \end{aligned} \tag{37}$$

so rearranging implies

$$\frac{\alpha}{6} \cdot \mathbb{E}[\mathbb{D}^\theta_{XYR}(M_{<C})] + \frac{\mathbb{E}[\mathbb{D}^\theta_{XYRM}(T_{\leq C})]}{8 \cdot 2c_\tau \log(2 \cdot 2c_\tau(\|\theta\|+1))} \leq \lceil 4\alpha\rho \rceil + \alpha + 1 \leq 5\alpha\rho.$$

Since both terms in the LHS are non-negative, (37) implies in particular that each of them is at most $5\alpha\rho$, therefore:

- $\frac{\alpha}{6} \cdot \mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] \leq 5\alpha\rho \implies \mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] \leq 5\alpha\rho \cdot \frac{6}{\alpha} = 30\rho.$
- $\mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})] \leq (8 \cdot 2c_{\tau} \log(2 \cdot 2c_{\tau}(\|\theta\| + 1))) \cdot 5\alpha\rho \leq \lambda\alpha\rho \cdot \log((\|\theta\| + 1)),$

for a sufficiently large constant $\lambda > 0$. Finally, recalling that $\alpha = \ln(1/\epsilon)/\rho$, we conclude that

$$\mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] + \mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\leq C})] \leq 30\rho + \lambda \log(1/\epsilon) \log(\|\theta\| + 1) \tag{38}$$

which finishes the proof of the claim.

It therefore remains to prove that $\mathcal{L}$ is indeed a sub-martingale. This argument is the heart of the proof, as it is the only place where we use the "low information" guarantee of our odometer (second proposition of Corollary 20).

Proof of Claim 46. Suppose j is odd (an analogues argument follows for even j). Set:

- $K = \|\theta\| + 1.$
- $X_t = \mathbf{1}^t_{\tau_\ell}.$
- $Y_t = \mathsf{D}\left(\frac{T_t|XYR\Pi_{<t}}{T_t|YR\Pi_{<t}}\right) + \mathsf{D}\left(\frac{T_t|XYR\Pi_{<t}}{T_t|XR\Pi_{<t}}\right) - c_{\tau}\alpha \cdot H(2^{-\ell}).$
- $Z_t = \mathsf{D}\left(\frac{M_t|XYRM_{<t}}{M_t|YRM_{<t}}\right) - 6 \cdot 2^{-\ell}.$
- $\lambda = \alpha/6.$
- $\beta = 2c_{\tau}.$

By the definitions in (33) and since τ_ℓ is executed with probability α at each round, in this notation (34) can be rewritten as

$$L_j = \sum_{t=1}^{j} X_t - \lambda \cdot \sum_{t=1}^{j} Z_t - \frac{\sum_{t=1}^{j} Y_t - \frac{j}{K}}{8\beta \cdot \log(2\beta K)}. \tag{39}$$

We are thus in the setup of Lemma 44. To apply the Lemma, we must show that:

1. $\mathbb{E}[Y_t|L_{t-1}] \leq \beta \cdot H(\mathbb{E}[X_j|L_{t-1}]).$
2. $\lambda \cdot \mathbb{E}[Z_t|L_{t-1}] \leq \frac{1}{2} \cdot \mathbb{E}[X_j|L_{t-1}].$

First proposition :

$$\mathbb{E}[Y_t \mid L_{t-1}] = \alpha \cdot \mathbb{E}\left[\mathsf{D}\left(\frac{T_t|xyr\pi_{<t}}{T_t|yr\pi_{<t}}\right) + \mathsf{D}\left(\frac{T_t|xyr\pi_{<t}}{T_t|xr\pi_{<t}}\right)\right] - c_\tau\alpha \cdot H(2^{-\ell})$$

$$= \alpha \cdot \mathbb{E}\left[\mathsf{D}\left(\frac{T_t|xyrm_{<t}}{T_t|yr\pi_{<t}}\right) + \mathsf{D}\left(\frac{T_t|xyrm_{<t}}{T_t|xr\pi_{<t}}\right)\right] - c_\tau\alpha \cdot H(2^{-\ell})$$

(Since T_t is independent of $T_{<t}$ given $XYRM_{<t}$, and therefore $\pi(T_t|xyr\pi_{<t}) = \pi(T_t|xyrm_{<t})$)

$$= \alpha \cdot \mathbb{E}\left[\mathsf{D}\left(\frac{T_t|p_tq_t}{T_t|yr\pi_{<t}}\right) + \mathsf{D}\left(\frac{T_t|p_tq_t}{T_t|xr\pi_{<t}}\right)\right] - c_\tau\alpha \cdot H(2^{-\ell})$$

(Since $\pi(T_t|xyrm_{<t}) = \pi(T_t|p_tq_t)$ by definition of p_t, q_t)

$$\leq \alpha \cdot \mathbb{E}\left[\mathsf{D}\left(\frac{T_t|p_tq_t}{T_t|q_t}\right) + \mathsf{D}\left(\frac{T_t|p_tq_t}{T_t|p_t}\right)\right] - c_\tau\alpha \cdot H(2^{-\ell})$$

(By Fact 12, setting $T = T_t, X = p_tq_t, Y = yr\pi_{<t}, Z(yr\pi_{<t}) = T_t|p_tq_t$)

$$\leq \alpha c_\tau \cdot \mathbb{E}[H((p_t - q_t)^2)] - c_\tau\alpha \cdot H(2^{-\ell}) \quad \text{(second proposition of Corollary 20)}$$

$$\leq c_\tau\alpha \cdot H(\mathbb{E}(p_t - q_t)^2) - c_\tau\alpha \cdot H(2^{-\ell}) \quad \text{(by concavity of } H(x))$$

$$= c_\tau\alpha \cdot (H(\mathbb{E}[(p_t - q_t)^2]) - H(2^{-\ell}))$$

$$\leq c_\tau\alpha \cdot H(\mathbb{E}[\mathbf{1}^t_{\tau_\ell}]/2) \tag{40}$$

$$\leq 2c_\tau\alpha \cdot H(\mathbb{E}[\mathbf{1}^t_{\tau_\ell}]) \tag{41}$$

$$\leq 2c_\tau \cdot H(\alpha \cdot \mathbb{E}[\mathbf{1}^t_{\tau_\ell}]) \quad \text{(by Proposition 13/}(ii)\text{ taken with } y = \alpha)$$

$$= 2c_\tau \cdot H(\mathbb{E}[X_t|L_{t-1}]) \tag{42}$$

where transition (40) follows from Proposition 13/(iv) taken with $x = (p_t - q_t)^2$,$y = \mathbb{E}[\mathbf{1}^t_{\tau_\ell}]/2$ and $\epsilon = 2^{-\ell}$, and Corollary 20) which implies $|\mathbb{E}[\mathbf{1}^t_{\tau_\ell}]/2 - (p_t - q_t)^2| \leq 2^{-\ell}$. Note that in the last inequality we used the fact that $\mathbf{1}^t_{\tau_\ell}$ is independent from L_{t-1} given $m_{<t}, x, y, r$.

Second proposition : Recall that by the first proposition of Corollary 20, $\mathbb{E}[\mathbf{1}^t_{\tau_\ell}] \in 2(p_t - q_t)^2 \pm 2^{-\ell}$. We have

$$\lambda \cdot \mathbb{E}[Z_t \mid L_{t-1}] = \frac{\alpha}{6}\left(\mathsf{D}\left(\frac{M_t|xyrm_{<t}}{M_t|yrm_{<t}}\right) - 6 \cdot 2^{-\ell}\right) = \frac{\alpha}{6}\left(\mathsf{D}\left(\frac{M_t|p_t}{M_t|yrm_{<t}}\right) - 6 \cdot 2^{-\ell}\right)$$

$$\leq \frac{\alpha}{6}\left(\mathsf{D}\left(\frac{M_t|p_t}{M_t|q_t}\right) - 6 \cdot 2^{-\ell}\right) \quad \text{(By Fact 12, setting } T = M_t, X = p_t, Y = yr\pi_{<t}, Z(yr\pi_{<t}) = M_t|q_t)$$

$$\frac{\alpha}{6}\left(\mathsf{D}\left(\frac{p_t}{q_t}\right) - 6 \cdot 2^{-\ell}\right) \quad \text{(By definition of } p_t, q_t)$$

$$\leq \frac{\alpha}{6}\left(\frac{9}{2} \cdot (p_t - q_t)^2 - 6 \cdot 2^{-\ell}\right) \quad \text{(By Proposition 14 and smoothness of } \theta)$$

$$= \frac{\alpha \cdot \left(\frac{3}{2}(p_t - q_t)^2 - 2^{-\ell}\right)}{2} \leq \frac{\alpha \cdot \mathbb{E}[\mathbf{1}^t_{\tau_\ell}]}{2} = \frac{1}{2} \cdot \mathbb{E}[X_t \mid L_{t-1}] \tag{43}$$

where we used again the fact that $\mathbf{1}^t_{\tau_\ell}$ is independent from the history given $m_{<}t$.

We may now apply Lemma 44 to conclude that $\mathcal{L}$ is a sub-martingale, which completes the proof. □

In conclusion, Claims 43 and 45 together imply that the information cost of π is at most

$$\mathsf{IC}(\pi) \leq 120\rho + \lambda \log(1/\epsilon)\log(\|\theta\|) + 4\log(\|\theta\|+1) \leq O\left(\frac{I_q + \log(\|\theta\|+1)}{\epsilon^2}\right)$$

since $\rho = \frac{2I_q + 4/(e\ln 2) + 3\log(\|\theta\|+1)}{\epsilon^2} + \frac{2\log(1/\epsilon)}{\epsilon}$. This concludes the whole proof of Theorem 3.

□

B Proof of Lemma 19

Recall that our goal is to show that for any $p, p' \in (1/3, 2/3)$, $\mathsf{D}\left(\frac{Z_p}{Z_{p'}}\right) \leq 8(p-p')^2$.

Proof. Assume w.l.o.g that $p < p' < 1/2$ (recall that for $p > 1/2$, $\mu_{1-p}(1-z) = \mu_p(z)$, so the following argument applies to all ranges of p, p'). Recall that $Supp(Z) = [0,1]$. We divide $[0,1]$ into regions, and bound the contribution of each separate region to the divergence. In what follows, we use natural base logs (this will only affect the divergence calculation by a multiplicative constant $\log_2(e)$). Throughout the proof let us denote $\alpha := p' - p$.

Region 1: $z \in [0,p] \cup [p'+1/2, 1]$. Note that by definition, for $z \in [0,p]$, $\mu_{p'}(z) - \mu_p(z) = 4(p'-z) - 4(p-z) = 4(p'-p)$, and also for $z \in [p'+1/2, 1]$, $\mu_{p'}(z) - \mu_p(z) = 2 - 4(z-p'-1/2) - [2-4(z-p-1/2)] = 4(p'-p)$. So we have that in the region above,

$$\mu_{p'}(z) - \mu_p(z) = 4\alpha.$$

Therefore, the contribution of this region to the KL divergence $\mathsf{D}\left(\frac{Z_p}{Z_{p'}}\right)$ is:

$$\begin{aligned} C(1) &:= \int_0^p \mu_p(z)\log\left(\frac{\mu_p(z)}{\mu_p(z)+\alpha}\right)dz + \int_{p'+1/2}^1 \mu_p(z)\log\left(\frac{\mu_p(z)}{\mu_p(z)+\alpha}\right)dz \\ &= \int_0^p 4(p-z)\log\left(\frac{p-z}{p-z+\alpha}\right)dz + \int_{p'+1/2}^1 [2-4(z-p-1/2)]\log\left(\frac{2-4(z-p-1/2)}{2-4(z-p-1/2)+\alpha}\right)dz. \end{aligned}$$

Using Wolfram Mathematica, the above integral is equal to

$$C(1) = 2\left(-\alpha(\frac{1}{2}-\alpha) + (\frac{1}{2}-\alpha)^2 \cdot \log(1-2\alpha) + \alpha^2 \cdot \log(\frac{1}{2\alpha})\right).$$

Region 2: $z \in (p', p+1/2]$. Note that in this region, $\mu_{p'}(z) - \mu_p(z) = 4(z-p') - 4(z-p) = -4\alpha$. Similarly to the calculation in Region 1, the contribution in this region is

$$C(2) = \int_{p'}^{p+1/2} 4(z-p)\log\left(\frac{4(z-p)}{4(z-p)-4\alpha}\right)dz.$$

Using Wolfram Mathematica, the above integral is equal to

$$C(2) = 2 \cdot \left(\alpha(\frac{1}{2} - \alpha) - (\frac{1}{2} - \alpha)(\frac{1}{2} + \alpha) \cdot \log(1 - 2\alpha) + \alpha^2 \cdot \log(\frac{1}{2\alpha}) \right).$$

Region 3: $z \in (p, p']$. By definition of μ_p, in this region: $\mu_p(z) = 4(z - p)$, $\mu_{p'}(z) = 4(p' - z)$. Therefore,

$$C(3) := \int_p^{p'} \mu_p(z) \log\left(\frac{\mu_p(z)}{\mu_{p'}(z)}\right) dz = \int_p^{p'} 4(z-p) \log\left(\frac{z-p}{p'-z}\right) dz \leq 4 \int_p^{p'} (z-p) \log\left(\frac{p'-p}{p'-z}\right) dz$$

where the inequality follows from the fact that $\mu_p(z) = z - p < p' - p$ in this region. For convenience, set $y = z - p$. Recalling that $\alpha = p' - p$, then the above becomes

$$\begin{aligned} &= 4 \int_0^{\alpha} y \log\left(\frac{\alpha}{\alpha - y}\right) dy \leq 4\alpha \int_0^{\alpha} \log\left(\frac{\alpha}{\alpha - y}\right) = 4\alpha \cdot [(y - \alpha)\log(\alpha/(\alpha - y) + y))]_0^{\alpha} = \\ &= 4\alpha \cdot [\text{“}0 \log(0)\text{”} + \alpha + \alpha \log(\alpha/\alpha)] = 4\alpha^2 = 4(p' - p)^2. \end{aligned}$$

where we used the fact that $\lim_{x \to 0} x \log(x) = 0$.

Region 4: $z \in (p + 1/2, p' + 1/2]$. In this region, $\mu_p(z) = 4(1 + p - z) \leq 2$ (for $z = p + 1/2$), and $\mu_{p'}(z) = 4(z - p') \geq 4(p + 1/2 - p') = 2 - 4(p' - p)$ (for $z = p + 1/2$). Therefore

$$\begin{aligned} &C(4) \leq \int_{p+1/2}^{p'+1/2} 4(1 + p - z) \log\left(\frac{2}{2 - 4(p' - p)}\right) dz = -4 \log[1 - 2(p' - p)] \int_{p+1/2}^{p'+1/2} 4(1 + p - z) dz = \\ &= -4 \log[1 - 2(p' - p)] \cdot [(1 + p)z - z^2/2]_{p+1/2}^{p'+1/2} = \\ &= -4 \log[1 - 2(p' - p)] \cdot [(p' - p)(1 + p) - \frac{1}{2} \cdot (p' - p)(p' + p + 1/2)] = \\ &= -4 \log[1 - 2(p' - p)] \cdot [(p' - p)(1 + p - p'/2 - p/2 - 1/2] \leq \\ &\leq -4 \log[1 - 2(p' - p)] \cdot \frac{p' - p}{2} \leq 4 \cdot 2(p' - p) \cdot \frac{p' - p}{2} = 4(p' - p)^2. \end{aligned}$$

where in the first inequality in the last line we used the fact that $p - p'/2 - p/2 < 0$ since $p < p'$ by assumption.

Concluding: It can be directly verified that

$$C(1) + C(2) = 2\left(-\alpha \cdot \log(1 - 2\alpha) + 2 \cdot \alpha^2 \cdot \log(1 - 2\alpha) + 2 \cdot \alpha^2 \cdot \log(1/2\alpha)\right) \leq c_1 \cdot \alpha^2 \log(1/\alpha),$$

for a sufficiently large constant $0 < c_1 < 40$. Accounting for the base change of the logs, we therefore conclude that

$$\begin{aligned} &\mathsf{D}\left(\frac{Z_p}{Z_{p'}}\right) = \log_2(e) \cdot [C(1) + C(2) + C(3) + C(4)] \leq \\ &\leq \log_2(e)[8\alpha^2 + c_1 \cdot \alpha^2 \log(1/\alpha)] \leq 2(c_1 + 8) \cdot (p' - p)^2 \cdot \log(1/p' - p) \leq c \cdot H((p' - p)^2), \end{aligned}$$

for a sufficiently large constant $c = 4(c_1 + 8)$. □

C Proof of Proposition (4)

Proof. By linearity of expectation and definition of the divergence cost, we have

$$\begin{aligned}
\mathbb{E}_{xyrm\sim\theta}\left[\mathbb{D}^{\theta}_{xyr}(m)\right] &= \sum_{t=1}^{\|\theta\|}\mathbb{E}_{xyrm\sim\theta}\left[\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}xr)}\right)+\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}yr)}\right)\right] = \\
&= \sum_{t=1}^{\|\theta\|}\mathbb{E}_{xyrm_{<t}\sim\theta}\left[\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}xr)}\right)+\mathsf{D}\left(\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}yr)}\right)\right] = \\
&= \sum_{t=1}^{\|\theta\|} I_\theta(M_t;Y|XRM_{<t}) + I_\theta(M_t;X|YRM_{<t}) = \\
&= I_\theta(M;Y|XR) + I_\theta(M;X|YR) = \mathsf{IC}(\theta),
\end{aligned}$$

where in the second equality we used the fact that the t'th term is independent of $m_{>t}$, the third equality follows from the definition of (conditional) mutual information, and the fourth equality follows from the chain rule (Fact 11). □

D Proof of Lemma 41

We will need the following lemma from [BRWY12], which shows that if a variable A is statistically close to having low information, then some prefix $A_{\le K}$ of A usually has low information: to $A_{\le K}$, one can obtain a new variable that is statistically close to the old one, yet actually has low information.

Lemma 47 (Truncation Lemma [BRWY12])**.** *Let $p(a,b,c) \overset{\epsilon}{\approx} q(a,b,c)$ where $a = a_1,\dots,a_s$. For every a,b,c, define k to be the minimum number j in $[s]$ such that*

$$\log\frac{p(a_{\le j}|bc)}{p(a_{\le j}|c)} > \beta.$$

If no such index exists, set $k = s+1$. Then,

$$p(k < s+1) < \frac{I_q(A;B|C) + \log(s+1) + 1/(e\ln 2)}{\beta - 2} + 9\epsilon/2.$$

Proof of Lemma 41. Let

$$\beta := \frac{I_q + 1/(e\ln 2) + \log(\|\theta\|+1)}{\epsilon} + \log(1/\epsilon) + 2.$$

For any x,y,r,m, let k denote the smallest index j such that either

$$\sum_{t\le j,t \text{ odd}} \log\frac{\theta(m_t|xrm_{<t})}{\theta(m_t|yrm_{<t})} > \beta \quad \text{or} \quad \sum_{t\le j,t \text{ even}} \log\frac{\theta(m_t|yrm_{<t})}{\theta(m_t|xrm_{<t})} > \beta. \tag{44}$$

If no such index, define $k = \|\theta\|+1$. Note that the random variable[10] K is a function of X,Y,R,M.

We prove the following three Claims in order:

[10]Since it can be ambiguous whether the expression $p(m_k)$ refers to $p(M_K = m_k)$ or $p(M_k = m_k)$, we shall be more explicit with the notation in the rest of this section. However, observe that $p(m_k,k)$ has only one interpretation, so in such cases we use the more concise notation.

Claim 48. $\mathbb{E}_\theta[\mathbb{D}^\theta_{xyr}(m_{<k})] \le 2(\beta + \log(\|\theta\|+1))$.

Claim 49. *The probability of "truncation" is small:* $\theta(k \le \|\theta\|) < 13\epsilon$.

Claim 50. *For any* $\rho > 0$, $\theta(\mathbb{D}^\theta_{xyr}(m) > \rho) \le \theta(\mathbb{D}^\theta_{xyr}(m_{<k}) > \rho) + (k \le \|\theta\|)$.

To prove the lemma using the above claims, set $\rho := [2\beta + 2\log(\|\theta\|+1)]/\epsilon$. Since divergence is non-negative (Lemma 10), $\mathbb{D}^\theta_{xyr}(m_{<k}) \ge 0$. Thus it follows from Markov's inequality that $\theta(\mathbb{D}^\theta_{xyr}(m_{<k}) > \rho) \le \mathbb{E}_\theta[\mathbb{D}^\theta_{xyr}(m_{<k})]/\rho$. By Claims 48,49 and 50, we conclude that

$$\theta\left(\mathbb{D}^\theta_{xyr}(m) > \frac{2I_q + 4/(e\ln 2) + 3\log(\|\theta\|+1)}{\epsilon^2} + \frac{2\log(1/\epsilon)}{\epsilon}\right) \le \theta\left(\mathbb{D}^\theta_{xyr}(m) > \frac{2\beta + 2\log(\|\theta\|+1)}{\epsilon}\right)$$
$$= \theta(\mathbb{D}^\theta_{xyr}(m) > \rho) \le \theta(\mathbb{D}^\theta_{xyr}(m_{<k}) > \rho) + (k \le \|\theta\|) \le \frac{2(\beta + \log(\|\theta\|+1))}{2(\beta+\log(\|\theta\|+1))/\epsilon} + 13\epsilon = 14\epsilon.$$

We turn to prove Claims 48,49 and 50.

Proof of Claim 48. By definition, $\mathbb{E}_\theta[\mathbb{D}^\theta_{xyr}(m_{<k})] = I_\theta(X; M_{<K}|YR) + I_\theta(Y; M_{<K}|XR)$.

$$I_\theta(X; M_{<K}|YR) \le I_\theta(X; KM_{<K}|YR) = \sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \log \frac{\theta(km_{<k}|xyr)}{\theta(km_{<k}|yr)}$$
$$= \sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \left(\log \frac{\theta(M_{<k}=m_{<k}|xyr)}{\theta(M_{<k}=m_{<k}|yr)} + \log \frac{\theta(K=k|M_{<k}=m_{<k},xyr)}{\theta(K=k|M_{<k}=m_{<k},yr)}\right). \quad (45)$$

The second term can be bounded as follows:

$$\sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \log \frac{\theta(K=k|M_{<k}=m_{<k},xyr)}{\theta(K=k|M_{<k}=m_{<k},yr)}$$
$$\le \sum_{y,r,k,m_{<k}} \theta(yrkm_{<k}) \log \frac{1}{\theta(K=k|M_{<k}=m_{<k},yr)}$$
$$\le \log \sum_{y,r,k,m_{<k}} \frac{\theta(yrkm_{<k})}{\theta(K=k|M_{<k}=m_{<k},yr)} \qquad \text{by concavity of log}$$
$$= \log \sum_{y,r,k,m_{<k}} \theta(M_{<k}=m_{<k},yr) = \log(\|\theta\|+1). \quad (46)$$

As for the first term,

$$\log\left(\frac{\theta(M_{<k}=m_{<k}|xyr)}{\theta(M_{<k}=m_{<k}|yr)}\right) = \sum_{j<k, j \text{ odd}} \log \frac{\theta(m_j|xrm_{<j})}{\theta(m_j|yrm_{<j})} + \sum_{j<k, j \text{ even}} \log \frac{\theta(m_j|yrm_{<j})}{\theta(m_j|yrm_{<j})},$$

where here we used the fact that since θ is a protocol, each (odd) message m_j sent by Alice satisfies $\theta(m_j|xyrm_{<j}) = \theta(m_j|xrm_{<j})$, and that a similar statement holds for Bob's messages. Thus by the definition of K,

$$\sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \log \frac{\theta(M_{<k}=m_{<k}|xyr)}{\theta(M_{<k}=m_{<k}|r)} \le \beta. \quad (47)$$

Combining (45), (46) and (47), with an analogues argument for $I_\theta(Y; M_{<K}|XR)$, we conclude that

$$\mathbb{E}_\theta[\mathbb{D}^\theta_{xyr}(m_{<k})] = I_\theta(X; M_{<K}|YR) + I_\theta(Y; M_{<K}|XR) \leq 2(\beta + \log(\|\theta\| + 1)).$$

□

Next, we prove Claim 49, showing that the probability that the index k "truncates" the protocol is small:

Proof of Claim 49. For any x, y, r, m, let k' denote the smallest index such that

$$\sum_{j \leq k', j \text{ odd}} \log \frac{\theta(m_j|xrm_{<j})}{\theta(m_j|yrm_{<j})} > \beta - \log(1/\epsilon) \quad \text{or} \quad \sum_{j \leq k', j \text{ even}} \log \frac{\theta(m_j|yrm_{<j})}{\theta(m_j|xrm_{<j})} > \beta - \log(1/\epsilon).$$

If no such index, define $k' = \|\theta\| + 1$. By Lemma 47, applied once with $a := m, b := x, c := yr$, and once with $a := m, b := y, c := xr$ (using the fact that $\theta(xyrm) \overset{\epsilon}{\approx} q(xyrm)$), we have

$$\theta(k' \leq \|\theta\|) < 2\left(\frac{I_q + 1/(e \ln 2) + \log(\|\theta\| + 1)}{\beta - 2 - \log(1/\epsilon)} + 9\epsilon/2\right) \leq 11\epsilon, \tag{48}$$

by choice of β. We shall show that $\theta(k < k') < 2\epsilon$, which will complete the proof. Define

$$S_1 = \left\{(x, y, r, m) : k(x, y, r, m) \leq \|\theta\| \text{ and } \sum_{d \leq k, d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|yrm_{<d})} \leq -\log(1/\epsilon)\right\},$$

$$S_2 = \left\{(x, y, r, m) : k(x, y, r, m) \leq \|\theta\| \text{ and } \sum_{d \leq k, d \text{ even}} \log \frac{\theta(m_d|yrm_{<d})}{\theta(m_d|xrm_{<d})} \leq -\log(1/\epsilon)\right\}.$$

Observe that $k < k'$ implies that $(x, y, r, m) \in S_1 \cup S_2$. We shall prove that $\theta(S_1) \leq \epsilon$ and $\theta(S_2) \leq \epsilon$. Consider the distribution

$$\theta'(xyrm) = \theta(xyr) \cdot \prod_{d \text{ odd}} \theta(m_d|yrm_{<d}) \cdot \prod_{d \text{ even}} \theta(m_d|yrm_{<d}).$$

Fix any $(x, y, r, m) \in S_1$, and let $k = k(x, y, r, m)$ be defined as above. We have:

$$\begin{aligned}
&\log \frac{\theta(km_{\leq k}|xyr)}{\theta'(km_{\leq k}|xyr)} \\
&= \sum_{d \leq k, d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|yrm_{<d})} + \sum_{d \leq k, d \text{ even}} \log \frac{\theta(m_d|yrm_{<d})}{\theta(m_d|yrm_{<d})} + \log \frac{\theta(K = k|M_{\leq k} = m_{\leq k}, xyr)}{\theta'(K = k|M_{\leq k} = m_{\leq k}, xyr)} \\
&= \sum_{d \leq k, d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|yrm_{<d})} \leq -\log(1/\epsilon).
\end{aligned}$$

Thus $\theta(xyrkm_{\leq k}) \leq \epsilon \cdot \theta'(xyrkm_{\leq k})$. So (here we set $k = k(x, y, r, m)$ in the sum):

$$\begin{aligned}
\theta(S_1) &= \sum_{(x,y,r,m) \in S_1} \theta(xyrm) \\
&= \sum_{(x,y,r,m) \in S_1} \theta(xyrkm_{\leq k}) \cdot \theta(m|xyrkm_{\leq k}) \\
&\leq \epsilon \sum_{(x,y,r,m) \in S_1} \theta'(xyrkm_{\leq k}) \cdot \theta(m|xyrkm_{\leq k}) \leq \epsilon.
\end{aligned}$$

A similar argument proves $\theta(S_2) \leq \epsilon$. Thus, by (48), we have that $\theta(k \leq \|\theta\|) \leq \theta(k' \leq \|\theta\|) + \theta(k < k') < 11\epsilon + 2\epsilon = 13\epsilon$ as required. □

Proof of Claim 50.

$$\begin{aligned}
&\theta(\mathbb{D}^{\theta}_{xyr}(m) > \rho) = \\
&= \theta(\mathbb{D}^{\theta}_{xyr}(m) > \rho | k = \|\theta\| + 1)\theta(k = \|\theta\| + 1) + \theta(\mathbb{D}^{\theta}_{xyr}(m) > \rho | k \leq \|\theta\|)\theta(k \leq \|\theta\|) \leq \\
&\leq \theta(\mathbb{D}^{\theta}_{xyr}(m_{<k}) > \rho) + \theta(k \leq \|\theta\|)
\end{aligned}$$

where the last transition is by Claim 49 and since $\mathbb{D}^{\theta}_{xyr}(m) = \mathbb{D}^{\theta}_{xyr}(m_{<k})$ whenever $k = \|\theta\| + 1$. □

□

E Proof of Claim 43

Proof. Recall that $\Pi = M_{<C}T_{\leq C}$. By the chain rule,

$$\begin{aligned}
&I(\Pi; X|YR) = I(M_{<C}T_{\leq C}; X|YR) = \\
&I(M_{<C}; X|YR) + I(T_{\leq C}; X|YRM_{<C}) \\
&= I(M_{<C}; X|YR) + I(T_{\leq C}; X|YRM_{<C}, C) \qquad \text{(Since } M_{<C} \text{ determines } C\text{)} \\
&\leq I(M_{<C}; X|YR) + I(T_{\leq C}; M_{\geq C}X|YRM_{<C}, C) \\
&= I(M_{<C}; X|YR) + I(T_{\leq C}; M_{\geq C}|YRM_{<C}, C) + I(T_{\leq C}; X|YRM, C) \\
&= I(M_{<C}; X|YR) + I(T_{\leq C}; X|YRM, C) \qquad (49) \\
&\leq I(M_{<C}; X|YR) + I(T_{\leq C}, C; X|YRM) \\
&\leq I(M_{<C}; X|YR) + I(T_{\leq C}; X|YRM) + I(C; X|YRM, T_{\leq C}) \\
&= I(M_{<C}; X|YR) + I(T_{\leq C}; X|YRM) \qquad (50)
\end{aligned}$$

where inequality (49) follows from Proposition 7, as $I(M_{\geq C}; T_{\leq C}|M_{<C}XYR, C) = 0$, and the last inequality follows since $I(C; X|YRM, T_{\leq C}) = 0$ as $H(C|T_{\leq C}) = 0$. Repeating the same argument for $I(\Pi; Y|XR)$, we have

$$\mathsf{IC}(\pi) \leq I(M_{<C}; X|YR) + I(M_{<C}; Y|XR) + I(T_{\leq C}; X|YRM) + I(T_{\leq C}; Y|XRM). \qquad (51)$$

We first show that $I(M_{<C}; X|YR) + I(M_{<C}; Y|XR) \leq \mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] + 2\log(\|\theta\| + 1)$. Note that since $M_{<C}$ determines C, we have

$$\begin{aligned}
&I(M_{<C}; X|YR) + I(M_{<C}; Y|XR) = I(CM_{<C}; X|YR) + I(CM_{<C}; Y|XR) = \\
&= \sum_{x,y,r,c,m_{<c}} \pi(xyrcm_{<c}) \left[\log \frac{\pi(cm_{<c}|xyr)}{\pi(cm_{<c}|yr)} + \log \frac{\pi(cm_{<c}|xyr)}{\pi(cm_{<c}|xr)} \right] \\
&= \sum_{x,y,r,c,m_{<c}} \pi(xyrcm_{<c}) \Bigg[\log \frac{\pi(M_{<c} = m_{<c}|xyr)}{\pi(M_{<c} = m_{<c}|yr)} + \log \frac{\pi(M_{<c} = m_{<c}|xyr)}{\pi(M_{<c} = m_{<c}|xr)} + \\
&+ \log \frac{\pi(C = c|M_{<c} = m_{<c}, xyr)}{\pi(C = c|M_{<c} = m_{<c}, yr)} + \log \frac{\pi(C = c|M_{<c} = m_{<c}, xyr)}{\pi(C = c|M_{<c} = m_{<c}, xr)} \Bigg]. \qquad (52)
\end{aligned}$$

Each of the the last two terms contributes at most $\log(\|\theta\|+1)$ (See the calculation in (46)). We shall show that the contribution of the first two terms is exactly $\mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})]$, that is

$$\mathbb{E}_{xyrcm}\left[\log\frac{\pi(M_{<c}=m_{<c}|xyr)}{\pi(M_{<c}=m_{<c}|yr)}+\log\frac{\pi(M_{<c}=m_{<c}|xyr)}{\pi(M_{<c}=m_{<c}|xr)}\right]=\mathbb{E}_{xyrcm}[\mathbb{D}^{\theta}_{XYR}(M_{<C})]. \tag{53}$$

To this end, we define for any $j\in[\|\theta\|]$, the random variable

$$L_j:=\mathbb{D}^{\theta}_{XYR}(M_{\le j})-\left[\log\frac{\theta(M_{\le j}|XYR)}{\theta(M_{\le j}|YR)}+\log\frac{\theta(M_{\le j}|XYR)}{\theta(M_{\le j}|XR)}\right].$$

Note that $\mathcal{L}:=\{L_j\}_{j=0}^{\|\theta\|+1}$ is a stochastic process. In fact:

Claim 51. *$\mathcal{L}$ is a martingale: $\mathbb{E}[L_j|L_{j-1}]=L_{j-1}$.*

Proof. We show that $\mathbb{E}[L_j\mid L_{j-1}]-L_{j-1}=0$. Indeed, suppose w.l.o.g that j is odd, and fix the choice of random variables up to L_{j-1} (i.e, fix $x,y,r,m_{<j}$). Since $\log\frac{\theta(m_{\le j}|xyr)}{\theta(m_{\le j}|yr)}=\sum_{t=1}^{j}\log\frac{\theta(m_t|m_{<t}xyr)}{\theta(m_t|m_{<t}yr)}$, and for odd j $\theta(m_t|m_{<t}xyr)=\theta(m_t|m_{<t}xr)$, we have

$$\mathbb{E}[L_j\mid L_{j-1}]=\mathbb{E}_{m_j}\left[\mathsf{D}\left(\frac{\theta(m_j|m_{<j}xyr)}{\theta(m_j|m_{<j}yr)}\right)\right]-\mathbb{E}_{m_j}\left[\log\frac{\theta(m_j|m_{<j}xyr)}{\theta(m_j|m_{<j}yr)}\right]+L_{j-1}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[L_j\mid L_{j-1}]-L_{j-1}&=\mathbb{E}_{m_j}\left[\mathsf{D}\left(\frac{\theta(m_j|m_{<j}xyr)}{\theta(m_j|m_{<j}yr)}\right)\right]-\mathbb{E}_{m_j}\left[\log\frac{\theta(m_j|m_{<j}xyr)}{\theta(m_j|m_{<j}yr)}\right]\\&=\mathbb{E}_{m_j}\left[\mathsf{D}\left(\frac{\theta(m_j|m_{<j}xyr)}{\theta(m_j|m_{<j}yr)}\right)\right]-\mathsf{D}\left(\frac{\theta(m_j|m_{<j}xyr)}{\theta(m_j|m_{<j}yr)}\right)=0,\end{aligned}$$

where the second equality is by definition of KL-divergence, and in the last transition we may drop the expectation over m_j since it is already accounted for in the divergence term. □

Now, since the index C is a stopping rule for the martingale $\mathcal{L}$, the optional stopping theorem [Doo75] implies that $\mathbb{E}[L_C]=L_0=0$, in other words

$$0=\mathbb{E}[L_C]=\mathbb{E}_{xyrcm}\left[\log\frac{\pi(M_{<c}=m_{<c}|xyr)}{\pi(M_{<c}=m_{<c}|yr)}+\log\frac{\pi(M_{<c}=m_{<c}|xyr)}{\pi(M_{<c}=m_{<c}|xr)}\right]-\mathbb{E}_{xyrcm}[\mathbb{D}^{\theta}_{XYR}(M_{<C})] \tag{54}$$

which proves (53), and together with (52), we conclude that

$$I(M_{<C};X|YR)+I(M_{<C};Y|XR)\le\mathbb{E}[\mathbb{D}^{\theta}_{XYR}(M_{<C})]+2\log(\|\theta\|+1)$$

as desired. Since C is also a stopping rule with respect to

$$L'_j:=\mathbb{D}^{\theta}_{XYRM}(T_{\le j})-\left[\log\frac{\theta(T_{\le j}|XYRM_{<C})}{\theta(T_{\le j}|YRM)}+\log\frac{\theta(T_{\le j}|XYRM)}{\theta(T_{\le j}|XRM)}\right],$$

repeating an analogues proof shows that $I(T_{\le C};X|YRM)+I(T_{\le C};Y|XRM)\le\mathbb{E}[\mathbb{D}^{\theta}_{XYRM}(T_{\le C})]+2\log(\|\theta\|+1)$, completing the proof of the entire Lemma. □

F Proof of Lemma 44

Proof. We show $\mathbb{E}[L_j \mid L_{j-1}] - L_{j-1} \geq 0$. To this end, Fix a choice of random variables up to L_{j-1} (i.e, fix $x_{<j}, y_{<j}$). Since the first $j-1$ terms cancel out,

$$\mathbb{E}[L_j \mid L_{j-1}] - L_{j-1} = \mathbb{E}[X_j|L_{j-1}] - \lambda \cdot \mathbb{E}[Z_t|L_{t-1}] - \frac{\mathbb{E}[Y_j|L_{j-1}] - \frac{1}{K}}{8\beta \cdot \log(2\beta K)}. \tag{55}$$

If $\mathbb{E}[Y_j|L_{j-1}] < \frac{1}{K}$, we are done, since the contribution of the third term is non-negative, and by the second premise of the lemma, $\mathbb{E}[X_j|L_{j-1}] - \lambda \cdot \mathbb{E}[Z_t|L_{t-1}] \geq \mathbb{E}[X_j|L_{j-1}]/2 \geq 0$, since $X_t \geq 0$ for all t. Otherwise, $\frac{1}{K} \leq \mathbb{E}[Y_j|L_{j-1}]$. Denoting $\Delta_j := \mathbb{E}[X_j|L_{j-1}]$, assumption (i) of the lemma implies that $\mathbb{E}[Y_j|L_{j-1}] \leq \beta \cdot H_{\Delta_j}$, so we have

$$\begin{aligned} &1/K \leq \mathbb{E}[Y_j|L_{j-1}] \leq \beta \cdot H_{\Delta_j} \leq 2\beta\sqrt{\Delta_j} \qquad \text{(by Proposition 13/}(i)\text{)} \\ &\implies \quad \log(1/\Delta_j) \leq 2\log(2\beta K). \end{aligned} \tag{56}$$

Therefore, (55) is at least

$$\begin{aligned} &\geq \Delta_j - \frac{\Delta_j}{2} - \lambda \cdot \mathbb{E}[Z_t|L_{t-1}] - \frac{\beta \cdot H_{\Delta_j} - \frac{1}{K}}{8\beta \cdot \log(2\beta K)} \qquad (\text{as } \mathbb{E}[Y_j|L_{j-1}] \leq \beta \cdot H_{\Delta_j} \text{ , } \lambda \cdot \mathbb{E}[Z_t|L_{t-1}] \leq \Delta_j/2) \\ &\geq \frac{\Delta_j}{2} - \frac{2\beta\Delta_j \log(1/\Delta_j)}{8\beta \cdot \log(2\beta K)} \geq \frac{\Delta_j}{2} - \frac{4\beta\Delta_j \log(2\beta K)}{8\beta \cdot \log(2\beta K)} \quad \text{(By (56))} \\ &= \frac{\Delta_j}{2} - \frac{\Delta_j}{2} = 0. \end{aligned}$$

This concludes that $\mathbb{E}[L_j \mid L_{j-1}] - L_{j-1} \geq 0$ and thus finishes the proof. □

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il