Electronic Colloquium on Computational Complexity, Report No. 81 (2016)

COMPRESSING INTERACTIVE COMMUNICATION UNDER PRODUCT DISTRIBUTIONS

ALEXANDER A. SHERSTOV

Abstract. We study the problem of compressing interactive communication to its information content I, defined as the amount of information that the participants learn about each other's inputs. We focus on the case when the participants' inputs are distributed independently and show how to compress the communication to $O(I \log^2 I)$ bits, with no dependence on the original communication cost. This result improves quadratically on previous work by Kol (STOC 2016) and essentially matches the well-known lower bound $\Omega(I)$.

[*] Computer Science Department, UCLA, Los Angeles, CA 90095. ✉ sherstov@cs.ucla.edu Supported by NSF CAREER award CCF-1149018 and an Alfred P. Sloan Foundation Research Fellowship.

ISSN 1433-8092

Contents

1. Introduction

Classic work by Shannon [23, 24] shows how to optimally compress one-way communication to its information content, achieving in the limit a transmission cost equal to the entropy of the message. The corresponding problem for *interactive* communication has attracted increasing attention over the past two decades. Consider two computationally unbounded parties, Alice and Bob, with inputs $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$, respectively, where $\mathscr{X}$ and $\mathscr{Y}$ are finite sets and the pair (X, Y) is distributed according to some known probability distribution on $\mathscr{X} \times \mathscr{Y}$. Alice and Bob exchange messages back and forth according to an agreed-upon randomized *protocol* in order to implement some functionality that depends on both inputs. One distinguishes between *public-coin* and *private-coin* protocols, corresponding to communication with or without a shared source of random bits. Information complexity theory [11, 3, 2, 4] studies a protocol's *information cost*, defined as the amount of information that Alice and Bob learn on average about each other's inputs from the history of messages exchanged between them (the *protocol transcript*). This complexity measure is quite different from *communication cost*, studied in Yao's communication complexity theory [25] and defined as the number of bits exchanged between Alice and Bob in the worst case on any input.

Basic properties of the entropy function ensure that a protocol's communication cost is always at least as large as its information cost, and the gap between the two quantities can be arbitrary. In this light, it is natural to ask whether the communication in every protocol π can be compressed to its information content while approximately preserving the protocol's functionality. In more detail, the approximate simulation of a given protocol π on given inputs X and Y by another protocol π' involves running π' on (X, Y) and interpreting the resulting transcript as a transcript of π. Alice and Bob may base their interpretations on their respective inputs X and Y, potentially arriving at distinct conclusions. In an accurate simulation, we require that their interpretations almost always agree and approximately follow the distribution of π's transcript on the input in question. Formally, π' *simulates* π *with error* ϵ if there exist a pair of "transcript interpretation" functions $a\colon \{0,1\}^* \to \{0,1\}^*$ and $b\colon \{0,1\}^* \to \{0,1\}^*$ for Alice and Bob such that the random variables (X, Y, Π, Π) and $(X, Y, a(X, \Pi'), b(Y, \Pi'))$ are at statistical distance at most ϵ, where Π and Π' denote the transcripts of π and π', respectively, on input (X, Y). The compression problem for interactive communication is the problem of simulating, with small error ϵ, a given protocol π by a protocol with communication cost as close as possible to the information cost of π. In addition to its basic importance, protocol compression is closely related to *direct sum theorems* in communication complexity theory [11, 19, 7].

Protocol compression has been actively studied [17, 4, 7, 6, 9, 10, 5, 22] over the past two decades. In a groundbreaking paper, Barak et al. [4] showed how to compress any protocol with information cost I and communication cost C to a protocol with communication cost $\sqrt{IC}\,\mathrm{polylog}(C)$. Since the original communication cost C can be essentially infinite, it is natural to ask if compression independent of C is a possibility. The influential results of Braverman [6] and Braverman and Weinstein [9] answer this question in the affirmative, showing how to compress the communication in any protocol to $2^{O(I)}$ bits. Despite much subsequent research, these two incomparable bounds remain the strongest results for general protocol compression. On the lower bounds side, Ganor, Kol, and Raz [13, 14, 15] prove that Braverman's $2^{O(I)}$ compression is in general the best possible bound that does not depend on the original communication cost C. It is consistent with our current knowledge, however, that any protocol can be compressed to $I\operatorname{polylog}(C)$ bits, with only a nominal dependence on the original communication cost.

In this paper, we focus on the well-studied special case [4, 8, 21] of the protocol compression problem when Alice and Bob's inputs X and Y are distributed independently. The resulting joint probability distribution μ of the inputs is called a *product distribution*, in reference to its representation as $\mu = \mu_{\mathscr{X}} \times \mu_{\mathscr{Y}}$ for some distributions $\mu_{\mathscr{X}}$ and $\mu_{\mathscr{Y}}$ on Alice and Bob's input sets, respectively. Table 1 gives a quantitative summary of this line of work. Braverman's $2^{O(I)}$ compression [6] of course applies to this special case as well, whereas Barak et al. [4] are able to strengthen their compression bound to $I\operatorname{polylog}(C)$ bits. These two bounds have complementary strengths, namely, independence of C and moderate growth with I. In a remarkable recent paper, Kol [21] shows how to achieve these desiderata simultaneously, for a compressed communication cost of $I^2\operatorname{polylog}(I)$ bits. We obtain a quadratic improvement on Kol's work, achieving a compressed communication cost of $O(I\log^2 I)$ bits and essentially matching the well-known lower bound of $\Omega(I)$.

THEOREM 1.1 (Main result). *Let $0 < \epsilon < 1/2$ be given. Fix any public- or private-coin protocol π with input space $\mathscr{X} \times \mathscr{Y}$. Let μ be a product distribution on $\mathscr{X} \times \mathscr{Y}$, and let I be the information cost of π under μ. Then there is a public-coin protocol π' that simulates π with error ϵ under μ and has worst-case communication cost*

$$O\left(\frac{I}{\epsilon}\log^2\frac{I}{\epsilon}\right).$$

Theorem 1.1 improves on previous compression schemes for product distributions with respect to all parameters. Our proof is inspired by the work of Barak et al. [4] and Kol [21], which we will describe shortly and contrast with our approach.

Compressed cost	**Reference**
$I\operatorname{polylog}(C)$	Barak et al. [4]
$2^{O(I)}$	Braverman [6], Braverman and Weinstein [9]
$I^2\operatorname{polylog}(I)$	Kol [21]
$O(I\log^2 I)$	This paper

Table 1: Compression of protocols with original communication cost C and information cost I under a product distribution ($I \leqslant C$).

1.1. Background for protocol compression. We start with a brief review of relevant terminology and background; a thorough treatment of these technical preliminaries is available in Section 2. Throughout this paper, we consider binary strings to be ordered by the prefix ordering $\preceq$. The terms *minimal* and *maximal*, when applied to strings, refer to this ordering $\preceq$. All trees in our work are binary and finite. We identify the vertices of a tree with binary strings in the usual manner, namely, the root corresponds to the empty string ε, and inductively the left child and right child of a vertex v correspond to the strings $v0$ and $v1$, respectively. A *cut* in a binary tree is any subset of the tree's vertices that intersects every root-to-leaf path in exactly one vertex. For example, the leaves of the tree form a cut. More generally, by truncating a given tree arbitrarily and considering the resulting set of leaves, one obtains a cut in the original tree. Given our identification of tree vertices with binary strings, we view cuts as subsets of $\{0,1\}^*$. The *floor* of cuts $\mathscr{C}_1$ and $\mathscr{C}_2$, denoted $\lfloor \mathscr{C}_1, \mathscr{C}_2 \rfloor$, is the set of minimal elements of $\mathscr{C}_1 \cup \mathscr{C}_2$. Analogously, the *ceiling* of cuts $\mathscr{C}_1$ and $\mathscr{C}_2$, denoted $\lceil \mathscr{C}_1, \mathscr{C}_2 \rceil$, is the set of maximal elements of $\mathscr{C}_1 \cup \mathscr{C}_2$. These definitions generalize in the obvious way to three or more cuts. For any collection of cuts in a binary tree, their floor and ceiling are also cuts in the same binary tree (Proposition 3.4 and 3.5). For visual correspondence with the floor and ceiling operations, we draw binary trees in this paper with the root at the bottom and leaves at the top.

Consider a randomized protocol with input space $\mathscr{X} \times \mathscr{Y}$. Assume for simplicity that it is a private-coin protocol, meaning that Alice and Bob do not have access to a shared source of random bits. They communicate by sending one bit at a time. A multibit message corresponds to several consecutive single-bit transmissions by the same sender. For any given history of previously transmitted bits, the protocol specifies which of the participants must send the next bit, which in turn is a function of the sender's private random string, the sender's input, and the history of previously transmitted bits. Formally, a private-coin protocol is given by a finite binary tree and a function $\pi\colon (\mathscr{A} \times \mathscr{X}) \cup (\mathscr{B} \times \mathscr{Y}) \to [0,1]$, where the sets $\mathscr{A}$ and $\mathscr{B}$ form a partition of the tree's internal vertices. We identify the protocol with its corresponding function π and use the same symbol for both. The vertices in $\mathscr{A}$ and $\mathscr{B}$ are said to be *owned* by Alice and Bob, respectively. The execution of π on a fixed pair of inputs (x, y) corresponds to a random walk on the protocol tree that starts at the root and proceeds one edge at a time, as follows. On reaching a vertex v owned by Alice, the walk proceeds to the left child with probability $\pi(v, x)$ and right child with the complementary probability $1 - \pi(v, x)$. Analogously, on reaching a vertex v owned by Bob, the walk proceeds to the left subtree with probability $\pi(v, y)$ and right subtree with probability $1 - \pi(v, y)$. The walk terminates upon reaching a leaf vertex, which represents a *transcript* of Alice and Bob's joint computation on input (x, y). In view of our identification of tree vertices with binary strings, the transcript on a given input (x, y) is a random variable with range $\{0,1\}^*$.

In the rest of the introduction, let π be an arbitrary but fixed private-coin protocol, and let μ be a product distribution on the protocol's input space $\mathscr{X} \times \mathscr{Y}$. Let I denote the information cost of π with respect to μ. Let X and Y be a pair of inputs with joint distribution μ, and let Π be the transcript of π on input (X, Y). For fixed values $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, define $P, P_x, P_y, P_{x,y}$ to be the probability distributions that govern the random variables

$$\begin{aligned}&\Pi,\\&\Pi \mid X = x,\\&\Pi \mid Y = y,\\&\Pi \mid X = x, Y = y,\end{aligned}$$

respectively. Thus, $P, P_x, P_y, P_{x,y}$ are probability distributions on the leaves of the protocol tree. For a leaf or internal vertex v, we define $P(v), P_x(v), P_y(v), P_{x,y}(v)$ to be the corresponding probabilities of reaching a leaf in the subtree rooted at v. With this convention, $P, P_x, P_y, P_{x,y}$ are nonnegative functions defined at every vertex of the protocol tree. Observe that the restriction of any one of these functions to a cut of the protocol tree is a probability distribution. We further use the shorthands $P(v \mid u), P_x(v \mid u), P_y(v \mid u), P_{x,y}(v \mid u)$ to refer to the probabilities of reaching a leaf in the subtree rooted at v conditioned on reaching a leaf in the subtree rooted at u. Using the fact that μ is a product distribution, one easily verifies the identity $P(v)P_{x,y}(v) = P_x(v)P_y(v)$ for all inputs $x \in \mathscr{X}$ and $y \in \mathscr{Y}$ and all vertices v.

With this setup in place, we now describe the work of Barak et al. [4] and Kol [21]. We alert the reader that our descriptions are somewhat adapted and reinterpreted versions of the original papers [4, 21]. In both cases, we have sought to convey the main ideas as simply and clearly as possible while maintaining consistency with the notation and methodology of this manuscript.

1.2. Sampling algorithm of Barak et al. For $x \in \mathscr{X}$ and an internal vertex v, define $\mathbb{D}_x(v)$ to be the Kullback–Leibler divergence between the Bernoulli distributions $(P_x(v0 \mid v), P_x(v1 \mid v))$ and $(P(v0 \mid v), P(v1 \mid v))$. Similarly, define $\mathbb{D}_y(v)$ to be the Kullback–Leibler divergence between the Bernoulli distributions $(P_y(v0 \mid v), P_y(v1 \mid v))$ and $(P(v0 \mid v), P(v1 \mid v))$. Let $0 < \delta < 1$ be a small parameter, with order of magnitude $\delta = O(1/\log I)$. Without loss of generality [4], we may assume that $\mathbb{D}_x(v) \leqslant \delta$ and $\mathbb{D}_y(v) \leqslant \delta$ for all v, x, y. A key notion introduced by Barak et al. is that of a *δ-frontier*, defined separately for Alice and Bob. Alice's δ-frontier $\mathscr{F}_{x,\delta}$ is the set of minimal vertices v such that either v is a leaf or the sum of the $\mathbb{D}_x$ values of v's proper ancestors is at least δ. Analogously, Bob's δ-frontier $\mathscr{F}_{y,\delta}$ is the set of minimal vertices v such that either v is a leaf or the sum of the $\mathbb{D}_y$ values of v's proper ancestors is at least δ. A moment's reflection shows that $\mathscr{F}_{x,\delta}$ and $\mathscr{F}_{y,\delta}$ are cuts in the protocol tree.

Execution of π on input X, Y corresponds to sampling a random leaf of the protocol tree according to the probability distribution $P_{X,Y}$. Unfortunately, neither Alice nor Bob knows $P_{X,Y}$. Indeed, Alice only knows P and P_X, and Bob only knows P and P_Y. As the technical centerpiece of their analysis, Barak et al. prove that the restrictions of P_X and P_Y to the cut $\lfloor \mathscr{F}_{X,\delta}, \mathscr{F}_{Y,\delta} \rfloor$ are within a multiplicative constant c_0 of P almost at every vertex. We assume in this overview that the multiplicative bound holds everywhere. Under this simplifying assumption, the sampling procedure is as follows. Alice and Bob start by computing their respective frontiers $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$. They then use the shared randomness to sample a vertex V of the cut $\lfloor \mathscr{F}_{X,\delta}, \mathscr{F}_{Y,\delta} \rfloor$ according to the probability distribution P, by sampling a leaf according to P and sending each other its ancestors in $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$, respectively. To adjust for any multiplicative disparity between P and $P_{X,Y}$, they use *rejection sampling* [17, 18, 20, 4], whereby Alice accepts V with probability $P_X(V)/c_0P(V)$ and Bob independently accepts V with probability $P_Y(V)/c_0P(V)$. Conditioned on both parties accepting, which happens with probability $1/c_0^2$, the vertex V is a random element of the cut $\lfloor \mathscr{F}_{X,\delta}, \mathscr{F}_{Y,\delta} \rfloor$ governed by the correct probability distribution:

$$P(V) \cdot \frac{P_X(V)}{P(V)} \cdot \frac{P_Y(V)}{P(V)} = \frac{P_X(V)P_Y(V)}{P(V)} = P_{X,Y}(V).$$

By generating V in this manner, Barak et al. execute the initial part of π that corresponds to the shaded region of the protocol tree in Figure 1.1 (left). They then run their algorithm recursively on the protocol subtree rooted at V, eventually outputting a leaf distributed according to $P_{X,Y}$. For the cost analysis, consider the intermediate vertices generated by the algorithm as it works its way from the root to a leaf. The path segment between any two of them contributes at least δ toward the path's cumulative $\mathbb{D}_X$ or $\mathbb{D}_Y$ value. By the chain rule for the Kullback–Leibler divergence, it follows that the process terminates on average after $O(I/\delta) = O(I \log I)$ recursive calls. The communication cost of a single recursive call is $O(\log C)$, where C is the height of the protocol tree for π. As a result, the overall simulation has communication cost $I \operatorname{polylog}(C)$.

1.3. Kol's sampling algorithm. The most expensive step in the algorithm of Barak et al. is the transmission of the intersection points of $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$ with the root-to-leaf path sampled according to P. Their implementation involves the exchange of the actual intersection points, for a communication cost of $\Theta(\log C)$ bits, which can be essentially infinite even when the information cost I is small.

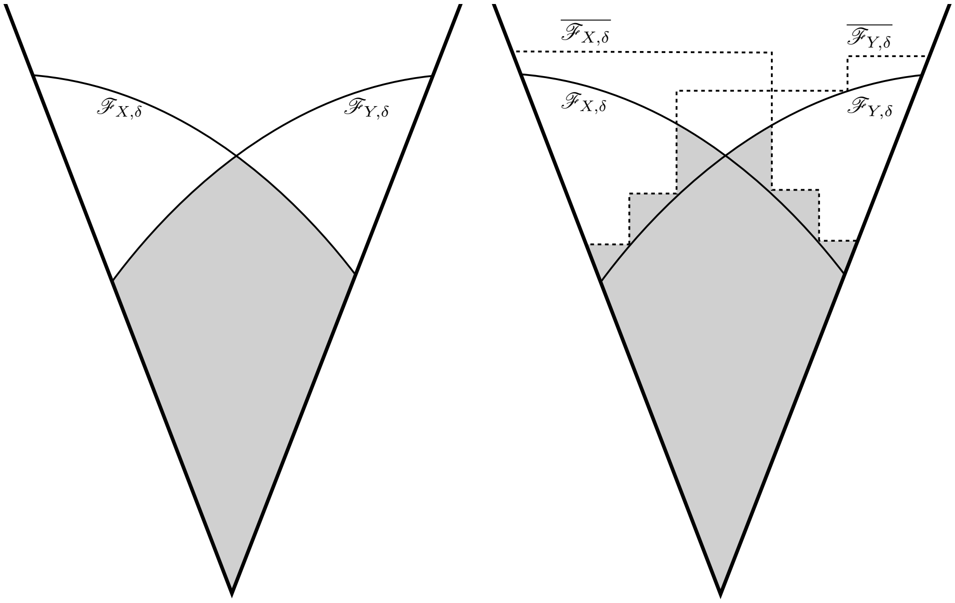

Figure 1.1: The sampling step in the algorithms of Barak et al. (left) and Kol (right). The shaded area corresponds to the sampling subtree.

Kol [21] proposed an alternate sampling procedure, based on discretization, that ingeniously eliminates the dependence of the cost on C. Specifically, Kol rounds the frontiers $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$ up with respect to a small and fixed collection of cuts known to both Alice and Bob, resulting in a pair of approximate frontiers $\overline{\mathscr{F}_{X,\delta}}$ and $\overline{\mathscr{F}_{Y,\delta}}$. Figure 1.1 (right) illustrates Kol's construction, with the approximate frontiers shown as dashed lines. Instead of sampling from the cut $\lfloor \mathscr{F}_{X,\delta}, \mathscr{F}_{Y,\delta} \rfloor$ as Barak et al. do, Kol samples from the cut $\lceil \lfloor \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{Y,\delta} \rfloor, \lfloor \overline{\mathscr{F}_{Y,\delta}}, \mathscr{F}_{X,\delta} \rfloor \rceil$. Using the fact that μ is a product distribution, Kol shows that this new sampling cut coincides almost always with $\lfloor \overline{\mathscr{F}_{X,\delta}}, \overline{\mathscr{F}_{Y,\delta}} \rfloor$ and therefore enables the efficient transmission of the intersection points with any root-to-leaf path.

Assuming for simplicity that Alice and Bob's frontiers $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$ are disjoint, Kol's complete sampling algorithm is as follows. First, one of the parties is randomly designated as the *leader*. Under Alice's leadership, the algorithm starts by sampling a root-to-leaf path according to P_X. This step uses the correlated sampling algorithm of Braverman and Rao [7] for the probability distributions P_X and P, with expected communication cost $O(\mathbf{E}\,\mathrm{KL}(P_X \parallel P)) \leqslant O(I)$. If Bob's frontier $\mathscr{F}_{Y,\delta}$ precedes Alice's frontier $\mathscr{F}_{X,\delta}$ along the sampled path, they reject the path and go back to randomly choosing a leader. Otherwise, they compute the path's intersection V with the cut $\lfloor \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{Y,\delta} \rfloor$, and Bob performs rejection sampling on V as in the work of Barak et al. If Bob rejects V, they go back to randomly choosing a leader; otherwise they accept V and run the algorithm recursively on the subtree rooted at V. This completes the description of the algorithm when Alice is the leader. Under Bob's leadership, the roles of Alice and Bob, and the roles of X and Y, are reversed. The cost analysis is similar to that of Barak et al., with the difference that the expected cost of a recursive call is now $O(I)$ rather than $O(\log I)$. Since the expected number of recursive calls does not exceed $I\,\mathrm{polylog}(I)$, the overall algorithm has communication cost $I^2\,\mathrm{polylog}(I)$.

1.4. Our sampling algorithm. Kol's algorithm incurs essentially its entire communication cost at the beginning of a recursive call, when sampling a root-to-leaf path. The expected communication cost $\Theta(I)$ of this operation far exceeds its expected contribution $\Theta(1/\log I)$ to the cumulative $\mathbb{D}_X$ or $\mathbb{D}_Y$ value of the path that the algorithm eventually outputs. There are two reasons for this inefficiency. First, the portion of the sampled path beyond the sampling cut is always discarded, forfeiting the corresponding sampling effort. Second, the entire sampled path is discarded if the follower's frontier precedes the leader's along that path. We eliminate both sources of inefficiency and obtain an algorithm in which every step has communication cost proportional to that step's expected contribution to the progress measure.

We address the first problem by sampling the root-to-leaf path according to a "hybrid" distribution. The portion of the path up to the *leader's* sampling cut is distributed according to either P_X or P_Y as in Kol's algorithm, whereas the rest of the path is distributed according to the publicly known distribution P. The effect of this modification is that the segment of the path beyond the leader's sampling cut does not contribute to the sampling cost. To address the second source of inefficiency, we use a sampling cut different from Kol's. Let $\mathscr{R}_{\mathscr{X},\delta,1/2}$ denote the set of minimal vertices v such that the frontier $\mathscr{F}_{x,\delta}$ is encountered on the path from the root to v for at least half of the inputs $x \in \mathscr{X}$ weighted according to μ. Define $\mathscr{R}_{\mathscr{Y},\delta,1/2}$ analogously, and abbreviate $\mathscr{R}_{\delta,1/2} = \lfloor \mathscr{R}_{\mathscr{X},\delta,1/2}, \mathscr{R}_{\mathscr{Y},\delta,1/2} \rfloor$.

These definitions ensure that for random X and Y, neither of the frontiers $\mathscr{F}_{X,\delta}$ or $\mathscr{F}_{Y,\delta}$ is very likely to precede $\mathscr{R}_{\delta,1/2}$ along a fixed root-to-leaf path. This motivates the use of $\lceil \lfloor \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{Y,\delta}, \mathscr{R}_{\delta,1/2} \rfloor, \lfloor \overline{\mathscr{F}_{Y,\delta}}, \mathscr{F}_{X,\delta}, \mathscr{R}_{\delta,1/2} \rfloor \rceil$ as the sampling cut, instead of Kol's $\lceil \lfloor \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{Y,\delta} \rfloor, \lfloor \overline{\mathscr{F}_{Y,\delta}}, \mathscr{F}_{X,\delta} \rfloor \rceil$. Figure 1.2 illustrates the resulting sampling subtree. To be precise, the sampling cut that we actually use is $\lceil \lfloor \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{X,\Delta}, \mathscr{F}_{Y,\delta}, \mathscr{R}_{\delta,1/2} \rfloor, \lfloor \overline{\mathscr{F}_{Y,\delta}}, \mathscr{F}_{Y,\Delta}, \mathscr{F}_{X,\delta}, \mathscr{R}_{\delta,1/2} \rfloor \rceil$ for a large parameter $\Delta \gg 1$, but the distinction can be ignored on a first reading.

Summarizing, our modifications ensure that the sampling cost of every step in the algorithm is a constant plus a quantity proportional to the step's expected contribution to the progress measure. To prove that the overall sampling cost is at most $I \operatorname{polylog}(I)$, we must further argue that every step of the algorithm contributes on average $1/\operatorname{polylog}(I)$ to the progress measure. The corresponding claims in the work of Barak et al. and Kol were trivial to prove. In particular, the leader in Kol's algorithm is always guaranteed to contribute at least δ to the progress measure. Our situation is different because our choice of sampling cut effectively truncates the tree at $\mathscr{R}_{\delta,1/2}$, making a zero contribution a possibility for both the leader and the follower. Information-theoretically, the difficulty is as follows. For any *fixed* vertex $v \in \mathscr{R}_{\delta,1/2}$ and random X and Y, the probability that at least one of the frontiers $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$ is encountered on the path from the root to v is at least $1/2$. However, the sampled vertex V in the sampling cut is neither fixed nor independent of X or Y. We solve the problem by showing that any correlation between V and the protocol inputs causes information to be revealed about X and Y in a way that on average contributes to the progress measure instead of defeating it. We complete the proof of our main result with an amortized analysis of the cost versus progress, which too is more demanding than in previous work.

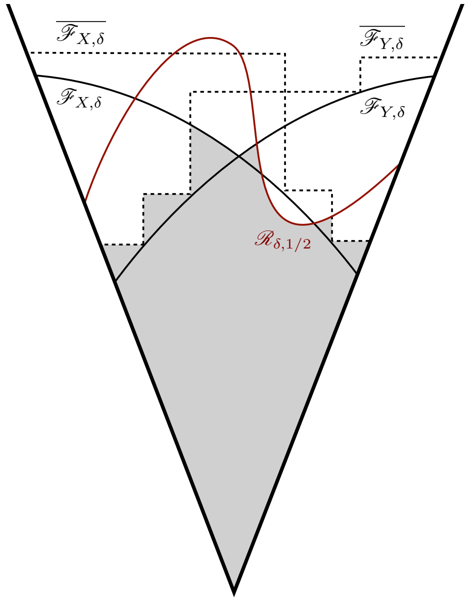
$\overline{\mathscr{F}_{X,\delta}}$
$\overline{\mathscr{F}_{Y,\delta}}$
$\mathscr{F}_{X,\delta}$
$\mathscr{F}_{Y,\delta}$
$\mathscr{R}_{\delta,1/2}$

Figure 1.2: The sampling step in this paper. The shaded area corresponds to the sampling subtree.

2. Preliminaries

We let $\log x$ denote the logarithm of x to base 2. We adopt the convention that $0/0 = 0$, justified throughout this paper by continuity arguments. For a binary string v, the shorthand $|v|$ stands for the length of v. We use calligraphic letters for finite sets ($\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{X}, \mathscr{Y}$), lowercase letters for set elements (x, y, u, v, w), and uppercase letters for random variables (X, Y, U, V, W). For a random variable X and an event E in the probability space, we let $X \mid E$ denote the random variable obtained from X by conditioning on E. The notation $X \sim \mu$ means that the random variable X is governed by the probability distribution μ. For random variables X and Y with a certain joint probability distribution, recall that $\mathbf{E}[Y \mid X]$ is not a specific number but a random variable defined as a function of X. Specifically, $\mathbf{E}[Y \mid X] = f(X)$ where f is given by $f(x) = \mathbf{E}[Y \mid X = x]$. Analogously, $\mathbf{P}[E \mid X]$ for an event E is not a specific number but a random variable defined as a function of X.

2.1. Strings. Recall that $\{0,1\}^*$ and $\{0,1\}^+$ refer to the set of binary strings and the set of nonempty binary strings, respectively. The empty string is denoted ε. The concatenation of the strings u and v is denoted uv. Consider the standard partial order $\prec$ on $\{0,1\}^*$, whereby $u \prec v$ if and only if $uw = v$ for a nonempty string w. The derived relations $\succ, \preceq, \succeq$ are defined as usual by

$$\begin{aligned} u \succ v \qquad &\Leftrightarrow \qquad v \prec u, \\ u \succeq v \qquad &\Leftrightarrow \qquad v \prec u \text{ or } v = u, \\ u \preceq v \qquad &\Leftrightarrow \qquad u \prec v \text{ or } v = u. \end{aligned}$$

Strings u and v are called *comparable* if $u \preceq v$ or $u \succeq v$, and *incomparable* otherwise. In addition to their role as relational operators, we use $\prec, \succ, \preceq, \succeq$ as the unary operators given by

$$\begin{aligned} {\prec}v &= \{u : u \prec v\}, \\ {\succ}v &= \{u : u \succ v\}, \\ {\preceq}v &= \{u : u \preceq v\}, \\ {\succeq}v &= \{u : u \succeq v\}. \end{aligned}$$

We refer to the elements of ${\preceq}v$ and ${\succeq}v$ as the *ancestors of* v and the *descendants of* v, respectively. Analogously, we call the elements of ${\prec}v$ and ${\succ}v$ the *proper ancestors of* v and the *proper descendants of* v, respectively. These unary operators naturally extend from strings to *sets* of strings, according to

$${\prec}\mathscr{V} = \bigcup_{v \in \mathscr{V}} {\prec}v, \qquad {\succ}\mathscr{V} = \bigcup_{v \in \mathscr{V}} {\succ}v, \qquad {\preceq}\mathscr{V} = \bigcup_{v \in \mathscr{V}} {\preceq}v, \qquad {\succeq}\mathscr{V} = \bigcup_{v \in \mathscr{V}} {\succeq}v.$$

In their unary capacity, the operators $\prec, \succ, \preceq, \succeq$ have the highest precedence. To illustrate,

$$\begin{aligned} &{\prec}u \setminus {\prec}v = ({\prec}u) \setminus ({\prec}v), \\ &{\preceq}v \cap \mathscr{V} = ({\preceq}v) \cap \mathscr{V}. \end{aligned}$$

2.2. Kullback–Leibler divergence. In this subsection and the next, we provide relevant background from information theory. All definitions and facts referenced here are well-known and can be found in any standard text on information theory, such as Cover and Thomas [12]. For the reader's convenience, we provide proofs for the more specialized of the facts that we use.

All probability distributions in this work are defined on finite sets. For a probability distribution p on a set $\mathscr{X}$, its *support* is given by $\operatorname{supp} p = \{x \in \mathscr{X} : p(x) \neq 0\}$. For a subset $\mathscr{X}' \subseteq \mathscr{X}$, we let $p|_{\mathscr{X}'}$ denote the probability distribution induced by p on $\mathscr{X}'$. For probability distributions p and q on $\mathscr{X}$, their *Kullback–Leibler divergence* is given by

$$\mathrm{KL}(p \parallel q) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)}.$$

In the context of the Kullback–Leibler divergence, we frequently identify a real number $0 \leqslant p \leqslant 1$ with the corresponding Bernoulli distribution $(p, 1-p)$ and use the shorthand $\mathrm{KL}(p \parallel q) = \mathrm{KL}((p, 1-p) \parallel (q, 1-q))$. The following estimate can be verified using elementary calculus:

$$\mathrm{KL}\left(\frac{p}{3} \;\Big\|\; p\right) \geqslant \frac{p}{3}, \qquad\qquad 0 \leqslant p \leqslant 1. \tag{2.1}$$

The Kullback–Leibler divergence $\mathrm{KL}(p \parallel q)$ is a measure of distance for probability distributions p and q in that it is always nonnegative, with $\mathrm{KL}(p \parallel q) = 0$ if and only if $p = q$. It falls short of being a metric because in general, it is not symmetric and does not obey the triangle inequality. The Kullback–Leibler divergence does, however, have the following approximate symmetry property, which too can be verified using elementary calculus:

$$\sup_{\substack{p,q \in [1/3, 2/3] \\ p \neq q}} \frac{\mathrm{KL}(p \parallel q)}{\mathrm{KL}(q \parallel p)} < \frac{21}{20}. \tag{2.2}$$

For the sake of completeness, we note that this qualitative phenomenon holds in greater generality.

Proposition 2.1. *For all* $0 < \epsilon < 1/2$,

$$\sup_{\substack{p,q \in [\epsilon, 1-\epsilon] \\ p \neq q}} \frac{\mathrm{KL}(p \parallel q)}{\mathrm{KL}(q \parallel p)} < \infty. \tag{2.3}$$

Proof. Let M denote the left-hand side of (2.3). By compactness, there is a sequence $\{(p_n, q_n)\}_{n=1}^{\infty}$ in $[\epsilon, 1-\epsilon]^2$ such that $\mathrm{KL}(p_n \parallel q_n)/\mathrm{KL}(q_n \parallel p_n) \to M$ and $(p_n, q_n) \to (p, q)$. If $p \neq q$, then $\mathrm{KL}(p \parallel q)$ and $\mathrm{KL}(q \parallel p)$ are finite and positive, whence $\mathrm{KL}(p_n \parallel q_n)/\mathrm{KL}(q_n \parallel p_n) \to \mathrm{KL}(p \parallel q)/\mathrm{KL}(q \parallel p) < \infty$. In the complementary case $p = q$, the Taylor series for the logarithm gives

$$\mathrm{KL}(a \parallel b) = \frac{(a-b)^2}{b(1-b)\ln 4} \cdot (1 + O(a-b)) \tag{2.4}$$

for all $a, b \in [\epsilon, 1-\epsilon]$, so that $\mathrm{KL}(p_n \,\|\, q_n)/\mathrm{KL}(q_n \,\|\, p_n) \to 1$. □

As the next result shows [12, Theorem 2.7.2], the Kullback–Leibler divergence $\mathrm{KL}(p \,\|\, q)$ is convex in the pair (p, q).

Fact 2.2. *Fix probability distributions $p_1, p_2, \ldots, p_k$ and $q_1, q_2, \ldots, q_k$ on a given finite set $\mathscr{X}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be nonnegative reals with $\sum \lambda_i = 1$. Then*

$$\sum_{i=1}^{k} \lambda_i \,\mathrm{KL}(p_i \,\|\, q_i) \geqslant \mathrm{KL}\left(\sum_{i=1}^{k} \lambda_i p_i \,\middle\|\, \sum_{i=1}^{k} \lambda_i q_i \right).$$

We now recall a basic optimization problem pertaining to the Kullback–Leibler divergence. Let $p_1, p_2, \ldots, p_k$ be probability distributions on a given finite set. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be nonnegative weights with $\sum \lambda_i = 1$. Consider the problem of finding a probability distribution p that minimizes the weighted sum

$$\sum_{i=1}^{k} \lambda_i \,\mathrm{KL}(p_i \,\|\, p).$$

The optimal distribution, $p^* = \sum \lambda_i p_i$, is easy to guess based on convexity considerations (Fact 2.2). There are both analytic and information-theoretic ways to verify the optimality of p^*. For the reader's convenience, we include a proof that is at once short and self-contained.

Fact 2.3. *Fix probability distributions $p_1, p_2, \ldots, p_k$ on a given finite set $\mathscr{X}$ and nonnegative reals $\lambda_1, \lambda_2, \ldots, \lambda_k$ with $\sum \lambda_i = 1$. Then the minimum*

$$\min_{p} \left\{ \sum_{i=1}^{k} \lambda_i \,\mathrm{KL}(p_i \,\|\, p) \right\}$$

is achieved at

$$p^* = \sum_{i=1}^{k} \lambda_i p_i.$$

Proof. For any probability distribution p,

$$\begin{aligned}\sum_{i=1}^k \lambda_i \operatorname{KL}(p_i \parallel p) - \sum_{i=1}^k \lambda_i \operatorname{KL}(p_i \parallel p^*)&\\ &= \sum_{i=1}^k \lambda_i \sum_{x\in\mathscr{X}} p_i(x)\log\frac{p_i(x)}{p(x)} - \sum_{i=1}^k \lambda_i \sum_{x\in\mathscr{X}} p_i(x)\log\frac{p_i(x)}{p^*(x)}\\ &= \sum_{i=1}^k \lambda_i \sum_{x\in\mathscr{X}} p_i(x)\log\frac{1}{p(x)} - \sum_{i=1}^k \lambda_i \sum_{x\in\mathscr{X}} p_i(x)\log\frac{1}{p^*(x)}\\ &= \sum_{x\in\mathscr{X}} p^*(x)\log\frac{p^*(x)}{p(x)}\\ &= \operatorname{KL}(p^* \parallel p)\\ &\geqslant 0,\end{aligned}$$

where the final step uses the nonnegativity of the Kullback–Leibler divergence. □

The Kullback–Leibler divergence satisfies the following *chain rule*, which is particularly useful when analyzing stochastic processes with hierarchical structure such as random walks in trees.

FACT 2.4 (Chain rule). *Let p and q be probability distributions on a given finite set $\mathscr{X}$. Then for any partition $\mathscr{X} = \bigcup_{i=1}^k \mathscr{X}_i$,*

$$\begin{aligned}\operatorname{KL}(p \parallel q) = \operatorname{KL}((p(\mathscr{X}_1),\ldots,p(\mathscr{X}_k)) \parallel (q(\mathscr{X}_1),\ldots,q(\mathscr{X}_k)))&\\ + \sum_{i=1}^k p(\mathscr{X}_i)\operatorname{KL}(p|_{\mathscr{X}_i} \parallel q|_{\mathscr{X}_i}).&\end{aligned}$$

2.3. Statistical distance. Another distance measure for probability distributions is *statistical distance*, also known as *total variation distance* and defined for p and q by

$$\operatorname{TV}(p,q) = \max_{\mathscr{E}\subseteq\mathscr{X}} |p(\mathscr{E}) - q(\mathscr{E})|.$$

Unlike the Kullback–Leibler divergence, statistical distance is an actual metric. The following fundamental inequality relates the two notions.

FACT 2.5 (Pinsker's inequality). *For any probability distributions p and q on a given set $\mathscr{X}$,*

$$\operatorname{TV}(p,q)^2 \leqslant \frac{\ln 2}{2}\operatorname{KL}(p \parallel q).$$

We will also need the following first-principles bound on statistical distance.

FACT 2.6. *Let p and q be probability distributions on $\mathscr{X}$ such that $p(x) \leqslant c \cdot q(x)$ for all $x \in \mathscr{X}$. Then*

$$\mathrm{TV}(p,q) \leqslant 1 - \frac{1}{c}.$$

Proof:

$$\begin{aligned}
\mathrm{TV}(p,q) &= \sum_{x : p(x) \geqslant q(x)} (p(x) - q(x)) \\
&\leqslant \sum_{x : p(x) \geqslant q(x)} \left(p(x) - \frac{1}{c} \cdot p(x) \right) \\
&= \left(1 - \frac{1}{c} \right) \sum_{x : p(x) \geqslant q(x)} p(x) \\
&\leqslant 1 - \frac{1}{c}. \qquad \square
\end{aligned}$$

In the context of the Kullback–Leibler divergence and statistical distance, we identify random variables with their corresponding probability distributions. For example, the notation $\mathrm{TV}(X,Y)$ refers to the statistical distance between the probability distributions of X and Y.

2.4. Mutual information. While the Kullback–Leibler divergence and statistical distance measure the distance between two probability distributions, mutual information measures how far two random variables are from being independent. Let X and Y be random variables with domains $\mathscr{X}$ and $\mathscr{Y}$, respectively, governed by some joint probability distribution. The *mutual information* of X and Y is defined as

$$\begin{aligned}
I(X;Y) &= \sum_{y \in \mathscr{Y}} \mathbf{P}[Y = y] \, \mathrm{KL}(X \mid Y = y \parallel X) \qquad (2.5) \\
&= \sum_{x \in \mathscr{X}} \mathbf{P}[X = x] \, \mathrm{KL}(Y \mid X = x \parallel Y),
\end{aligned}$$

where second equality is straightforward to verify. In particular, the mutual information $I(X;Y)$ is always nonnegative, with $I(X;Y) = 0$ if and only if X and Y are independent random variables. It is also clear that mutual information is symmetric:

$$I(X;Y) = I(Y;X).$$

Given an additional random variable Z with domain $\mathscr{Z}$, the *conditional mutual information* $I(X;Y \mid Z)$ is given by

$$I(X;Y \mid Z) = \sum_{z \in \mathscr{Z}} \mathbf{P}[Z = z] \, I(X \mid Z = z;\ Y \mid Z = z).$$

In particular, the conditional mutual information $I(X; Y \mid Z)$ is always nonnegative, with $I(X; Y \mid Z) = 0$ if and only if the random variables X and Y are conditionally independent given Z. A moment's thought reveals that the mutual information may increase, decrease, or remain unchanged as a result of conditioning. The symmetry of mutual information continues to hold in the presence of conditioning:

$$I(X; Y \mid Z) = I(Y; X \mid Z).$$

Mutual information satisfies the chain rule

$$I(X_1 X_2 \ldots X_k; Y) = \sum_{i=1}^{k} I(X_i; Y \mid X_1 X_2 \ldots X_{i-1}).$$

2.5. Probability distributions in binary trees. Fix a binary tree T and let μ be a probability distribution on the leaves of T. Throughout this paper, we identify the vertices of T with binary strings in the usual manner: the root vertex corresponds to the empty string ε, and inductively the left child and right child of a vertex v correspond to $v0$ and $v1$, respectively. For a vertex v of the tree, which can be either a leaf or an internal vertex, we let $\mu(v)$ stand for the probability of reaching a leaf in the subtree of v. Similarly, $\mu(v \mid u)$ denotes the probability of reaching a leaf in the subtree of v conditioned on reaching a leaf in the subtree of u. The following theorem exploits the hierarchical structure of a binary tree to give an alternate expression for the Kullback–Leibler divergence of two probability distributions on the tree leaves.

THEOREM 2.7. *Let μ and $\tilde{\mu}$ be probability distributions on the leaves of a given binary tree. For an internal vertex v, abbreviate*

$$\mathbb{D}(v) = \mathrm{KL}(\mu(v0 \mid v) \parallel \tilde{\mu}(v0 \mid v)).$$

Then

$$\mathrm{KL}(\mu \parallel \tilde{\mu}) = \mathop{\mathbf{E}}_{V \sim \mu} \left[\sum_{v : v \prec V} \mathbb{D}(v) \right].$$

Proof. Immediate by induction on tree depth, with the inductive step following from the chain rule for the Kullback–Leibler divergence (Fact 2.4). ∎

The next theorem states that two probability distributions on the leaves of a binary tree are multiplicatively close if the Kullback–Leibler divergence on any root-to-leaf path is small. The theorem is a minor adaptation of a result due to Barak et al. [4].

THEOREM 2.8. *Let μ and $\tilde{\mu}$ be probability distributions on the leaves of a binary tree. For an internal vertex v, abbreviate*

$$\mathbb{D}(v) = \mathrm{KL}(\mu(v0 \mid v) \parallel \tilde{\mu}(v0 \mid v)).$$

Assume that:

(i) $\mu(v0 \mid v), \tilde{\mu}(v0 \mid v) \in [1/3, 2/3]$ *for every internal vertex* v;

(ii) $\sum_{u : u \prec v} \mathbb{D}(u) \leqslant \theta$ *for every leaf* v.

Then:

$$\mathop{\mathbf{P}}_{V \sim \mu} \left[\mu(V) \geqslant 2^{c+\theta} \tilde{\mu}(V)\right] \leqslant \exp\left(-\frac{c^2}{52\theta}\right), \qquad c \geqslant 0, \tag{2.6}$$

$$\mathop{\mathbf{P}}_{V \sim \tilde{\mu}} \left[\tilde{\mu}(V) \geqslant 2^{c+(21/20)\theta} \mu(V)\right] \leqslant \exp\left(-\frac{c^2}{55\theta}\right), \qquad c \geqslant 0. \tag{2.7}$$

Part (2.6) of the theorem conclusion was proved, with different constants, by Barak et al. [4]. Our applications require the additional upper bound (2.7), which we infer by appealing to (2.2). For the reader's convenience, we provide a complete and self-contained proof of Theorem 2.8 in Appendix A.

2.6. Communication protocols. We consider communication between two computationally unbounded parties, called Alice and Bob, each with an input from some fixed finite set and with a private source of random bits. They send messages back and forth according to an agreed-upon protocol, where each message is a function of the sender's input, the sender's private random bits, and previously exchanged messages. Formally, a *private-coin communication protocol* is a tuple $(\mathscr{X}, \mathscr{Y}, T, \mathscr{A}, \mathscr{B}, \pi)$, where $\mathscr{X}$ and $\mathscr{Y}$ are the sets of possible inputs for Alice and Bob, respectively; T is a finite nonempty binary tree; $\mathscr{A}$ and $\mathscr{B}$ are disjoint sets that form a partition of the internal vertices of T; and $\pi \colon (\mathscr{A} \times \mathscr{X}) \cup (\mathscr{B} \times \mathscr{Y}) \to [0,1]$ is any function. The tree T is called the *protocol tree*. The vertices of $\mathscr{A}$ are said to be *owned by Alice*, and those of $\mathscr{B}$ are said to be *owned by Bob*. For brevity, we will identify a communication protocol with its corresponding function π since the other components $\mathscr{X}, \mathscr{Y}, T, \mathscr{A}, \mathscr{B}$ of the tuple can all be recovered from the domain of π.

The operational interpretation of a protocol π on a given pair of inputs $x \in \mathscr{X}$ and $y \in \mathscr{Y}$ is in terms of a random walk from the root of the protocol tree to a leaf. Specifically, at an internal vertex $v \in \mathscr{A}$, Alice sends 0 with probability $\pi(v, x)$ and sends 1 with the complementary probability $1 - \pi(v, x)$, directing the random walk to the left or right subtree, respectively. At an internal vertex $v \in \mathscr{B}$, Bob analogously sends 0 with probability $\pi(v, y)$, directing the random walk to the left subtree, and sends 1 with complementary probability. A *transcript* is the complete sequence of bits sent by Alice and Bob on a given pair of inputs over the course of the random walk from the root of the protocol tree to a leaf. Given our identification of tree vertices with binary strings, we identify the transcript with the leaf reached by the random walk. The *communication cost* of protocol π, denoted $|\pi|$, is the height of the protocol tree, or equivalently the maximum number of bits exchanged by Alice and Bob in the worst case on any input. We let $\mathscr{V}(\pi)$ denote the set of vertices of the protocol tree for π, which includes both the internal vertices and the leaves. The set of leaves of the protocol tree is denoted $\mathscr{L}(\pi)$. We regard $\mathscr{V}(\pi)$ and $\mathscr{L}(\pi)$ as subsets of $\{0,1\}^*$.

A *public-coin communication protocol* is a probability distribution over a finite number of private-coin communication protocols, each with its own protocol tree. In a public-coin protocol, Alice and Bob use a shared source of random bits (a "public coin") to sample a random string R and then proceed to execute the private-coin protocol that corresponds to R. The *communication cost* of a public-coin protocol π, denoted $|\pi|$, is the maximum communication cost of the associated private-coin protocols. In particular, the length of the shared random string R does not count toward the communication cost of a public-coin protocol. The shared string is, however, always considered to be a part of the protocol transcript. Unless indicated otherwise, the term *protocol* throughout this paper refers to a private-coin protocol.

Recall from the introduction that a *product distribution* on $\mathscr{X} \times \mathscr{Y}$ is any probability distribution of the form $\mu(x,y) \equiv \mu_{\mathscr{X}}(x)\mu_{\mathscr{Y}}(y)$, where $\mu_{\mathscr{X}}$ and $\mu_{\mathscr{Y}}$ are probability distributions on $\mathscr{X}$ and $\mathscr{Y}$, respectively. The following fact is well-known.

FACT 2.9. *For any private-coin protocol π, product distribution μ, and vertex $v \in \mathscr{V}(\pi)$,*

$$\mathbf{P}[\Pi \succeq v]\,\mathbf{P}[\Pi \succeq v \mid X, Y] = \mathbf{P}[\Pi \succeq v \mid X]\,\mathbf{P}[\Pi \succeq v \mid Y], \tag{2.8}$$

where X, Y are random variables with joint distribution μ, and Π is the protocol transcript on input X, Y.

Proof. For notational convenience, we will assume that $v = 0^k$ for some k. Then by definition,

$$\mathbf{P}[\Pi \succeq v \mid X, Y] = \prod_{u \in \mathscr{A} \cap \prec v} \pi(u, X) \cdot \prod_{u \in \mathscr{B} \cap \prec v} \pi(u, Y). \tag{2.9}$$

Passing to expectations,

$$\begin{aligned}\mathbf{P}[\Pi \succeq v] &= \mathbf{E}\left[\prod_{u \in \mathscr{A} \cap \prec v} \pi(u, X) \cdot \prod_{u \in \mathscr{B} \cap \prec v} \pi(u, Y)\right] \\ &= \mathbf{E}\left[\prod_{u \in \mathscr{A} \cap \prec v} \pi(u, X)\right] \mathbf{E}\left[\prod_{u \in \mathscr{B} \cap \prec v} \pi(u, Y)\right], \end{aligned} \tag{2.10}$$

where the last step uses the independence of X and Y. By an analogous argument, (2.9) yields

$$\mathbf{P}[\Pi \succeq v \mid X] = \prod_{u \in \mathscr{A} \cap \prec v} \pi(u, X) \cdot \mathbf{E}\left[\prod_{u \in \mathscr{B} \cap \prec v} \pi(u, Y)\right] \tag{2.11}$$

and

$$\mathbf{P}[\Pi \succeq v \mid Y] = \mathbf{E}\left[\prod_{u \in \mathscr{A} \cap \prec v} \pi(u, X)\right] \cdot \prod_{u \in \mathscr{B} \cap \prec v} \pi(u, Y). \tag{2.12}$$

The claimed relationship (2.8) is immediate from (2.9)–(2.12). ∎

2.7. Information cost. Fix a private-coin communication protocol π and a probability distribution μ on the input space of π. Let X and Y be random variables with joint distribution μ, corresponding to Alice and Bob's inputs, and let Π be the transcript of π on inputs X and Y. The *internal information cost of* π *with respect to* μ is defined as

$$\mathrm{IC}_\mu(\pi) = I(\Pi; X \mid Y) + I(\Pi; Y \mid X).$$

Introduced by Barak et al. [4], this quantity measures the amount of information that Alice and Bob learn on average about each other's inputs by executing the protocol. A closely related notion is the *external information cost*, defined for π with respect to μ as

$$\mathrm{IC}^*_\mu(\pi) = I(\Pi; XY).$$

This alternate quantity was introduced several years earlier by Chakrabarti et al. [11], with implicit uses in several other works. External information cost measures the amount of information that the protocol transcript reveals to an outside observer about the inputs X and Y. The chain rule for mutual information implies that $\mathrm{IC}_\mu(\pi) \leqslant 2\,\mathrm{IC}^*_\mu(\pi)$. The following sharper result was proved by Barak et al. [4].

Theorem 2.10 (Barak et al.). *For any private-coin protocol* π *and any probability distribution* μ,

$$\mathrm{IC}_\mu(\pi) \leqslant \mathrm{IC}^*_\mu(\pi),$$

with equality for product distributions.

For general (nonproduct) distributions μ, the gap between the internal and external information cost can be arbitrary [4]. The internal and external information cost of a *public-coin* protocol π are defined in a natural way by conditioning on the shared random string R. Formally,

$$\mathrm{IC}_\mu(\pi) = I(\Pi; X \mid RY) + I(\Pi; Y \mid RX),$$
$$\mathrm{IC}^*_\mu(\pi) = I(\Pi; XY \mid R).$$

Put another way, the information cost of a public-coin protocol is the average information cost of the associated private-coin protocols.

2.8. Local view of information cost. External information cost admits a useful alternate characterization, based on the chain rule for mutual information. As before, fix a private-coin communication protocol π with input space $\mathscr{X} \times \mathscr{Y}$ and consider a probability distribution μ on $\mathscr{X} \times \mathscr{Y}$. Let X and Y be random variables with joint distribution μ, and let Π be the transcript of π on input X, Y. For $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, define $P, P_x, P_y, P_{x,y}$ to be the probability distributions that govern the random variables

$$\begin{aligned}&\Pi,\\&\Pi \mid X = x,\\&\Pi \mid Y = y,\\&\Pi \mid X = x, Y = y,\end{aligned}$$

respectively. Thus, $P, P_x, P_y, P_{x,y}$ are probability distributions on the leaves of the protocol tree. For a leaf or internal vertex v, recall from Section 2.5 that the shorthands $P(v), P_x(v), P_y(v), P_{x,y}(v)$ refer to the probability of reaching a leaf in the subtree of v. Similarly, $P(v \mid u), P_x(v \mid u), P_y(v \mid u), P_{x,y}(v \mid u)$ refer to the probability of reaching a leaf in the subtree of v conditioned on reaching a leaf in the subtree of u. For any vertex v of the protocol tree and inputs $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, define

$$\begin{aligned}\mathbb{D}_x^{\pi,\mu}(v) &= \begin{cases}\mathrm{KL}(P_x(v0 \mid v) \parallel P(v0 \mid v)) & \text{if } v \in \mathscr{A},\\ 0 & \text{otherwise}\end{cases}\\ &= \begin{cases}\mathrm{KL}(P_{x,y}(v0 \mid v) \parallel P(v0 \mid v)) & \text{if } v \in \mathscr{A},\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

and analogously

$$\begin{aligned}\mathbb{D}_y^{\pi,\mu}(v) &= \begin{cases}\mathrm{KL}(P_y(v0 \mid v) \parallel P(v0 \mid v)) & \text{if } v \in \mathscr{B},\\ 0 & \text{otherwise}\end{cases}\\ &= \begin{cases}\mathrm{KL}(P_{x,y}(v0 \mid v) \parallel P(v0 \mid v)) & \text{if } v \in \mathscr{B},\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

where as usual $\mathscr{A}$ and $\mathscr{B}$ stand for the sets of vertices owned by Alice and Bob, respectively. These quantities, introduced by Barak et al. [4], measure the information revealed about the protocol inputs locally due to the bit transmission at vertex v. Observe that for an internal vertex v, at most one of the quantities $\mathbb{D}_x^{\pi,\mu}(v), \mathbb{D}_y^{\pi,\mu}(v)$ is nonzero, whereas for every leaf vertex v, both quantities are zero. Define

$$\begin{aligned}\mathbb{D}_{x,y}^{\pi,\mu}(v) &= \mathbb{D}_x^{\pi,\mu}(v) + \mathbb{D}_y^{\pi,\mu}(v)\\ &= \mathrm{KL}(P_{x,y}(v0 \mid v) \parallel P(v0 \mid v)). \qquad (2.13)\end{aligned}$$

For $\mathscr{S} \subseteq \mathscr{V}(\pi)$, we abbreviate

$$\begin{aligned}\mathbb{D}_x^{\pi,\mu}(\mathscr{S}) &= \sum_{v \in \mathscr{S}} \mathbb{D}_x^{\pi,\mu}(v),\\ \mathbb{D}_y^{\pi,\mu}(\mathscr{S}) &= \sum_{v \in \mathscr{S}} \mathbb{D}_y^{\pi,\mu}(v),\\ \mathbb{D}_{x,y}^{\pi,\mu}(\mathscr{S}) &= \sum_{v \in \mathscr{S}} \mathbb{D}_{x,y}^{\pi,\mu}(v).\end{aligned}$$

We are now in a position to state the alternate characterization of external information cost, due to Barak et al. [4].

THEOREM 2.11. *For any private-coin protocol π and distribution μ,*

$$\mathrm{IC}^*_\mu(\pi) = \mathbf{E}\,\mathbb{D}^{\pi;\mu}_{X,Y}(\prec\Pi),$$

where X and Y are random variables with joint distribution μ, and Π is the protocol transcript of π on input X,Y.

The lower bound in this theorem was proved in [4].

Proof of Theorem 2.11. Recall that P and $P_{x,y}$ stand for the probability distributions that govern the random variables Π and $\Pi \mid X = x, Y = y$, respectively. Theorem 2.7 guarantees that

$$\mathrm{KL}(P_{x,y} \parallel P) = \mathbf{E}[\mathbb{D}^{\pi;\mu}_{X,Y}(\prec\Pi) \mid X = x, Y = y],$$

whence

$$\mathbf{E}\,\mathrm{KL}(P_{X,Y} \parallel P) = \mathbf{E}\,\mathbb{D}^{\pi;\mu}_{X,Y}(\prec\Pi).$$

The left-hand side of this equation is by definition $I(\Pi; XY) = \mathrm{IC}^*_\mu(\pi)$. ∎

The following related result is inspired by Fact 2.3.

THEOREM 2.12. *For every private-coin protocol π and distributions μ and $\tilde\mu$,*

$$\mathbf{E}\,\mathbb{D}^{\pi;\mu}_{X,Y}(\prec\Pi) \leqslant \mathbf{E}\,\mathbb{D}^{\pi;\tilde\mu}_{X,Y}(\prec\Pi), \tag{2.14}$$

where X and Y are random variables with joint distribution μ, and Π is the protocol transcript of π on input X,Y.

Proof. As usual, let $P_{x,y}$ stand for the probability distribution that governs the random variable $\Pi \mid X = x, Y = y$. Then

$$\begin{aligned}
\mathbf{E}\,\mathbb{D}^{\pi;\mu}_{X,Y}(\prec\Pi) &= \sum_{x,y} \mu(x,y)\,\mathbf{E}[\mathbb{D}^{\pi;\mu}_{X,Y}(\prec\Pi) \mid X = x, Y = y] \\
&= \sum_{x,y} \mu(x,y)\,\mathrm{KL}\left(P_{x,y} \,\middle\|\, \sum_{x',y'} \mu(x',y') P_{x',y'}\right) \\
&\leqslant \sum_{x,y} \mu(x,y)\,\mathrm{KL}\left(P_{x,y} \,\middle\|\, \sum_{x',y'} \tilde\mu(x',y') P_{x',y'}\right) \\
&= \sum_{x,y} \mu(x,y)\,\mathbf{E}[\mathbb{D}^{\pi;\tilde\mu}_{X,Y}(\prec\Pi) \mid X = x, Y = y] \\
&= \mathbf{E}\,\mathbb{D}^{\pi;\tilde\mu}_{X,Y}(\prec\Pi),
\end{aligned}$$

where the second and fourth steps use Theorem 2.7, and the third step is valid by Fact 2.3. ∎

2.9. Protocol simulation. Simulation of a given protocol π on given inputs x, y by another protocol π' involves running π' on x, y and interpreting its transcript as a transcript of π. Alice and Bob may base their interpretations on their respective inputs x and y, potentially arriving at distinct conclusions. In an accurate simulation, we require that their interpretations almost always agree and approximately follow the distribution of π's transcript on the input in question. Naturally, we are interested in simulating π efficiently, with communication cost as close to $\mathrm{IC}_\mu(\pi)$ as possible. This challenge is referred to as *protocol compression*.

Formally, let π be a private- or public-coin communication protocol with input space $\mathscr{X} \times \mathscr{Y}$, and let μ be a probability distribution on $\mathscr{X} \times \mathscr{Y}$. We say that π' *simulates* π *with error* ϵ *with respect to* μ, denoted

$$\pi' \hookrightarrow_{\mu,\epsilon} \pi,$$

if there exist a pair of functions $a\colon \{0,1\}^* \to \{0,1\}^*$ and $b\colon \{0,1\}^* \to \{0,1\}^*$ such that $\mathrm{TV}((X,Y,\Pi,\Pi),(X,Y,a(X,\Pi'),b(Y,\Pi'))) \leqslant \epsilon$, where X and Y are random variables with joint distribution μ, and Π and Π' are the transcripts of π and π', respectively, on input X, Y. We remind the reader that for public-coin protocols, the protocol transcript always includes the shared random string. The notion of protocol simulation is transitive in the following sense.

Theorem 2.13. *Let π, π', π'' be private- or public-coin protocols with input space $\mathscr{X} \times \mathscr{Y}$. Let μ be a probability distribution on $\mathscr{X} \times \mathscr{Y}$. Assume that*

$$\pi'' \hookrightarrow_{\mu,\epsilon} \pi',$$
$$\pi' \hookrightarrow_{\mu,\delta} \pi.$$

Then

$$\pi'' \hookrightarrow_{\mu,\epsilon+\delta} \pi.$$

Proof. Immediate from the triangle inequality for statistical distance. ∎

The following well-known result shows that any public-coin protocol can be faithfully simulated by a private-coin protocol with no increase in information cost. Thus, private- and public-coin protocols can be regarded as equivalent notions from the point of view of information cost.

Theorem 2.14 (Folklore). *Let π be a public-coin protocol with input space $\mathscr{X} \times \mathscr{Y}$. Let μ be a probability distribution on $\mathscr{X} \times \mathscr{Y}$. Then there is a private-coin protocol π' such that*

$$\pi' \hookrightarrow_{\mu,0} \pi, \tag{2.15}$$
$$\mathrm{IC}_\mu(\pi') = \mathrm{IC}_\mu(\pi), \tag{2.16}$$
$$\mathrm{IC}^*_\mu(\pi') = \mathrm{IC}^*_\mu(\pi). \tag{2.17}$$

Proof. Recall that Alice and Bob execute π by sampling a bit string R from a shared source of random bits and executing the corresponding private-coin protocol π_R. To simulate this behavior with a private-coin protocol π', Alice will privately sample a bit string R to be used as shared randomness and send it to Bob, at which point they will run the private-coin protocol π_R as before. On any given input, the transcript of π' has the same distribution as the transcript of π, settling (2.15).

It remains to analyze the information cost of π'. Let X and Y be random variables with joint distribution μ, and let Π_R be the transcript of π_R on input X, Y. Then

$$\begin{aligned}\mathrm{IC}^*_\mu(\pi') &= I(R\Pi_R; XY)\\ &= I(R\Pi_R; XY \mid R) + I(R; XY)\\ &= I(R\Pi_R; XY \mid R)\\ &= \mathrm{IC}^*_\mu(\pi)\end{aligned}$$

and

$$\begin{aligned}\mathrm{IC}_\mu(\pi') &= I(R\Pi_R; X \mid Y) + I(R\Pi_R; Y \mid X)\\ &= I(R\Pi_R; X \mid RY) + I(R\Pi_R; Y \mid RX) + I(R; X \mid Y) + I(R; Y \mid X)\\ &= I(R\Pi_R; X \mid RY) + I(R\Pi_R; Y \mid RX)\\ &= \mathrm{IC}_\mu(\pi),\end{aligned}$$

where the third step in both derivations uses the independence of R and XY. ∎

A private-coin protocol $\pi\colon (\mathscr{A}\times\mathscr{X})\cup(\mathscr{B}\times\mathscr{Y}) \to [0,1]$ is β-*balanced* if the range of π is contained in $[\frac{1}{2}-\beta, \frac{1}{2}+\beta]$. The following result, obtained by Barak et al. [4] and revisited recently by Kol [21], shows that any protocol can be simulated by a β-balanced protocol at the expense of an infinitesimal increase in information cost.

THEOREM 2.15 (Barak et al., Kol). *Let π be a private-coin protocol with input space $\mathscr{X}\times\mathscr{Y}$. Let μ be a probability distribution on $\mathscr{X}\times\mathscr{Y}$. Then for every $\beta > 0$ and $\epsilon > 0$, there exists a private-coin β-balanced protocol π' such that*

$$\begin{aligned}&\pi' \hookrightarrow_{\mu,\epsilon} \pi,\\ &\mathrm{IC}_\mu(\pi') \leqslant \mathrm{IC}_\mu(\pi) + \epsilon,\\ &\mathrm{IC}^*_\mu(\pi') \leqslant \mathrm{IC}^*_\mu(\pi) + \epsilon.\end{aligned}$$

For the reader's convenience, we provide a proof of this result in Appendix B. Collectively, Theorems 2.13–2.15 reduce the protocol compression problem to private-coin β-balanced protocols, where $\beta > 0$ can be taken arbitrarily small relative to the protocol's information cost.

Another useful tool in protocol compression is *correlated sampling*, which makes it possible for two parties to sample a random element according to a probability distribution known to only one of them. The following theorem, due to Braverman and Rao [7, Section IV], gives an efficient communication protocol for correlated sampling in the public-coin randomized model.

THEOREM 2.16 (Braverman and Rao). *Fix a finite set $\mathscr{X}$ and an error parameter $0 < \epsilon < 1/2$. There is a two-party public-coin communication protocol with the following properties.*

(i) *Alice and Bob receive as input probability distributions μ and $\tilde{\mu}$, respectively, on $\mathscr{X}$.*

(ii) *At the end of the protocol, Alice and Bob privately generate elements $Z \in \mathscr{X}$ and $\tilde{Z} \in \mathscr{X}$, respectively, such that $Z \sim \mu$ and $\mathbf{P}[Z = \tilde{Z}] > 1 - \epsilon$.*

(iii) *The communication cost on input pair $(\mu, \tilde{\mu})$ is $O(\log \frac{1}{\epsilon}) + C$, where C is a nonnegative random variable with expected value $O(\mathrm{KL}(\mu \parallel \tilde{\mu}))$.*

3. Cuts, floors, and ceilings

Cuts are special families of binary strings that arise in the analysis of binary trees. This section defines cuts, establishes their basic properties, and examines natural relations and operations on cuts.

3.1. Cuts defined.

A *cut* in a binary tree is any subset $\mathscr{C}$ of the tree's vertices such that $|\mathscr{C} \cap {\preceq}v| = 1$ for every leaf v. In other words, a cut $\mathscr{C}$ is a set that intersects any root-to-leaf path in exactly one vertex. Since we identify tree vertices with binary strings, we view cuts as subsets of $\{0,1\}^*$. A frequently used fact in our proofs is that for any cut $\mathscr{C}$, the restriction of the binary tree to ${\preceq}\mathscr{C}$ is again a binary tree, namely, a truncation of the original. The following two propositions settle basic properties of cuts.

PROPOSITION 3.1. *Let $\mathscr{C}$ be a cut in a given binary tree. Then any two distinct vertices in $\mathscr{C}$ are incomparable.*

Proof. A pair of vertices are comparable if and only if there is a root-to-leaf path that passes through both of them. If $\mathscr{C}$ contained a pair of distinct vertices that were comparable, then some root-to-leaf path would intersect $\mathscr{C}$ in more than one vertex, violating the cut property. ∎

PROPOSITION 3.2. *Let $\mathscr{C}$ be a cut in a given binary tree. Then ${\prec}\mathscr{C}$, $\mathscr{C}$, and ${\succ}\mathscr{C}$ are pairwise disjoint sets whose union is $\{0,1\}^*$.*

Proof. If any two of the sets ${\prec}\mathscr{C}, \mathscr{C}, {\succ}\mathscr{C}$ had nonempty intersection, then $\mathscr{C}$ would contain a pair of distinct vertices that are comparable, contradicting Proposition 3.1.

It remains to show that the union ${\prec}\mathscr{C} \cup \mathscr{C} \cup {\succ}\mathscr{C}$ contains every binary string. Fix $v \in \{0,1\}^*$ arbitrarily. By extending the tree if necessary, we can view v as a tree vertex. Now consider any root-to-leaf path that passes through v. By the cut property, the path intersects $\mathscr{C}$ in some vertex u, which by definition is comparable to v. Put another way, v is contained in at least one of the sets ${\prec}\mathscr{C}, \mathscr{C}, {\succ}\mathscr{C}$. ∎

3.2. Relations on cuts. We define the relations $\prec, \succ, \preceq, \succeq$ on cuts by

$$\begin{aligned}
\mathscr{C} \prec \mathscr{D} \quad &\Leftrightarrow \quad \mathscr{C} \subseteq {\prec}\mathscr{D},\\
\mathscr{C} \succ \mathscr{D} \quad &\Leftrightarrow \quad \mathscr{C} \subseteq {\succ}\mathscr{D},\\
\mathscr{C} \preceq \mathscr{D} \quad &\Leftrightarrow \quad \mathscr{C} \subseteq {\preceq}\mathscr{D},\\
\mathscr{C} \succeq \mathscr{D} \quad &\Leftrightarrow \quad \mathscr{C} \subseteq {\succeq}\mathscr{D}.
\end{aligned}$$

Each of the relations $\prec, \succ, \preceq, \succeq$ is clearly transitive, a fact that we will often use in our proofs without explicit mention.

Proposition 3.3. *Let $\mathscr{C}$ and $\mathscr{D}$ be cuts in a given binary tree. Then*

$$\mathscr{C} \prec \mathscr{D} \quad \Leftrightarrow \quad \mathscr{D} \succ \mathscr{C}, \tag{3.1}$$

$$\mathscr{C} \preceq \mathscr{D} \quad \Leftrightarrow \quad \mathscr{D} \succeq \mathscr{C}. \tag{3.2}$$

Proof. Assume that $\mathscr{C} \prec \mathscr{D}$. Then $(\mathscr{D} \cap {\preceq}\mathscr{C}) \subseteq {\preceq}\mathscr{C} \subseteq {\prec}\mathscr{D}$. Thus, $\mathscr{D} \cap {\preceq}\mathscr{C} \neq \varnothing$ would imply that $\mathscr{D}$ contains a pair of distinct vertices that are comparable, contradicting Proposition 3.1. We conclude that $\mathscr{D} \cap {\preceq}\mathscr{C} = \varnothing$, which in view of Proposition 3.2 forces $\mathscr{D} \subseteq {\succ}\mathscr{C}$. Summarizing, $\mathscr{C} \prec \mathscr{D} \implies \mathscr{D} \succ \mathscr{C}$. One similarly proves $\mathscr{C} \preceq \mathscr{D} \implies \mathscr{D} \succeq \mathscr{C}$, as well as the converses of these two implications. □

For a leaf v and cuts $\mathscr{C}$ and $\mathscr{D}$ in a given binary tree, we write $\mathscr{C} \prec \mathscr{D} \pmod{v}$ to mean that $(\mathscr{C} \cap {\preceq}v) \prec (\mathscr{D} \cap {\preceq}v)$, i.e., the cut $\mathscr{C}$ precedes the cut $\mathscr{D}$ along the root-to-leaf path ${\preceq}v$. We analogously define

$$\begin{aligned}
&\mathscr{C} \preceq \mathscr{D} \pmod{v},\\
&\mathscr{C} = \mathscr{D} \pmod{v},\\
&\mathscr{C} \succeq \mathscr{D} \pmod{v},\\
&\mathscr{C} \succ \mathscr{D} \pmod{v}.
\end{aligned}$$

For a vertex v and a cut $\mathscr{C}$, we adopt the abbreviations

$$\begin{aligned}
v \prec \mathscr{C} \quad &\Leftrightarrow \quad v \in {\prec}\mathscr{C},\\
v \succ \mathscr{C} \quad &\Leftrightarrow \quad v \in {\succ}\mathscr{C},\\
v \preceq \mathscr{C} \quad &\Leftrightarrow \quad v \in {\preceq}\mathscr{C},\\
v \succeq \mathscr{C} \quad &\Leftrightarrow \quad v \in {\succeq}\mathscr{C}.
\end{aligned}$$

Proposition 3.2 ensures that for every v and $\mathscr{C}$, precisely one of the following conditions holds: $v \prec \mathscr{C}$, $v \in \mathscr{C}$, $v \succ \mathscr{C}$.

3.3. Floors and ceilings. A set $\mathscr{S} \subseteq \{0,1\}^*$ is *downward closed* if $\mathscr{S} = {\preceq}\mathscr{S}$. Analogously, $\mathscr{S}$ is *upward closed* if $\mathscr{S} = {\succeq}\mathscr{S}$. Observe that for any $\mathscr{S} \subseteq \{0,1\}^*$, the sets ${\preceq}\mathscr{S}$ and ${\prec}\mathscr{S}$ are downward closed, whereas ${\succeq}\mathscr{S}$ and ${\succ}\mathscr{S}$ are upward closed. For a set $\mathscr{S} \subseteq \{0,1\}^*$, we let $\lfloor \mathscr{S} \rfloor$ and $\lceil \mathscr{S} \rceil$ denote the subset of minimal elements of $\mathscr{S}$ and the subset of maximal elements of $\mathscr{S}$, respectively, relative to the $\prec$ ordering:

$$\begin{aligned}\lfloor \mathscr{S} \rfloor &= \{s \in \mathscr{S} : \mathscr{S} \cap {\prec} s = \varnothing\},\\ \lceil \mathscr{S} \rceil &= \{s \in \mathscr{S} : \mathscr{S} \cap {\succ} s = \varnothing\}.\end{aligned}$$

We emphasize that the floor and ceiling operations are defined for arbitrary sets $\mathscr{S}$ of binary strings and not just for cuts. A moment's reflection shows that for any $\mathscr{S} \subseteq \{0,1\}^*$,

$$\lfloor \mathscr{S} \rfloor = \mathscr{S} \setminus {\succ}\mathscr{S}, \tag{3.3}$$

$$\lceil \mathscr{S} \rceil = \mathscr{S} \setminus {\prec}\mathscr{S}. \tag{3.4}$$

For $\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \subseteq \{0,1\}^*$, we abbreviate

$$\begin{aligned}\lfloor \mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \rfloor &= \lfloor \mathscr{S}_1 \cup \mathscr{S}_2 \cup \ldots \cup \mathscr{S}_k \rfloor,\\ \lceil \mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \rceil &= \lceil \mathscr{S}_1 \cup \mathscr{S}_2 \cup \ldots \cup \mathscr{S}_k \rceil.\end{aligned}$$

By definition, $\lfloor \mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \rfloor$ and $\lceil \mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \rceil$ are invariant under permutations of the given k sets. It is also straightforward to verify the alternate representations

$$\lfloor \mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \rfloor = \lfloor \mathscr{S}_1, \lfloor \mathscr{S}_2, \lfloor \ldots, \lfloor \mathscr{S}_{k-1}, \mathscr{S}_k \rfloor \ldots \rfloor \rfloor \rfloor, \tag{3.5}$$

$$\lceil \mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_k \rceil = \lceil \mathscr{S}_1, \lceil \mathscr{S}_2, \lceil \ldots, \lceil \mathscr{S}_{k-1}, \mathscr{S}_k \rceil \ldots \rceil \rceil \rceil. \tag{3.6}$$

We now show that cuts are closed under the floor and ceiling operations.

PROPOSITION 3.4. *If $\mathscr{C} \subseteq \{0,1\}^*$ is a cut in a given binary tree, then so is $\lfloor \mathscr{C}, \mathscr{S} \rfloor$ for any $\mathscr{S} \subseteq \{0,1\}^*$.*

Proof. Let v be a leaf. Then $|\mathscr{C} \cap {\preceq} v| \geqslant 1$ by the cut property, and therefore $|\lfloor \mathscr{C}, \mathscr{S} \rfloor \cap {\preceq} v| \geqslant 1$ by the downward closure of ${\preceq} v$. For the matching upper bound, any two vertices in ${\preceq} v$ are comparable and therefore $\lfloor \mathscr{C}, \mathscr{S} \rfloor$ contains at most one of them. ∎

Proposition 3.4 has no direct counterpart for the ceiling operator. For example, the ceiling of $\mathscr{S} = \{0\}$ and of the cut $\mathscr{C} = \{\varepsilon\}$ is $\lceil \mathscr{C}, \mathscr{S} \rceil = \{0\}$, which is not a cut. However, if *both* of the original sets are cuts, then their ceiling is also a cut.

PROPOSITION 3.5. *If $\mathscr{C}_1, \mathscr{C}_2 \subseteq \{0,1\}^*$ are cuts in a given binary tree, then so is $\lceil \mathscr{C}_1, \mathscr{C}_2 \rceil$.*

Proof. Let v be a leaf. By the cut property, there is a pair of vertices $v_1 \in \mathscr{C}_1 \cap {\preceq} v$ and $v_2 \in \mathscr{C}_2 \cap {\preceq} v$. Proposition 3.1 implies that v_1 has no proper descendant in $\mathscr{C}_1$, and likewise v_2 has no proper descendant in $\mathscr{C}_2$. This means that the vertex $\max\{v_1, v_2\} \in \mathscr{C}_1 \cup \mathscr{C}_2$ has no proper descendant in $\mathscr{C}_1 \cup \mathscr{C}_2$ and is therefore an element of $\lceil \mathscr{C}_1, \mathscr{C}_2 \rceil$. For the matching upper bound, any two vertices in ${\preceq} v$ are comparable and therefore $\lceil \mathscr{C}_1, \mathscr{C}_2 \rceil$ contains at most one of them. ∎

Propositions 3.4 and 3.5 generalize to three or more cuts, using induction and the alternate representations (3.5) and (3.6). We conclude this section with useful explicit formulas for the floor and ceiling of a pair of cuts.

Proposition 3.6. *Let $\mathscr{C}$ and $\mathscr{D}$ be cuts in a given binary tree. Then*

$$\begin{aligned}\lfloor \mathscr{C},\mathscr{D}\rfloor &= (\mathscr{C}\cap{\preceq}\mathscr{D})\cup(\mathscr{D}\cap{\preceq}\mathscr{C}),\\ \lceil \mathscr{C},\mathscr{D}\rceil &= (\mathscr{C}\cap{\succeq}\mathscr{D})\cup(\mathscr{D}\cap{\succeq}\mathscr{C}).\end{aligned}$$

Proof. We have

$$\begin{aligned}\lfloor \mathscr{C},\mathscr{D}\rfloor &= (\mathscr{C}\cup\mathscr{D})\setminus({\succ}\mathscr{C}\cup{\succ}\mathscr{D})\\ &= (\mathscr{C}\cup\mathscr{D})\cap{\preceq}\mathscr{C}\cap{\preceq}\mathscr{D}\\ &= (\mathscr{C}\cap{\preceq}\mathscr{C}\cap{\preceq}\mathscr{D})\cup(\mathscr{D}\cap{\preceq}\mathscr{C}\cap{\preceq}\mathscr{D})\\ &= (\mathscr{C}\cap{\preceq}\mathscr{D})\cup(\mathscr{D}\cap{\preceq}\mathscr{C}),\end{aligned}$$

where the first two steps use (3.3) and Proposition 3.2, respectively. Analogously,

$$\begin{aligned}\lceil \mathscr{C},\mathscr{D}\rceil &= (\mathscr{C}\cup\mathscr{D})\setminus({\prec}\mathscr{C}\cup{\prec}\mathscr{D})\\ &= (\mathscr{C}\cup\mathscr{D})\cap{\succeq}\mathscr{C}\cap{\succeq}\mathscr{D}\\ &= (\mathscr{C}\cap{\succeq}\mathscr{C}\cap{\succeq}\mathscr{D})\cup(\mathscr{D}\cap{\succeq}\mathscr{C}\cap{\succeq}\mathscr{D})\\ &= (\mathscr{C}\cap{\succeq}\mathscr{D})\cup(\mathscr{D}\cap{\succeq}\mathscr{C}),\end{aligned}$$

where the first two steps use (3.4) and Proposition 3.2, respectively. ∎

4. Partial simulation

Let π be a given communication protocol with information cost I under a product distribution μ. Recall that the goal of this paper is to construct a public-coin randomized protocol that accurately simulates π with respect to μ and has communication cost $O(I\log^2 I)$. In keeping with now-standard practice, we start by developing a public-coin randomized procedure that simulates a nontrivial initial portion of the protocol π. The complete simulation, analyzed in a later section, will involve repeated execution of this partial procedure until the communication allotment is reached.

Theorem 4.1 (Partial simulation, $\sigma_{\pi,\mu,\epsilon}$). *Let $0<\epsilon<1/2$ be given. For $\beta=\beta(\epsilon)>0$ sufficiently small, fix any β-balanced private-coin protocol π with input space $\mathscr{X}\times\mathscr{Y}$, and any product distribution μ on $\mathscr{X}\times\mathscr{Y}$. Then there is a public-coin randomized protocol $\sigma_{\pi,\mu,\epsilon}$ with input space $\mathscr{X}\times\mathscr{Y}$ whose execution allows Alice and Bob to agree on a vertex of the protocol tree for π, subject to the following* *properties:*

$$
\begin{aligned}
&\sum_{w \preceq v} \frac{\mathbf{P}[W = w \mid X,Y]}{\mathbf{P}[\Pi \succeq w \mid X,Y]} \leqslant 1+\epsilon \qquad \forall v \in \mathscr{V}(\pi) && \textbf{(accuracy)}\\
&\mathbf{P}[W \in \mathscr{L}(\pi)] + \log\left(\frac{1}{\epsilon}\right) \mathbf{E}\,\mathbb{D}^{\pi,\mu}_{X,Y}(\prec W) \geqslant \frac{1}{c} && \textbf{(progress)}\\
&\left.\begin{aligned}
&C \leqslant C' + C'' + c\log\frac{1}{\epsilon}\\
&\mathbf{E}\,C' \leqslant c(\mathbf{E}\,\mathbb{D}^{\pi,\mu}_{X,Y}(\prec W) + \epsilon\,\mathbf{E}\,\mathbb{D}^{\pi,\mu}_{X,Y}(\prec \Pi))\\
&\mathbf{P}[C'' > 0] \leqslant \epsilon
\end{aligned}\right\} && \textbf{(cost)}
\end{aligned}
$$

where

(i) X, Y *are random variables with joint distribution* μ;

(ii) Π *is the transcript of* π *on input* X, Y;

(iii) $W \in \mathscr{V}(\pi)$ *is Alice and Bob's agreed-upon vertex after executing* $\sigma_{\pi,\mu,\epsilon}$ *on input* (X, Y), *and* $C \in \mathbb{N}$ *is the communication cost of that execution*;

(iv) $C', C'' \in \mathbb{N}$ *are auxiliary random variables*;

(v) W, C, C', C'' *are completely determined by the transcript of* $\sigma_{\pi,\mu,\epsilon}$;

(vi) $c > 1$ *is an absolute constant.*

The remainder of this section is devoted to the proof of Theorem 4.1. We have structured the proof around nine key milestones, corresponding to Sections 4.1–4.9 below.

4.1. Probability space and parameter list. We will assume that the protocol tree for π has more than one vertex, the theorem being trivial otherwise. Recall from the theorem statement that X, Y is a pair of inputs distributed according to μ, and Π is the protocol transcript of π on input X, Y. For a pair of inputs $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, consider the following familiar functions on the vertices of the protocol tree for π:

$$
\begin{aligned}
P(v) &= \mathbf{P}[\Pi \succeq v], && (4.1)\\
P_x(v) &= \mathbf{P}[\Pi \succeq v \mid X = x], && (4.2)\\
P_y(v) &= \mathbf{P}[\Pi \succeq v \mid Y = y], && (4.3)\\
P_{x,y}(v) &= \mathbf{P}[\Pi \succeq v \mid X = x, Y = y]. && (4.4)
\end{aligned}
$$

We use analogous notation for conditional probabilities:

$$\left.\begin{aligned} P(v \mid u) &= \frac{P(v)}{P(u)} = \mathbf{P}[\Pi \succeq v \mid \Pi \succeq u], \\ P_x(v \mid u) &= \frac{P_x(v)}{P_x(u)} = \mathbf{P}[\Pi \succeq v \mid \Pi \succeq u, X = x], \\ P_y(v \mid u) &= \frac{P_y(v)}{P_y(u)} = \mathbf{P}[\Pi \succeq v \mid \Pi \succeq u, Y = y], \\ P_{x,y}(v \mid u) &= \frac{P_{x,y}(v)}{P_{x,y}(u)} = \mathbf{P}[\Pi \succeq v \mid \Pi \succeq u, X = x, Y = y] \end{aligned}\right\} \qquad u \preceq v.$$

For a set $\mathscr{C}$ of vertices, we abbreviate

$$P(\mathscr{C}) = \sum_{v \in \mathscr{C}} P(v),$$

and likewise for $P_x(\mathscr{C}), P_y(\mathscr{C}), P_{x,y}(\mathscr{C})$. By definition, each of the functions in (4.1)–(4.4) is a probability distribution when restricted to the leaves $\mathscr{L}(\pi)$ of the protocol tree. More generally, (4.1)–(4.4) are probability distributions when restricted to any cut of the protocol tree:

$$P(\mathscr{C}) = P_x(\mathscr{C}) = P_y(\mathscr{C}) = P_{x,y}(\mathscr{C}) = 1, \qquad \mathscr{C} \text{ a cut.} \tag{4.5}$$

By Fact 2.9,

$$P(v)P_{x,y}(v) = P_x(v)P_y(v), \qquad v \in \mathscr{V}(\pi), \tag{4.6}$$

whence

$$P(v \mid u)P_{x,y}(v \mid u) = P_x(v \mid u)P_y(v \mid u), \qquad u \preceq v \in \mathscr{V}(\pi). \tag{4.7}$$

Throughout the proof, we abbreviate $\mathbb{D}_x^{\pi,\mu}, \mathbb{D}_y^{\pi,\mu}, \mathbb{D}_{x,y}^{\pi,\mu}$ to $\mathbb{D}_x, \mathbb{D}_y, \mathbb{D}_{x,y}$, respectively. We assume without loss of generality that Alice and Bob's input sets $\mathscr{X}$ and $\mathscr{Y}$ are disjoint, which eliminates the possibility of conflicting interpretations for $\mathbb{D}_x$ and $\mathbb{D}_y$. Among the many applications of (4.7) are the following simplified expressions for $\mathbb{D}_x$ and $\mathbb{D}_y$.

Claim 4.2. *Let $x \in \mathscr{X}$ and $y \in \mathscr{Y}$. Then*

$$\begin{aligned} \mathbb{D}_x(v) &= \mathrm{KL}(P_x(v0 \mid v) \,\|\, P(v0 \mid v)) \qquad (4.8) \\ &= \mathrm{KL}(P_{x,y}(v0 \mid v) \,\|\, P_y(v0 \mid v)). \qquad (4.9) \end{aligned}$$

Analogously,

$$\begin{aligned} \mathbb{D}_y(v) &= \mathrm{KL}(P_y(v0 \mid v) \,\|\, P(v0 \mid v)) \qquad (4.10) \\ &= \mathrm{KL}(P_{x,y}(v0 \mid v) \,\|\, P_x(v0 \mid v)). \qquad (4.11) \end{aligned}$$

Proof. By symmetry, it suffices to prove (4.8) and (4.9). There are two cases to consider. If v is owned by Alice, then $P_x(v0 \mid v) = P_{x,y}(v0 \mid v)$, which in turn implies in view of (4.7) that $P_y(v0 \mid v) = P(v0 \mid v)$. In light of these two identities, (4.8) and (4.9) are both equivalent to $\mathbb{D}_x(v) = \mathrm{KL}(P_{x,y}(v0 \mid v) \parallel P(v0 \mid v))$, which is the defining equation at vertices v owned by Alice.

In the complementary case when v is owned by Bob, we have $P_y(v0 \mid v) = P_{x,y}(v0 \mid v)$, which in turn implies in view of (4.7) that $P_x(v0 \mid v) = P(v0 \mid v)$. In light of these two identities, (4.8) and (4.9) are both equivalent to $\mathbb{D}_x(v) = 0$, which is the defining equation at vertices v owned by Bob. □

Our proof has three positive parameters δ, Δ, r, whose precise values will be determined later in terms of ϵ. Their orders of magnitude are given by (4.83)–(4.85). We will ensure that

$$0 < \delta < \frac{1}{16}, \tag{4.12}$$

$$\Delta \geqslant 1. \tag{4.13}$$

4.2. Frontiers and randomized frontiers. By taking $\beta > 0$ sufficiently small, we can ensure that

$$\max_{p,q \in [\frac{1}{2}-\beta, \frac{1}{2}+\beta]} \mathrm{KL}(p \parallel q) < \delta.$$

Indeed, (2.4) makes it clear that $\beta = \Theta(\sqrt{\delta})$ is an adequate setting. We may therefore assume that

$$\left.\begin{array}{r}\mathbb{D}_{x,y}(v) \leqslant \delta\\ \mathbb{D}_x(v) \leqslant \delta\\ \mathbb{D}_y(v) \leqslant \delta\end{array}\right\} \qquad x \in \mathscr{X}, y \in \mathscr{Y}, v \in \mathscr{V}(\pi). \tag{4.14}$$

For $\theta > 0$, define

$$\mathscr{F}_{x,\theta} = \lfloor \{v \in \mathscr{V}(\pi) : \mathbb{D}_x(\prec v) \geqslant \theta\}, \mathscr{L}(\pi) \rfloor, \tag{4.15}$$

$$\mathscr{F}_{y,\theta} = \lfloor \{v \in \mathscr{V}(\pi) : \mathbb{D}_y(\prec v) \geqslant \theta\}, \mathscr{L}(\pi) \rfloor. \tag{4.16}$$

Following previous work on protocol compression [4, 21], we refer to these sets as *θ-frontiers*. We will be interested in θ-frontiers for both $0 < \theta \ll 1$ and $\theta \gg 1$. The notion of a θ-frontier has a natural randomized counterpart due to Kol [21], obtained by passing from specific inputs $x \in \mathscr{X}$ and $y \in \mathscr{Y}$ to random inputs X and Y:

$$\begin{aligned}\mathscr{R}_{\mathscr{X},\theta,\rho} &= \lfloor \{v \in \mathscr{V}(\pi) : \mathbf{P}[\mathbb{D}_X(\prec v) \geqslant \theta] \geqslant \rho\}, \mathscr{L}(\pi) \rfloor & (4.17)\\ &= \lfloor \{v \in \mathscr{V}(\pi) : \mathbf{P}[v \succeq \mathscr{F}_{X,\theta}] \geqslant \rho\} \rfloor, & (4.18)\\ \mathscr{R}_{\mathscr{Y},\theta,\rho} &= \lfloor \{v \in \mathscr{V}(\pi) : \mathbf{P}[\mathbb{D}_Y(\prec v) \geqslant \theta] \geqslant \rho\}, \mathscr{L}(\pi) \rfloor & (4.19)\\ &= \lfloor \{v \in \mathscr{V}(\pi) : \mathbf{P}[v \succeq \mathscr{F}_{Y,\theta}] \geqslant \rho\} \rfloor, & (4.20)\end{aligned}$$

where $\theta > 0$ and $0 < \rho \leqslant 1$. We abbreviate

$$\mathscr{R}_{\theta,\rho} = \lfloor \mathscr{R}_{\mathscr{X},\theta,\rho}, \mathscr{R}_{\mathscr{Y},\theta,\rho} \rfloor.$$

In the two claims that follow, we settle basic combinatorial and measure-theoretic properties of these sets.

Claim 4.3. *Let $\theta > 0$ and $0 < \rho \leqslant 1$. Then $\mathscr{F}_{x,\theta}$, $\mathscr{F}_{y,\theta}$, $\mathscr{R}_{\mathscr{X},\theta,\rho}$, $\mathscr{R}_{\mathscr{Y},\theta,\rho}$, $\mathscr{R}_{\theta,\rho}$ for all $x \in \mathscr{X}$ and $y \in \mathscr{Y}$ are nonempty and are cuts in the protocol tree for π. Moreover,*

$$\varepsilon \notin \mathscr{F}_{x,\theta}, \tag{4.21}$$
$$\varepsilon \notin \mathscr{F}_{y,\theta}, \tag{4.22}$$
$$\varepsilon \notin \mathscr{R}_{\mathscr{X},\theta,\rho}, \tag{4.23}$$
$$\varepsilon \notin \mathscr{R}_{\mathscr{Y},\theta,\rho}, \tag{4.24}$$
$$\varepsilon \notin \mathscr{R}_{\theta,\rho}. \tag{4.25}$$

Proof. The floor operator in (4.15)–(4.19) is applied to sets that contain the cut $\mathscr{L}(\pi)$. We conclude that $\mathscr{F}_{x,\theta}$, $\mathscr{F}_{y,\theta}$, $\mathscr{R}_{\mathscr{X},\theta,\rho}$, $\mathscr{R}_{\mathscr{Y},\theta,\rho}$ are nonempty and by Proposition 3.4 are cuts. This in turn makes $\mathscr{R}_{\theta,\rho}$ a cut, again by Proposition 3.4.

By hypothesis, the protocol tree for π has more than one vertex and therefore ε is not a leaf. As a result, our definitions imply (4.21)–(4.25) directly. ∎

Claim 4.4. *Let $\theta > 0$ and $0 \leqslant \rho_1 < \rho_2 \leqslant 1$. Then for any leaf $v \in \mathscr{L}(\pi)$,*

$$\mathbf{P}[\mathscr{R}_{\mathscr{X},\theta,\rho_1} \prec \mathscr{F}_{X,\theta} \prec \mathscr{R}_{\mathscr{X},\theta,\rho_2} \pmod{v}] < \rho_2 - \rho_1,$$
$$\mathbf{P}[\mathscr{R}_{\mathscr{Y},\theta,\rho_1} \prec \mathscr{F}_{Y,\theta} \prec \mathscr{R}_{\mathscr{Y},\theta,\rho_2} \pmod{v}] < \rho_2 - \rho_1.$$

More generally,

$$\mathbf{P}[\lfloor \mathscr{R}_{\mathscr{X},\theta,\rho_1}, \mathscr{C} \rfloor \prec \lfloor \mathscr{F}_{X,\theta}, \mathscr{C} \rfloor \prec \lfloor \mathscr{R}_{\mathscr{X},\theta,\rho_2}, \mathscr{C} \rfloor \pmod{v}] < \rho_2 - \rho_1,$$
$$\mathbf{P}[\lfloor \mathscr{R}_{\mathscr{Y},\theta,\rho_1}, \mathscr{C} \rfloor \prec \lfloor \mathscr{F}_{Y,\theta}, \mathscr{C} \rfloor \prec \lfloor \mathscr{R}_{\mathscr{Y},\theta,\rho_2}, \mathscr{C} \rfloor \pmod{v}] < \rho_2 - \rho_1$$

for any fixed cut $\mathscr{C}$.

Proof. The first two bounds are immediate from the definitions of $\mathscr{R}_{\mathscr{X},\theta,\rho}$ and $\mathscr{R}_{\mathscr{Y},\theta,\rho}$ for $\theta > 0$ and $0 < \rho \leqslant 1$. The other two bounds follow directly from the first two. ∎

Analogous to previous analyses [4, 21], we will need the fact that the ratios $P_{x,y}(v)/P_x(v)$ and $P_{x,y}(v)/P_y(v)$ behave reasonably for most vertices below the δ-frontiers.

Claim 4.5. *Let $\mathscr{S} \subseteq \{0,1\}^*$ be arbitrary. Then for any $x \in \mathscr{X}$ and $y \in \mathscr{Y}$,*

$$P_{x,y}\left(\left\{v \in \lfloor \mathscr{S}, \mathscr{F}_{y,\delta} \rfloor : \frac{c_0 P_{x,y}(v)}{P_x(v)} \geqslant 1\right\}\right) \leqslant \exp\left(-\frac{1}{\delta}\right), \tag{4.26}$$

$$P_{x}\left(\left\{v \in \lfloor \mathscr{S}, \mathscr{F}_{y,\delta} \rfloor : \frac{c_0 P_{x}(v)}{P_{x,y}(v)} \geqslant 1\right\}\right) \leqslant \exp\left(-\frac{1}{\delta}\right) \tag{4.27}$$

and analogously

$$P_{x,y}\left(\left\{v \in \lfloor \mathscr{S}, \mathscr{F}_{x,\delta} \rfloor : \frac{c_0 P_{x,y}(v)}{P_y(v)} \geqslant 1\right\}\right) \leqslant \exp\left(-\frac{1}{\delta}\right), \tag{4.28}$$

$$P_{y}\left(\left\{v \in \lfloor \mathscr{S}, \mathscr{F}_{x,\delta} \rfloor : \frac{c_0 P_{y}(v)}{P_{x,y}(v)} \geqslant 1\right\}\right) \leqslant \exp\left(-\frac{1}{\delta}\right), \tag{4.29}$$

where $0 < c_0 < 1$ is an absolute constant.

Proof. By symmetry, it suffices to prove (4.26) and (4.27). By Proposition 3.4 and Claim 4.3, the set $\lfloor \mathscr{S}, \mathscr{F}_{y,\delta} \rfloor$ is a cut. As a result, (4.5) allows us to regard $P_{x,y}$ and P_x as probability distributions on $\lfloor \mathscr{S}, \mathscr{F}_{y,\delta} \rfloor$. We have

$$\begin{aligned} \sum_{u : u \prec v} \mathrm{KL}(P_{x,y}(u0 \mid u) \parallel P_x(u0 \mid u)) &= \mathbb{D}_y(\prec v) \\ &< 2\delta, \qquad\qquad v \in \lfloor \mathscr{S}, \mathscr{F}_{y,\delta} \rfloor, \end{aligned} \tag{4.30}$$

where the first step follows from (4.11) and the second step uses (4.14) and the definition of $\mathscr{F}_{y,\delta}$. Finally, the balance property of π implies that

$$\frac{1}{3} \leqslant P_{x,y}(u0 \mid u) \leqslant \frac{2}{3}, \tag{4.31}$$

$$\frac{1}{3} \leqslant P_{x}(u0 \mid u) \leqslant \frac{2}{3} \tag{4.32}$$

for all internal vertices u. Now (4.26) and (4.27) follow immediately from (4.12), (4.30)–(4.32), and Theorem 2.8, applied to the protocol tree truncated at the cut $\lfloor \mathscr{S}, \mathscr{F}_{y,\delta} \rfloor$. ∎

4.3. Rounded frontiers and sampling cuts. Consider a "rounded" version of the cut $\mathscr{F}_{x,\theta}$, given by

$$\overline{\mathscr{F}_{x,\theta}} = \left\lfloor \bigcup_{i=1}^{2r} \mathscr{R}_{\mathscr{Y},\theta,\frac{i}{2r}} \cap \succeq \mathscr{F}_{x,\theta}, \quad \mathscr{L}(\pi) \right\rfloor. \tag{4.33}$$

Put another way, $\overline{\mathscr{F}_{x,\theta}}$ is obtained by collecting, for each vertex of $\mathscr{F}_{x,\theta}$, its descendants in

$$\mathscr{R}_{\mathscr{Y},\theta,\frac{1}{2r}} \cup \cdots \cup \mathscr{R}_{\mathscr{Y},\theta,\frac{i}{2r}} \cup \cdots \cup \mathscr{R}_{\mathscr{Y},\theta,\frac{2r}{2r}} \cup \mathscr{L}(\pi) \tag{4.34}$$

and taking the floor of the resulting collection of descendants. In that sense, $\overline{\mathscr{F}_{x,\theta}}$ is the result of rounding $\mathscr{F}_{x,\theta}$ up with respect to the cuts $\mathscr{R}_{\mathscr{Y},\theta,\frac{1}{2r}} \preceq \cdots \preceq \mathscr{R}_{\mathscr{Y},\theta,\frac{i}{2r}} \preceq \cdots \preceq \mathscr{R}_{\mathscr{Y},\theta,\frac{2r}{2r}} \preceq \mathscr{L}(\pi)$. We analogously define

$$\overline{\mathscr{F}_{y,\theta}} = \left\lfloor \bigcup_{i=1}^{2r} \mathscr{R}_{\mathscr{X},\theta,\frac{i}{2r}} \cap {\succeq}\mathscr{F}_{y,\theta}, \quad \mathscr{L}(\pi) \right\rfloor.$$

Claim 4.6. *Let $\theta > 0$. Then for all $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, the sets $\overline{\mathscr{F}_{x,\theta}}$ and $\overline{\mathscr{F}_{y,\theta}}$ are cuts in the protocol tree for π. Moreover,*

$$\overline{\mathscr{F}_{x,\theta}} \succeq \mathscr{F}_{x,\theta}, \tag{4.35}$$

$$\overline{\mathscr{F}_{y,\theta}} \succeq \mathscr{F}_{y,\theta}. \tag{4.36}$$

Proof. By symmetry, it suffices to consider $\overline{\mathscr{F}_{x,\theta}}$. The floor operator in (4.33) is applied to a set that contains the cut $\mathscr{L}(\pi)$, which makes $\overline{\mathscr{F}_{x,\theta}}$ a cut by Proposition 3.4. To verify (4.35), observe that each of the sets

$$\begin{aligned}
&\mathscr{R}_{\mathscr{Y},\theta,\frac{1}{2r}} \cap {\succeq}\mathscr{F}_{x,\theta},\\
&\vdots\\
&\mathscr{R}_{\mathscr{Y},\theta,\frac{i}{2r}} \cap {\succeq}\mathscr{F}_{x,\theta},\\
&\vdots\\
&\mathscr{R}_{\mathscr{Y},\theta,\frac{2r}{2r}} \cap {\succeq}\mathscr{F}_{x,\theta},\\
&\mathscr{L}(\pi)
\end{aligned}$$

is by definition contained in ${\succeq}\mathscr{F}_{x,\theta}$. The same containment must therefore apply to the union of these $2r+1$ sets, as well as to the floor of that union. □

In what follows, we abbreviate

$$\begin{aligned}
\mathscr{S}_x &= \lfloor \overline{\mathscr{F}_{x,\delta}}, \mathscr{F}_{x,\Delta}, \mathscr{R}_{\delta,1/2} \rfloor, && x \in \mathscr{X},\\
\mathscr{S}_y &= \lfloor \overline{\mathscr{F}_{y,\delta}}, \mathscr{F}_{y,\Delta}, \mathscr{R}_{\delta,1/2} \rfloor, && y \in \mathscr{Y},\\
\mathscr{S}_{x,y} &= \lceil \lfloor \mathscr{S}_x, \mathscr{F}_{y,\delta} \rfloor, \lfloor \mathscr{S}_y, \mathscr{F}_{x,\delta} \rfloor \rceil, && x \in \mathscr{X}, y \in \mathscr{Y}.
\end{aligned}$$

For reasons that will become clear shortly, we think of these sets as Alice's sampling cut, Bob's sampling cut, and the joint sampling cut, respectively, on input x, y.

Claim 4.7. *For all $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, the sets $\mathscr{S}_x, \mathscr{S}_y, \mathscr{S}_{x,y}$ are cuts in the protocol tree for π. Moreover,*

$$\varepsilon \notin \mathscr{S}_x, \tag{4.37}$$

$$\varepsilon \notin \mathscr{S}_y, \tag{4.38}$$

$$\varepsilon \notin \mathscr{S}_{x,y}. \tag{4.39}$$

Proof. That $\mathscr{S}_x$ and $\mathscr{S}_y$ are cuts is immediate from Proposition 3.4 and Claim 4.3. Again by Proposition 3.4 and Claim 4.3, the set $\mathscr{S}_{x,y}$ is the ceiling of two cuts and is therefore itself a cut in view of Proposition 3.5. Finally, (4.37)–(4.39) follow from Claims 4.3 and 4.6 in view of $\delta > 0$ and $\Delta > 0$. $\square$

4.4. A stochastic process. The focal point of our proof is the following discrete stochastic process, which mimics the operation of π on input distributed according to μ. We will study its information-theoretic properties in Sections 4.4–4.7 and provide its implementation as an efficient two-party communication protocol in Section 4.8. In the pseudocode below, $0 < c_0 < 1$ is the absolute constant from Claim 4.5.

```
$X, Y \leftarrow$ random input with joint distribution $\mu$
*leader* $\leftarrow \begin{cases} \textit{Alice} & \text{with probability } 1/2, \\ \textit{Bob} & \text{with probability } 1/2 \end{cases}$
**if** *leader* $=$ *Alice* **then**
    $V_1 \leftarrow$ random vertex in $\mathscr{S}_X$ with probability distribution $P_X$
    $V_2 \leftarrow \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \preceq V_1$
    $V_3 \leftarrow \begin{cases} V_2 & \text{with probability } \min\{1, c_0 P_{X,Y}(V_2)/P_X(V_2)\}, \\ \varepsilon & \text{with probability } 1 - \min\{1, c_0 P_{X,Y}(V_2)/P_X(V_2)\} \end{cases}$
    $\eta \leftarrow \begin{cases} 1 & \text{if } V_3 \succ \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ 1/2 & \text{if } V_3 \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ 0 & \text{if } V_3 \prec \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \end{cases}$
    $W \leftarrow \begin{cases} V_3 & \text{with probability } \eta, \\ \varepsilon & \text{with probability } 1 - \eta \end{cases}$
**else**
    $V_1 \leftarrow$ random vertex in $\mathscr{S}_Y$ with probability distribution $P_Y$
    $V_2 \leftarrow \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \cap \preceq V_1$
    $V_3 \leftarrow \begin{cases} V_2 & \text{with probability } \min\{1, c_0 P_{X,Y}(V_2)/P_Y(V_2)\}, \\ \varepsilon & \text{with probability } 1 - \min\{1, c_0 P_{X,Y}(V_2)/P_Y(V_2)\} \end{cases}$
    $\eta \leftarrow \begin{cases} 1 & \text{if } V_3 \succ \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor, \\ 1/2 & \text{if } V_3 \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor, \\ 0 & \text{if } V_3 \prec \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \end{cases}$
    $W \leftarrow \begin{cases} V_3 & \text{with probability } \eta, \\ \varepsilon & \text{with probability } 1 - \eta \end{cases}$
**end**
```

As discussed in the introduction, this stochastic process is inspired by Kol's one-step protocol [21] but uses a different family of cuts. In particular, the initial vertex V_1 is always sampled at or below the cut $\mathscr{R}_{\delta,1/2}$, which has far-reaching implications and is key to our improved results. Further differences with [21] will be highlighted in

Sections 4.6–4.8 as we discuss the information-theoretic properties of the stochastic process and its implementation by a two-party communication protocol.

In what follows, we let A and B denote the complementary events

$$\begin{aligned} A &\equiv \textit{leader} = \textit{Alice}, \\ B &\equiv \textit{leader} = \textit{Bob}. \end{aligned}$$

The corresponding parts of the stochastic process (lines 4–8 and 10–14) are symmetric in that either part results from the other by interchanging the roles of X and Y. We will make frequent use of this symmetry in our analysis. A comment is in order on the stated distribution of V_1 in lines 4 and 10. Recall from Claim 4.7 that $\mathscr{S}_X$ and $\mathscr{S}_Y$ are cuts. It follows from (4.5) that the restrictions of P_X and P_Y to these cuts are probability distributions. In particular, the stated probability distributions of V_1 are legitimate. We now examine the probability distributions that govern the next two vertices in the stochastic process, V_2 and V_3.

Claim 4.8. *The random variable V_2 obeys*

$$\mathbf{P}[V_2 = v \mid X, Y, A] = \begin{cases} P_X(v) & \text{if } v \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor, \\ 0 & \text{otherwise}, \end{cases}$$

$$\mathbf{P}[V_2 = v \mid X, Y, B] = \begin{cases} P_Y(v) & \text{if } v \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ 0 & \text{otherwise}. \end{cases}$$

Proof. By symmetry, it suffices to prove the former claim. By the algorithmic definition of P_X, choosing a random element V of a given cut $\mathscr{C}$ with probability distribution P_X is equivalent to choosing a random element V' of a higher cut $\mathscr{C}' \succeq \mathscr{C}$ with probability distribution P_X and outputting the unique ancestor of V' in $\mathscr{C}$. This proves the claim because V_1 is a random element of the cut $\mathscr{S}_X$ with probability distribution P_X, and V_2 is the unique ancestor of V_1 in the lower cut $\lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor$. ∎

Claim 4.9. *Let $0 < c_0 < 1$ be the absolute constant from Claim 4.5. Then*

$$\operatorname{supp}(V_3 \mid X, Y, A) \subseteq \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cup \{\varepsilon\}, \tag{4.40}$$

$$\mathbf{P}[V_3 = v \mid X, Y, A] \leqslant c_0 P_{X,Y}(v), \qquad v \neq \varepsilon, \tag{4.41}$$

$$\mathbf{P}[V_3 \neq \varepsilon \mid X, Y, A] \geqslant c_0 - c_0 \exp\left(-\frac{1}{\delta}\right). \tag{4.42}$$

Analogously,

$$\operatorname{supp}(V_3 \mid X, Y, B) \subseteq \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \cup \{\varepsilon\}, \tag{4.43}$$

$$\mathbf{P}[V_3 = v \mid X, Y, B] \leqslant c_0 P_{X,Y}(v), \qquad v \neq \varepsilon, \tag{4.44}$$

$$\mathbf{P}[V_3 \neq \varepsilon \mid X, Y, B] \geqslant c_0 - c_0 \exp\left(-\frac{1}{\delta}\right). \tag{4.45}$$

Proof. By symmetry, it suffices to prove (4.40)–(4.42). Property (4.40) is immediate by Claim 4.8 and the definition of V_3. One similarly verifies

$$\begin{aligned}\mathbf{P}[V_3 = v \mid X, Y, A] &\leqslant P_X(v) \min\left\{1, \frac{c_0 P_{X,Y}(v)}{P_X(v)}\right\} \\ &\leqslant c_0 P_{X,Y}(v)\end{aligned}$$

for $v \neq \varepsilon$. It remains to settle (4.42). Recall from Claims 4.3 and 4.7 that $\varepsilon \notin \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor$. As a result,

$$\begin{aligned}\mathbf{P}[V_3 \neq \varepsilon \mid X, Y, A] &= \sum_{v \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor} \mathbf{P}[V_3 = v \mid X, Y, A] \\ &= \sum_{v \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor} P_X(v) \min\left\{1, \frac{c_0 P_{X,Y}(v)}{P_X(v)}\right\} \\ &\geqslant \sum_{\substack{v \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor: \\ c_0 P_{X,Y}(v) \leqslant P_X(v)}} P_X(v) \cdot \frac{c_0 P_{X,Y}(v)}{P_X(v)} \\ &= c_0 \left(\sum_{v \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor} P_{X,Y}(v) - \sum_{\substack{v \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor: \\ c_0 P_{X,Y}(v) > P_X(v)}} P_{X,Y}(v) \right) \\ &\geqslant c_0 \left(1 - \exp\left(-\frac{1}{\delta}\right)\right),\end{aligned}$$

where the last step uses (4.5) and Claim 4.5. ∎

4.5. Accuracy analysis. Building on the newly obtained facts about V_2 and V_3, we now show that the output W of the stochastic process is representative of the original protocol π in the sense of the accuracy requirement of Theorem 4.1.

CLAIM 4.10. *The random variable W obeys*

$$\operatorname{supp}(W \mid X, Y) \subseteq \mathscr{S}_{X,Y} \cup \{\varepsilon\}, \tag{4.46}$$

$$\sum_{w \preceq v} \frac{\mathbf{P}[W = w \mid X, Y]}{\mathbf{P}[\Pi \succeq w \mid X, Y]} \leqslant 1 + \exp\left(-\frac{1}{\delta}\right), \qquad v \in \mathscr{V}(\pi). \tag{4.47}$$

Proof. For any cuts $\mathscr{C}$ and $\mathscr{D}$, Proposition 3.6 implies that

$$\begin{aligned}\lfloor \mathscr{C}, \mathscr{D} \rfloor &= (\mathscr{C} \cap {\prec}\mathscr{D}) \cup (\mathscr{D} \cap {\prec}\mathscr{C}) \cup (\mathscr{C} \cap \mathscr{D}), \\ \lceil \mathscr{C}, \mathscr{D} \rceil &= (\mathscr{C} \cap {\succ}\mathscr{D}) \cup (\mathscr{D} \cap {\succ}\mathscr{C}) \cup (\mathscr{C} \cap \mathscr{D}),\end{aligned}$$

where in both cases the union on the right-hand side is disjoint by Proposition 3.2. We will use these facts without further mention for $\mathscr{C} = \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor$ and $\mathscr{D} = \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor$, which are cuts by Proposition 3.4 and Claim 4.3.

For $w \neq \varepsilon$,

$$\begin{aligned}
\mathbf{P}[W = w \mid X, Y, A] &= \mathbf{P}[V_3 = w \mid X, Y, A]\, \mathbf{P}[W = w \mid X, Y, A,\ V_3 = w] \\
&= \begin{cases} \mathbf{P}[V_3 = w \mid X, Y, A] & \text{if } w \succ \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ \mathbf{P}[V_3 = w \mid X, Y, A]/2 & \text{if } w \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ 0 & \text{otherwise} \end{cases} \\
&\leqslant \begin{cases} c_0 P_{X,Y}(w) & \text{if } w \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \succ\lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ c_0 P_{X,Y}(w)/2 & \text{if } w \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}$$

where the final step is valid by Claim 4.9. A symmetric line of reasoning shows that

$$\mathbf{P}[W = w \mid X, Y, B] \leqslant \begin{cases} c_0 P_{X,Y}(w) & \text{if } w \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \cap \succ\lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor, \\ c_0 P_{X,Y}(w)/2 & \text{if } w \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \cap \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor, \\ 0 & \text{otherwise} \end{cases}$$

for $w \neq \varepsilon$. Recall from the description of the stochastic process that the events A and B are complementary and equally likely conditioned on X, Y. Therefore,

$$\begin{aligned}
\mathbf{P}[W = w \mid X, Y] &\leqslant \begin{cases} c_0 P_{X,Y}(w)/2 & \text{if } w \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \succ\lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ c_0 P_{X,Y}(w)/2 & \text{if } w \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \cap \succ\lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor, \\ c_0 P_{X,Y}(w)/2 & \text{if } w \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor, \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} c_0 P_{X,Y}(w)/2 & \text{if } w \in \mathscr{S}_{X,Y}, \\ 0 & \text{otherwise} \end{cases} \qquad (4.48)
\end{aligned}$$

for $w \neq \varepsilon$. Among other things, this proves (4.46).

We now bound the probability that $W = \varepsilon$. The stochastic process description makes it clear that

$$\begin{aligned}
\mathbf{P}[W = \varepsilon \mid X, Y, A] = {} & \mathbf{P}[V_3 = \varepsilon \mid X, Y, A] \\
& + \frac{1}{2}\mathbf{P}[V_3 \in \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \text{ and } V_3 \neq \varepsilon \mid X, Y, A] \\
& + \mathbf{P}[V_3 \prec \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \text{ and } V_3 \neq \varepsilon \mid X, Y, A].
\end{aligned}$$

To simplify the right-hand side of this equation, recall from Claims 4.3, 4.7, and 4.9 that $\operatorname{supp}(V_3 \mid X, Y, A) \subseteq \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cup \{\varepsilon\}$ and $\varepsilon \notin \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor$. As a result,

$$\begin{aligned}
\mathbf{P}[W = \varepsilon \mid X, Y, A] = {} & \mathbf{P}[V_3 = \varepsilon \mid X, Y, A] \\
& + \frac{1}{2}\mathbf{P}[V_3 \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \mid X, Y, A] \\
& + \mathbf{P}[V_3 \in \lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor \cap \prec\lfloor \mathscr{S}_Y, \mathscr{F}_{X,\delta} \rfloor \mid X, Y, A].
\end{aligned}$$

Substituting the bounds from Claim 4.9,

$$\begin{aligned}\mathbf{P}[W=\varepsilon \mid X,Y,A] \leqslant 1-c_0\left(1-\exp\left(-\frac{1}{\delta}\right)\right)&\\ +\frac{c_0}{2}P_{X,Y}(\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor \cap \lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor)&\\ +c_0P_{X,Y}(\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor \cap \prec\lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor).&\end{aligned}$$

An analogous argument yields

$$\begin{aligned}\mathbf{P}[W=\varepsilon \mid X,Y,B] \leqslant 1-c_0\left(1-\exp\left(-\frac{1}{\delta}\right)\right)&\\ +\frac{c_0}{2}P_{X,Y}(\lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor \cap \lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor)&\\ +c_0P_{X,Y}(\lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor \cap \prec\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor).&\end{aligned}$$

Since the events A and B are complementary and equally likely conditioned on X,Y, we conclude from the last two equations that

$$\begin{aligned}\mathbf{P}[W=\varepsilon \mid X,Y] &\leqslant 1-c_0\left(1-\exp\left(-\frac{1}{\delta}\right)\right)\\ &\quad+\frac{c_0}{2}P_{X,Y}(\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor \cap \lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor)\\ &\quad+\frac{c_0}{2}P_{X,Y}(\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor \cap \prec\lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor)\\ &\quad+\frac{c_0}{2}P_{X,Y}(\lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor \cap \prec\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor)\\ &= 1-c_0\left(1-\exp\left(-\frac{1}{\delta}\right)\right)\\ &\quad+\frac{c_0}{2}P_{X,Y}(\lfloor\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor, \lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor\rfloor)\\ &= 1-c_0\left(1-\exp\left(-\frac{1}{\delta}\right)\right)+\frac{c_0}{2},\end{aligned}\tag{4.49}$$

where the final step follows from (4.5) and the fact that $\lfloor\lfloor \mathscr{S}_X,\mathscr{F}_{Y,\delta}\rfloor, \lfloor \mathscr{S}_Y,\mathscr{F}_{X,\delta}\rfloor\rfloor$ is a cut (Proposition 3.4 and Claim 4.3).

We are now in a position to settle (4.47) and thereby complete the proof. Let v be an arbitrary vertex of the protocol tree. Recall from Claim 4.7 that $\mathscr{S}_{X,Y}$ is a cut. Therefore, (4.46) shows that the intersection of the chain $\preceq v$ and the support of W is a subset of $\{\varepsilon,s\}$ for some vertex $s\neq\varepsilon$. Applying the newly obtained estimates (4.48) and (4.49),

$$\begin{aligned}\sum_{w\preceq v}\frac{\mathbf{P}[W=w\mid X,Y]}{P_{X,Y}(w)}&\leqslant\frac{\mathbf{P}[W=\varepsilon\mid X,Y]}{P_{X,Y}(\varepsilon)}+\frac{\mathbf{P}[W=s\mid X,Y]}{P_{X,Y}(s)}\\&\leqslant 1-c_0\left(1-\exp\left(-\frac{1}{\delta}\right)\right)+\frac{c_0}{2}+\frac{c_0}{2}\\&=1+c_0\exp\left(-\frac{1}{\delta}\right)\\&\leqslant 1+\exp\left(-\frac{1}{\delta}\right).\end{aligned}$$

□

4.6. Progress is proportional to cost. Key to our amortized analysis is the quantity $\mathbf{E}\,\mathbb{D}_{X,Y}(\prec W)$, which is naturally viewed as a measure of progress in simulating π on random input. To paraphrase Theorem 4.1, we must show that the progress remains proportional to the communication cost. This claim does not have an analogue in Kol's work [21], where the communication cost may significantly exceed the progress measure. We start with a claim that relates the distributions of the initial and final vertices in the stochastic process, V_1 and W.

CLAIM 4.11. *Let $0<c_0<1$ be the absolute constant from Claim 4.5. Then the inequality*

$$\mathbf{P}[W=V_1\mid A,X,V_1]\geqslant\frac{c_0^2}{8}\tag{4.50}$$

holds except with probability at most $4\exp(-1/\delta)$ conditioned on A. Analogously, the inequality

$$\mathbf{P}[W=V_1\mid B,Y,V_1]\geqslant\frac{c_0^2}{8}\tag{4.51}$$

holds except with probability at most $4\exp(-1/\delta)$ conditioned on B.

A comment is in order regarding the notation. The left-hand side of (4.50) is a real-valued function of the random variables X and V_1. Therefore, this inequality defines an *event* in the probability space, making it meaningful to speak about its *probability.* Similar reasoning applies to (4.51).

Proof. By symmetry, it suffices to prove (4.50). Let E_1 and E_2 stand for the events that $V_1 \preceq \mathscr{F}_{Y,\delta}$ and $P_{X,Y}(V_2)/P_X(V_2) \geqslant c_0$, respectively. Then

$$\begin{aligned}
&\mathbf{P}[W = V_1 \mid A, X, V_1, E_1 \wedge E_2] \\
&\qquad = \mathbf{P}\left[V_1 = V_2 = V_3 = W, \eta \geqslant \frac{1}{2} \,\middle|\, A, X, V_1, E_1 \wedge E_2\right] \\
&\qquad = \mathbf{P}[V_1 = V_2 \mid A, X, V_1, E_1 \wedge E_2] \\
&\qquad\qquad\quad \times \mathbf{P}[V_2 = V_3 \mid A, X, V_1, E_1 \wedge E_2, V_1 = V_2] \\
&\qquad\qquad\quad \times \mathbf{P}\left[\eta \geqslant \frac{1}{2} \,\middle|\, A, X, V_1, E_1 \wedge E_2, V_1 = V_2 = V_3\right] \\
&\qquad\qquad\quad \times \mathbf{P}\left[W = V_1 \,\middle|\, A, X, V_1, E_1 \wedge E_2, V_1 = V_2 = V_3, \eta \geqslant \frac{1}{2}\right] \\
&\qquad \geqslant 1 \cdot c_0^2 \cdot \mathbf{P}\left[\eta \geqslant \frac{1}{2} \,\middle|\, A, X, V_1, E_1 \wedge E_2, V_1 = V_2 = V_3\right] \cdot \frac{1}{2} \\
&\qquad = \frac{c_0^2}{2},
\end{aligned} \tag{4.52}$$

where the final step follows from

$$\begin{aligned}
V_1 &\in \mathscr{S}_X \\
&= \lfloor \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{X,\Delta}, \mathscr{R}_{\delta,1/2} \rfloor \\
&\succeq \lfloor \mathscr{F}_{X,\delta}, \mathscr{R}_{\delta,1/2} \rfloor \\
&\succeq \lfloor \mathscr{F}_{X,\delta}, \mathscr{R}_{\delta,1/2}, \overline{\mathscr{F}_{Y,\delta}}, \mathscr{F}_{Y,\Delta} \rfloor \\
&= \lfloor \mathscr{F}_{X,\delta}, \mathscr{S}_Y \rfloor.
\end{aligned}$$

We now analyze the probability of $E_1 \wedge E_2$ conditioned on A. By definition, $V_1 \in \mathscr{S}_X \preceq \mathscr{R}_{\delta,1/2} \preceq \mathscr{R}_{\mathscr{Y},\delta,1/2}$. As a result, the defining equation (4.20) for $\mathscr{R}_{\mathscr{Y},\delta,1/2}$ reveals that

$$\mathbf{P}[E_1 \mid A, X, V_1] > \frac{1}{2}. \tag{4.53}$$

Moving on to E_2, recall from Claim 4.8 that V_2 is a random element of $\lfloor \mathscr{S}_X, \mathscr{F}_{Y,\delta} \rfloor$ with probability distribution P_X. Claim 4.5 now ensures that $\mathbf{P}[\neg E_2 \mid A, X, Y] \leqslant \exp(-1/\delta)$ and in particular $\mathbf{P}[\neg E_2 \mid A] \leqslant \exp(-1/\delta)$. By Markov's inequality,

$$\mathbf{P}[\neg E_2 \mid A, X, V_1] \leqslant \frac{1}{4}$$

except with probability at most $4\exp(-1/\delta)$ conditioned on A. Combining this result with (4.53) shows that

$$\mathbf{P}[E_1 \wedge E_2 \mid A, X, V_1] > \frac{1}{4}$$

except with probability at most $4\exp(-1/\delta)$ conditioned on A. This completes the proof in light of (4.52). ∎

We are now in a position to prove our first lower bound on the progress measure $\mathbf{E}\,\mathbb{D}_{X,Y}(\prec W)$. Looking ahead, this claim will be crucial in arguing that the progress made by the sampling procedure remains proportional to the communication cost.

Claim 4.12. *Let $0 < c_0 < 1$ be the absolute constant from Claim 4.5. Then*

$$\begin{aligned}&\mathbf{E}[\mathbb{D}_X(\prec V_1) \mid A] + \mathbf{E}[\mathbb{D}_Y(\prec V_1) \mid B]\\&\qquad\qquad\leqslant \frac{16}{c_0^2}\,\mathbf{E}\,\mathbb{D}_{X,Y}(\prec W) + 8(\Delta+\delta)\exp\left(-\frac{1}{\delta}\right).\end{aligned}$$

Proof. It follows from Claim 4.11 that the lower bound

$$\frac{8}{c_0^2}\,\mathbf{E}[\mathbb{D}_X(\prec W) \mid A, X, V_1] \geqslant \mathbb{D}_X(\prec V_1) \tag{4.54}$$

holds with probability at least $1-4\exp(-1/\delta)$ with respect to (X, V_1) conditioned on A. The offending (X, V_1) pairs satisfy the weaker inequality

$$\frac{8}{c_0^2}\,\mathbf{E}[\mathbb{D}_X(\prec W) \mid A, X, V_1] \geqslant \mathbb{D}_X(\prec V_1) - (\Delta+\delta), \tag{4.55}$$

which can be verified by observing that $V_1 \in \mathscr{S}_X \preceq \mathscr{F}_{X,\Delta}$ and therefore $\mathbb{D}_X(\prec V_1) \leqslant \Delta+\delta$. Taking a weighted average of (4.54) and (4.55) according to the probabilities of the corresponding (X, V_1) pairs, we arrive at

$$\frac{8}{c_0^2}\,\mathbf{E}[\mathbb{D}_X(\prec W) \mid A] \geqslant \mathbf{E}[\mathbb{D}_X(\prec V_1) \mid A] - 4(\Delta+\delta)\exp\left(-\frac{1}{\delta}\right).$$

An analogous argument yields

$$\frac{8}{c_0^2}\,\mathbf{E}[\mathbb{D}_Y(\prec W) \mid B] \geqslant \mathbf{E}[\mathbb{D}_Y(\prec V_1) \mid B] - 4(\Delta+\delta)\exp\left(-\frac{1}{\delta}\right).$$

Since the events A and B are complementary and equally likely, the claim follows immediately from the last two inequalities. ∎

4.7. Progress is nonnegligible. In addition to proving that the progress remains proportional to the communication cost, Theorem 4.1 requires us to prove that the progress is nonnegligible. To be precise, we must show that with constant probability, the output vertex W is either a leaf or has nonnegligible progress measure $\mathbb{D}_{X,Y}(\prec W)$. The corresponding claims in the work of Barak et al. [4] and Kol [21] were trivial to show. Our situation is considerably more involved because W is always at or below the cut $\mathscr{R}_{\delta,1/2}$, which a priori makes it possible for the progress measure to be arbitrarily small. We start with a general claim about cuts in the protocol tree.

Claim 4.13. *Let $\mathscr{C}$ and $\mathscr{D}$ be cuts in the protocol tree for π. Then for any $x \in \mathscr{X}$ and $y \in \mathscr{Y}$,*

$$
\begin{align}
P_{x,y}(\mathscr{C} \cap \succeq\!\mathscr{D}) &= P_{x,y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor \cap \succeq\!\mathscr{D}), \tag{4.56}\\
P_{x}(\mathscr{C} \cap \succeq\!\mathscr{D}) &= P_{x}(\lfloor \mathscr{C}, \mathscr{D} \rfloor \cap \succeq\!\mathscr{D}), \tag{4.57}\\
P_{y}(\mathscr{C} \cap \succeq\!\mathscr{D}) &= P_{y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor \cap \succeq\!\mathscr{D}), \tag{4.58}\\
P(\mathscr{C} \cap \succeq\!\mathscr{D}) &= P(\lfloor \mathscr{C}, \mathscr{D} \rfloor \cap \succeq\!\mathscr{D}). \tag{4.59}
\end{align}
$$

Proof. It suffices to prove (4.56) since each of the equations (4.57)–(4.59) is a convex combination of (4.56) for appropriate x and y. We have

$$
\begin{aligned}
P_{x,y}(\mathscr{C} \cap \succeq\!\mathscr{D}) &= P_{x,y}(\mathscr{C}) - P_{x,y}(\mathscr{C} \cap \prec\!\mathscr{D})\\
&= 1 - P_{x,y}(\mathscr{C} \cap \prec\!\mathscr{D})\\
&= P_{x,y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor) - P_{x,y}(\mathscr{C} \cap \prec\!\mathscr{D})\\
&= P_{x,y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor) - P_{x,y}(((\mathscr{D} \cap \preceq\!\mathscr{C}) \cup (\mathscr{C} \cap \preceq\!\mathscr{D})) \cap \prec\!\mathscr{D})\\
&= P_{x,y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor) - P_{x,y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor \cap \prec\!\mathscr{D})\\
&= P_{x,y}(\lfloor \mathscr{C}, \mathscr{D} \rfloor \cap \succeq\!\mathscr{D}),
\end{aligned}
$$

where the first step uses Proposition 3.2, the second step is immediate from (4.5), the third step follows from (4.5) and Proposition 3.4, the fourth step uses Proposition 3.2 again, the fifth step uses Proposition 3.6, and the final step is valid yet again by Proposition 3.2. □

Key to our proof of nonnegligible progress is the following information-theoretic lemma, obtained by an application of the chain rule and convexity.

Claim 4.14. *For any cut $\mathscr{C}$ and parameters $\theta > 0$ and $0 < \rho \leqslant 1$,*

$$
\begin{align}
\mathbf{E}\, P_X(\mathscr{C} \cap \succeq\!\mathscr{F}_{X,\theta}) &\geqslant \frac{\rho}{3}\, P(\mathscr{C} \cap \succeq\!\mathscr{R}_{\mathscr{X},\theta,\rho}) - \theta, \tag{4.60}\\
\mathbf{E}\, P_Y(\mathscr{C} \cap \succeq\!\mathscr{F}_{Y,\theta}) &\geqslant \frac{\rho}{3}\, P(\mathscr{C} \cap \succeq\!\mathscr{R}_{\mathscr{Y},\theta,\rho}) - \theta. \tag{4.61}
\end{align}
$$

Proof. By symmetry, it suffices to establish (4.60). Recall from Claim 4.3 that $\mathscr{F}_{X,\theta}$ is a cut. We have

$$
\begin{align}
&\delta\, \mathbf{E}\, P_X(\mathscr{C} \cap \succeq\!\mathscr{F}_{X,\theta}) + \theta \notag\\
&\qquad = \delta\, \mathbf{E}\, P_X(\lfloor \mathscr{C}, \mathscr{F}_{X,\theta} \rfloor \cap \succeq\!\mathscr{F}_{X,\theta}) + \theta \notag\\
&\qquad \geqslant \mathbf{E}\, \mathbb{D}_X(\prec\!(\lfloor \mathscr{C}, \mathscr{F}_{X,\theta} \rfloor \cap \preceq\!\Pi)) \notag\\
&\qquad = \mathbf{E}\, \mathrm{KL}(P_X|_{\lfloor \mathscr{C}, \mathscr{F}_{X,\theta} \rfloor} \parallel P|_{\lfloor \mathscr{C}, \mathscr{F}_{X,\theta} \rfloor}) \notag\\
&\qquad \geqslant \mathbf{E}\, \mathrm{KL}(P_X(\lfloor \mathscr{C}, \mathscr{F}_{X,\theta} \rfloor \cap \succeq\!\mathscr{F}_{X,\theta}) \parallel P(\lfloor \mathscr{C}, \mathscr{F}_{X,\theta} \rfloor \cap \succeq\!\mathscr{F}_{X,\theta})) \notag\\
&\qquad = \mathbf{E}\, \mathrm{KL}(P_X(\mathscr{C} \cap \succeq\!\mathscr{F}_{X,\theta}) \parallel P(\mathscr{C} \cap \succeq\!\mathscr{F}_{X,\theta})) \notag\\
&\qquad \geqslant \mathrm{KL}(\mathbf{E}\, P_X(\mathscr{C} \cap \succeq\!\mathscr{F}_{X,\theta}) \parallel \mathbf{E}\, P(\mathscr{C} \cap \succeq\!\mathscr{F}_{X,\theta})), \tag{4.62}
\end{align}
$$

where the first step uses Claim 4.13, the second step is immediate from (4.14), the third step follows from (4.8) and Theorem 2.7, the fourth step uses the chain rule for the Kullback–Leibler divergence (Fact 2.4), the fifth step uses Claim 4.13 again, and the final step uses the convexity of the Kullback–Leibler divergence (Fact 2.2). Moreover,

$$\mathbf{E}\, P(\mathscr{C} \cap {\succeq}\mathscr{F}_{X,\theta}) \geqslant \rho P(\mathscr{C} \cap {\succeq}\mathscr{R}_{\mathscr{X},\theta,\rho}) \tag{4.63}$$

by the definition of $\mathscr{R}_{\mathscr{X},\theta,\rho}$. Now assume for the sake of contradiction that (4.60) is false. Then

$$\begin{aligned}
&\delta\, \mathbf{E}\, P_X(\mathscr{C} \cap {\succeq}\mathscr{F}_{X,\theta}) + \theta \\
&\qquad \geqslant \mathrm{KL}(\mathbf{E}\, P_X(\mathscr{C} \cap {\succeq}\mathscr{F}_{X,\theta}) \parallel \mathbf{E}\, P(\mathscr{C} \cap {\succeq}\mathscr{F}_{X,\theta})) \\
&\qquad \geqslant \mathrm{KL}\left(\frac{\rho}{3} P(\mathscr{C} \cap {\succeq}\mathscr{R}_{\mathscr{X},\theta,\rho}) \;\middle\|\; \rho P(\mathscr{C} \cap {\succeq}\mathscr{R}_{\mathscr{X},\theta,\rho})\right) \\
&\qquad \geqslant \frac{\rho}{3} P(\mathscr{C} \cap {\succeq}\mathscr{R}_{\mathscr{X},\theta,\rho}) \\
&\qquad > \mathbf{E}\, P_X(\mathscr{C} \cap {\succeq}\mathscr{F}_{X,\theta}) + \theta,
\end{aligned}$$

where the first inequality restates (4.62), the second uses (4.63) and the assumption, the third follows from (2.1), and the final step uses the assumption yet again. The promised contradiction results in view of (4.12). □

We are now in a position to give the promised proof that the sampling procedure simulates a nontrivial portion of protocol π.

Claim 4.15. *Let $0 < c_0 < 1$ be the absolute constant from Claim 4.5. Then*

$$\mathbf{P}[W \succeq \lfloor \mathscr{F}_{X,\delta}, \mathscr{F}_{Y,\delta} \rfloor] \geqslant \frac{c_0^2}{400}. \tag{4.64}$$

In particular,

$$\mathbf{P}[W \in \mathscr{L}(\pi)] + \frac{1}{\delta}\, \mathbf{E}\, \mathbb{D}_{X,Y}(\prec W) \geqslant \frac{c_0^2}{400}. \tag{4.65}$$

Proof. We have

$$\begin{aligned}
\mathbf{P}[V_1 \succeq \mathscr{F}_{X,\delta} \mid A] &= \mathbf{E}\, P_X(\mathscr{S}_X \cap {\succeq}\mathscr{F}_{X,\delta}) \\
&= \mathbf{E}\, P_X(\lfloor \mathscr{S}_X, \mathscr{F}_{X,\delta} \rfloor \cap {\succeq}\mathscr{F}_{X,\delta}) \\
&= \mathbf{E}\, P_X(\lfloor \mathscr{R}_{\delta,1/2}, \overline{\mathscr{F}_{X,\delta}}, \mathscr{F}_{X,\Delta}, \mathscr{F}_{X,\delta} \rfloor \cap {\succeq}\mathscr{F}_{X,\delta}) \\
&= \mathbf{E}\, P_X(\lfloor \mathscr{R}_{\delta,1/2}, \mathscr{F}_{X,\delta} \rfloor \cap {\succeq}\mathscr{F}_{X,\delta}) \\
&= \mathbf{E}\, P_X(\mathscr{R}_{\delta,1/2} \cap {\succeq}\mathscr{F}_{X,\delta}) \\
&\geqslant \frac{1}{6} P(\mathscr{R}_{\delta,1/2} \cap {\succeq}\mathscr{R}_{\mathscr{X},\delta,1/2}) - \delta,
\end{aligned}$$

where the second step follows from Claim 4.13 and the fact that $\mathscr{F}_{X,\delta}$ and $\mathscr{S}_X$ are cuts (Claims 4.3 and 4.7); the third step uses the defining equation for $\mathscr{S}_X$; the fourth step is valid by Claim 4.6 and our parameter settings (4.12) and (4.13); the fifth step uses Claim 4.13 and the fact that $\mathscr{R}_{\delta,1/2}$ and $\mathscr{F}_{X,\delta}$ are cuts (Claim 4.3); and the final step is immediate from Claim 4.14 for the cut $\mathscr{C} = \mathscr{R}_{\delta,1/2}$. In view of Claim 4.11, we conclude that

$$\begin{aligned}\mathbf{P}[W \succeq \mathscr{F}_{X,\delta} \mid A] &\geqslant \frac{c_0^2}{8}\left(\mathbf{P}[V_1 \succeq \mathscr{F}_{X,\delta} \mid A] - 4\exp\left(-\frac{1}{\delta}\right)\right)\\ &\geqslant \frac{c_0^2}{8}\left(\frac{1}{6}P(\mathscr{R}_{\delta,1/2} \cap {\succeq}\mathscr{R}_{\mathscr{X},\delta,1/2}) - \delta - 4\exp\left(-\frac{1}{\delta}\right)\right).\end{aligned}$$

An analogous argument applied to Bob yields

$$\mathbf{P}[W \succeq \mathscr{F}_{Y,\delta} \mid B] \geqslant \frac{c_0^2}{8}\left(\frac{1}{6}P(\mathscr{R}_{\delta,1/2} \cap {\succeq}\mathscr{R}_{\mathscr{Y},\delta,1/2}) - \delta - 4\exp\left(-\frac{1}{\delta}\right)\right).$$

Since the events A and B are complementary and equally likely, the last two inequalities imply that

$$\begin{aligned}&\mathbf{P}[W \succeq \lfloor \mathscr{F}_{X,\delta}, \mathscr{F}_{Y,\delta}\rfloor]\\ &\quad\geqslant \frac{c_0^2}{8}\left(\frac{1}{12}P(\mathscr{R}_{\delta,1/2} \cap {\succeq}\mathscr{R}_{\mathscr{X},\delta,1/2}) + \frac{1}{12}P(\mathscr{R}_{\delta,1/2} \cap {\succeq}\mathscr{R}_{\mathscr{Y},\delta,1/2})\right.\\ &\qquad\qquad\left. - \delta - 4\exp\left(-\frac{1}{\delta}\right)\right)\\ &\quad\geqslant \frac{c_0^2}{8}\left(\frac{1}{12}P(\mathscr{R}_{\mathscr{X},\delta,1/2} \cap {\preceq}\mathscr{R}_{\mathscr{Y},\delta,1/2}) + \frac{1}{12}P(\mathscr{R}_{\mathscr{Y},\delta,1/2} \cap {\preceq}\mathscr{R}_{\mathscr{X},\delta,1/2})\right.\\ &\qquad\qquad\left. - \delta - 4\exp\left(-\frac{1}{\delta}\right)\right)\\ &\quad\geqslant \frac{c_0^2}{8}\left(\frac{1}{12}P(\mathscr{R}_{\delta,1/2}) - \delta - 4\exp\left(-\frac{1}{\delta}\right)\right)\\ &\quad= \frac{c_0^2}{8}\left(\frac{1}{12} - \delta - 4\exp\left(-\frac{1}{\delta}\right)\right),\end{aligned}$$

where the second and third steps both follow from $\mathscr{R}_{\delta,1/2} = \lfloor \mathscr{R}_{\mathscr{X},\delta,1/2}, \mathscr{R}_{\mathscr{Y},\delta,1/2}\rfloor$ and Proposition 3.6, whereas the final step uses (4.5) and the fact that $\mathscr{R}_{\delta,1/2}$ is a cut (Claim 4.3). In view of (4.12), this completes the proof of (4.64). The remaining inequality (4.65) now follows by the definition of $\mathscr{F}_{X,\delta}$ and $\mathscr{F}_{Y,\delta}$. □

4.8. Implementation and cost analysis. A literal interpretation of the stochastic process as a two-party communication protocol requires Alice and Bob to transmit labels of vertices of the protocol tree for π, resulting in an arbitrarily high communication cost. We now give a more efficient implementation that keeps the communication cost low with respect to the input distribution μ. By symmetry, it suffices to implement lines 4–8, corresponding to the designation of Alice as the leader. As elsewhere in the proof, we let X and Y stand for Alice and Bob's inputs. We condition every step of the implementation on X and Y, keeping in mind for the purposes of cost and error analysis that X and Y are distributed independently according to the product distribution μ.

Correlated sampling. Given X, consider the following two-step randomized procedure for choosing a leaf of the protocol tree. First, one chooses a random vertex V_1 of the cut $\mathscr{S}_X$ according to the probability distribution P_X, exactly as in the stochastic process. Then, one chooses a random leaf in the protocol subtree rooted at V_1, according to the probability distribution induced by P on such leaves. Letting V denote the leaf so generated,

$$\begin{aligned}\mathrm{KL}(V \parallel P|_{\mathscr{L}(\pi)}) &= \mathrm{KL}(P_X|_{\mathscr{S}_X} \parallel P|_{\mathscr{S}_X}) \\ &= \mathbf{E}[\mathbb{D}_X(\prec V_1) \mid A, X],\end{aligned} \tag{4.66}$$

where the first step holds by the chain rule for the Kullback–Leibler divergence (Fact 2.4) and the second step uses (4.8) and Theorem 2.7. Alice and Bob begin their implementation of the stochastic process by executing the correlated sampling procedure of Theorem 2.16 with error parameter $1/r$ for the probability distributions of V and P, respectively, on the leaves of the protocol tree. This incurs a communication cost of $O(\log r) + C'$ for a nonnegative random variable C' with expected value

$$\mathbf{E}[C' \mid A] = O(\mathbf{E}[\mathbb{D}_X(\prec V_1) \mid A]), \tag{4.67}$$

and results in a pair of leaves $V^A = V$ and V^B for Alice and Bob, respectively, such that $V = V^A = V^B$ except with probability at most $1/r$. The piecewise definition of the distribution that governs V is a key departure from the work of Kol [21], where the correlated sampling is applied to the probability distributions P_X and P on the leaves of the protocol tree.

We may assume that the correlated sampling uses only public randomness because any private random bits can always be replaced with public ones without increasing the communication cost. Since Bob's input distribution P is public knowledge, we conclude that the transcript of the correlated sampling procedure reveals his computed vertex V^B. If $V^B = V$, Alice makes an announcement to that effect. In the complementary case, which by the previous paragraph occurs with probability no greater than $1/r$, she sends V to Bob, incurring an arbitrarily high communication cost. In either case, the resulting communication transcript uniquely identifies V.

Key vertices. With V known to both parties, consider the following vertices on the root-to-leaf path $\preceq V$:

$$V_X = \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \mathscr{F}_{X,\delta} \rfloor \cap {\preceq V}, \tag{4.68}$$

$$\overline{V_X} = \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \overline{\mathscr{F}_{X,\delta}} \rfloor \cap {\preceq V}, \tag{4.69}$$

$$V_Y = \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \mathscr{F}_{Y,\delta} \rfloor \cap {\preceq V}, \tag{4.70}$$

$$\overline{V_Y} = \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \overline{\mathscr{F}_{Y,\delta}} \rfloor \cap {\preceq V}, \tag{4.71}$$

$$\overline{\overline{V_Y}} = \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \mathscr{F}_{Y,\Delta} \rfloor \cap {\preceq V}, \tag{4.72}$$

$$R_i = \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \mathscr{R}_{\mathscr{Y},\delta,\frac{i}{2r}} \rfloor \cap {\preceq V}, \qquad i = 0, 1, 2, \ldots, r. \tag{4.73}$$

Any two of the vertices in (4.68)–(4.73) are comparable, a fact that we will use extensively without further mention. Equations (4.8) and (4.10) make it clear that Alice can compute the frontier $\mathscr{F}_{X,\theta}$ on her own for any θ, and similarly Bob can compute the frontier $\mathscr{F}_{Y,\theta}$ on his own. Moreover, the cuts $\mathscr{R}_{\mathscr{X},\theta,\rho}$ and $\mathscr{R}_{\mathscr{Y},\theta,\rho}$ for any θ and ρ are independent of Alice and Bob's inputs and are known to them both. In particular, Alice knows the vertices $V_X, \overline{V_X}$, Bob knows $V_Y, \overline{V_Y}, \overline{\overline{V_Y}}$, and they both know $R_0, R_1, \ldots, R_r$. They now additionally exchange $\overline{V_X}$ and $\overline{V_Y}$ by reporting the smallest indices $i^*, j^* \in \{1, 2, \ldots, r\}$ such that

$$\begin{aligned}\overline{V_X} &= \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \mathscr{R}_{\mathscr{Y},\delta,\frac{i^*}{2r}} \rfloor \cap {\preceq} V,\\ \overline{V_Y} &= \lfloor \mathscr{R}_{\delta,\frac{1}{2}}, \mathscr{R}_{\mathscr{X},\delta,\frac{j^*}{2r}} \rfloor \cap {\preceq} V.\end{aligned}$$

Comparing V_X and V_Y. Next, Alice and Bob establish the precise relation ($\prec$, $=$, or $\succ$) between V_X and V_Y. This step is implemented exactly as in Kol's work [21]. Specifically, recall that $R_{i^*-1} \prec V_X \preceq R_{i^*}$ where i^* is known to both Alice and Bob. In the following cases, to be detected and announced by Bob, the comparison of V_X and V_Y requires only constantly many bits of communication:

$$\begin{aligned}&V_Y \preceq R_{i^*-1},\\ &V_Y = R_{i^*},\\ &V_Y \succ R_{i^*}.\end{aligned}$$

The remaining case occurs with probability

$$\mathbf{P}[R_{i^*-1} \prec V_Y \prec R_{i^*} \mid V, i^*] \leqslant \frac{1}{2r}, \tag{4.74}$$

by Claim 4.4 and the independence of Y and (V, i^*). In this exceptional scenario, Alice and Bob compare V_X and V_Y by exchanging the labels of the corresponding vertices, for an arbitrarily high communication cost.

Main simulation. Alice checks whether $\overline{V_X} \succ \mathscr{F}_{X,\Delta}$ and informs Bob accordingly. This event occurs with probability

$$\begin{aligned}\mathbf{P}[\overline{V_X} \succ \mathscr{F}_{X,\Delta}] &\leqslant \mathbf{P}[\mathbb{D}_X({\prec}(\mathscr{S}_X \cap {\preceq} V)) \geqslant \Delta]\\ &= \mathbf{P}[\mathbb{D}_X({\prec}(\mathscr{S}_X \cap {\preceq}\Pi)) \geqslant \Delta]\\ &\leqslant \mathbf{P}[\mathbb{D}_X({\prec}\Pi) \geqslant \Delta]\\ &\leqslant \mathbf{P}[\mathbb{D}_{X,Y}({\prec}\Pi) \geqslant \Delta]\\ &\leqslant \frac{\mathbf{E}\,\mathbb{D}_{X,Y}({\prec}\Pi)}{\Delta},\end{aligned} \tag{4.75}$$

where the final step in the derivation uses Markov's inequality. If $\overline{V_X} \succ \mathscr{F}_{X,\Delta}$, then Alice sets $V_1 = \mathscr{F}_{X,\Delta} \cap {\preceq}\overline{V_X}$ and they proceed with a literal execution of lines 5–8 of the stochastic process, incurring an arbitrarily high communication cost. In what follows, we treat the complementary case $\overline{V_X} \preceq \mathscr{F}_{X,\Delta}$, whence $V_1 = \overline{V_X}$. Using

(4.6), (4.68)–(4.72), and the newly obtained equality $V_1 = \overline{V_X}$, the relevant part of the stochastic process simplifies as follows.

```
4    ⋮
5    $V_2 \leftarrow \lfloor \overline{V_X}, V_Y \rfloor$
6    $V_3 \leftarrow \begin{cases} V_2 & \text{with probability } \min\{1, c_0 P_Y(V_2)/P(V_2)\}, \\ \varepsilon & \text{with probability } 1 - \min\{1, c_0 P_Y(V_2)/P(V_2)\} \end{cases}$
7    $\eta \leftarrow \begin{cases} 1 & \text{if } \lfloor \overline{V_X}, V_Y \rfloor \succ \lfloor V_X, \overline{V_Y}, \overline{\overline{V_Y}} \rfloor, \\ 1/2 & \text{if } \lfloor \overline{V_X}, V_Y \rfloor = \lfloor V_X, \overline{V_Y}, \overline{\overline{V_Y}} \rfloor, \\ 0 & \text{if } \lfloor \overline{V_X}, V_Y \rfloor \prec \lfloor V_X, \overline{V_Y}, \overline{\overline{V_Y}} \rfloor \end{cases}$
8    $W \leftarrow \begin{cases} V_3 & \text{with probability } \eta, \\ \varepsilon & \text{with probability } 1 - \eta \end{cases}$
9    ⋮
```

Bob executes lines 5 and 6 without any help from Alice. For lines 7 and 8, he observes that

$$\lfloor \overline{V_X}, V_Y \rfloor \succeq \lfloor V_X, \overline{V_Y}, \overline{\overline{V_Y}} \rfloor \quad \Leftrightarrow \quad V_Y \succeq V_X \text{ or } V_Y = \overline{V_Y}, \tag{4.76}$$

$$\lfloor \overline{V_X}, V_Y \rfloor \preceq \lfloor V_X, \overline{V_Y}, \overline{\overline{V_Y}} \rfloor \quad \Leftrightarrow \quad V_Y \preceq V_X \text{ or } V_X = \overline{V_X}. \tag{4.77}$$

Therefore, Bob can execute lines 7 and 8 with a single bit of communication from Alice, indicating whether $V_X = \overline{V_X}$.

Sending back W. As a final step, Bob needs to send Alice W. If $W = \varepsilon$, he announces that fact and the protocol ends. If $W \neq \varepsilon$, then necessarily $W = V_2$ and $\lfloor \overline{V_X}, V_Y \rfloor \succeq \lfloor V_X, \overline{V_Y}, \overline{\overline{V_Y}} \rfloor$. Applying (4.76),

$$\begin{aligned} W &= \lfloor \overline{V_X}, V_Y \rfloor \\ &= \begin{cases} \overline{V_Y} & \text{if } V_Y \prec V_X, \\ \overline{V_X} & \text{if } V_Y \succeq \overline{V_X}, \\ V_Y & \text{otherwise.} \end{cases} \end{aligned}$$

Based on the communication so far, Bob knows which of the three cases applies. In the first two cases, Bob identifies W to Alice with constantly many bits by referring to the previously announced vertices $\overline{V_X}$ and $\overline{V_Y}$. In the third case, Bob sends W verbatim, incurring an arbitrarily high communication cost. By (4.74), this third case occurs with probability

$$\begin{aligned} \mathbf{P}[V_X \preceq V_Y \prec \overline{V_X} \mid X, V] &\leqslant \mathbf{P}[R_{i^*-1} \prec V_Y \prec R_{i^*} \mid V, i^*] \\ &\leqslant \frac{1}{2r}. \end{aligned} \tag{4.78}$$

In all three cases, W is fully determined by the communication transcript.

4.9. Summary and parameter settings. We now summarize our work so far and set the parameters so as to complete the proof of Theorem 4.1. The cost of the described implementation on input X, Y is

$$C \leqslant O(\log r) + C' + C'', \tag{4.79}$$

where C'' is the arbitrarily high communication cost that may be incurred due to failure in the correlated sampling step or in the exceptional cases (4.74), (4.75), (4.78). By above,

$$\mathbf{P}[C'' > 0] \leqslant \frac{\mathbf{E}\,\mathbb{D}_{X,Y}(\prec\Pi)}{\Delta} + \frac{2}{r}. \tag{4.80}$$

The other cost component, C', obeys the upper bound (4.67) and therefore by symmetry

$$\mathbf{E}[C' \mid B] = O(\mathbf{E}[\mathbb{D}_Y(\prec V_1) \mid B]). \tag{4.81}$$

Since the events A and B are complementary and equally likely, we obtain from (4.67), (4.81), and Claim 4.12 that

$$\mathbf{E}\,C' = O\left(\mathbf{E}\,\mathbb{D}_{X,Y}(\prec W) + (\Delta+\delta)\exp\left(-\frac{1}{\delta}\right)\right). \tag{4.82}$$

Now the accuracy, progress, and cost requirements of Theorem 4.1 follow from (4.47), (4.65), and (4.79)–(4.82), respectively, by taking

$$r = \Theta\left(\frac{1}{\epsilon}\right), \tag{4.83}$$

$$\delta = \Theta\left(\frac{1}{\log(1/\epsilon)}\right), \tag{4.84}$$

$$\Delta = \left\lceil \frac{2}{\epsilon}\,\mathbf{E}\,\mathbb{D}_{X,Y}(\prec\Pi) \right\rceil. \tag{4.85}$$

This concludes the proof of Theorem 4.1.

5. Complete simulation

Building on the sampling procedure of the previous section, we now prove the main result of this work.

Theorem 5.1 (Main theorem). *Let $0 < \epsilon < 1/2$ be given. Fix any public- or private-coin protocol π with input space $\mathscr{X} \times \mathscr{Y}$. Let μ be a product distribution on $\mathscr{X} \times \mathscr{Y}$, and abbreviate $I = \mathrm{IC}_\mu(\pi)$. Then there is a public-coin protocol π' with worst-case communication cost*

$$O\left(\frac{I}{\epsilon}\log^2\frac{I}{\epsilon}\right)$$

such that

$$\pi' \hookrightarrow_{\mu,\epsilon} \pi.$$

The remainder of this section is devoted to the proof of Theorem 5.1. Let $\delta = \delta(I, \epsilon) > 0$ be an accuracy parameter to be set later, and let $\beta = \beta(\delta) > 0$ be sufficiently small in the sense of Theorem 4.1. By Theorems 2.13–2.15, we may assume that π is a private-coin β-balanced protocol. Recall that our proof strategy in simulating π will be to repeatedly apply the partial sampling procedure of the previous section until a communication limit is exceeded. We will argue that the resulting simulation reaches a leaf with high probability and that its distribution is statistically close to the distribution of the transcript of π on the corresponding input. Following our methodology in Theorem 4.1, we will first define an abstract stochastic process, settle its information-theoretic properties, and then convert it to an efficient communication protocol for simulating π.

5.1. A stochastic process. Let X, Y be a pair of inputs with joint distribution μ. We define a discrete stochastic process given by the random variables X, Y, and

$$(\pi_t, \mu_t, R_t, M_t, W_t, C'_t, C''_t), \qquad\qquad t = 1, 2, 3, \ldots, \tag{5.1}$$

where $\mu_1, \mu_2, \mu_3, \ldots$ are product distributions on $\mathscr{X} \times \mathscr{Y}$. We let $\pi_1 = \pi$ and $\mu_1 = \mu$. For $t \geqslant 1$, the random variables X, Y, π_t, μ_t give rise to $R_t, M_t, W_t, C'_t, C''_t, \pi_{t+1}, \mu_{t+1}$ in an inductive manner as follows.

(i) Execute the public-coin protocol $\sigma_{\pi_t,\mu_t,\delta}$ from Theorem 4.1 on input X, Y. Let R_t and M_t denote the shared random string and the rest of the protocol transcript, respectively, from that execution. Let W_t, C'_t, C''_t be the corresponding additional random variables from Theorem 4.1, each of which is completely determined by the tuple (π_t, μ_t, R_t, M_t).

(ii) Define π_{t+1} to be the private-coin protocol corresponding to the protocol subtree of π_t rooted at W_t. Thus, vertex W_t of the protocol tree for π_t corresponds to vertex ε (the root) of the protocol tree for π_{t+1}.

(iii) Define μ_{t+1} to be the posterior probability distribution on $\mathscr{X} \times \mathscr{Y}$ obtained by conditioning μ_t on the transcript (R_t, M_t) of protocol $\sigma_{\pi_t,\mu_t,\delta}$. Recall that conditioning a product distribution on a protocol transcript results in a product distribution. Thus, μ_{t+1} is a product distribution, maintaining the promised invariant.

We let P denote the resulting infinite sequence (5.1) of random variables. For $t = 1, 2, 3, \ldots$, we let $P_{\leqslant t}$ denote the restriction of P to the first t stages of the stochastic process. In other words, $P_{\leqslant t}$ stands for

$$(\pi_1, \mu_1, R_1, M_1, W_1, C'_1, C''_1), \ldots, (\pi_t, \mu_t, R_t, M_t, W_t, C'_t, C''_t), \pi_{t+1}, \mu_{t+1},$$

where the inclusion of π_{t+1} and μ_{t+1} is motivated by the fact that they are fully determined by the previous tuple. In this notation, μ_{t+1} is the probability distribution that governs the random variable $XY \mid P_{\leqslant t}$.

We define the random variable Π as the transcript of π on input X, Y. More generally, we define Π_t as the transcript of π_t on input X, Y. We stress that the inputs X, Y and the auxiliary random variables $\Pi, \Pi_1, \Pi_2, \Pi_3, \ldots$ are not part of P and in particular do not appear in any $P_{\leqslant t}$. Observe also that Π_t is independent of P given X, Y, π_t.

5.2. Accuracy analysis. The focal point of the proof is the random walk

$$\begin{aligned}\varepsilon &\preceq W_1\\ &\preceq W_1W_2\\ &\preceq W_1W_2W_3\\ &\preceq \ldots\\ &\preceq W_1W_2\ldots W_t\\ &\preceq \ldots \end{aligned} \tag{5.2}$$

in the protocol tree for π. We start by studying how accurately this random walk models the actual protocol transcript, Π. For a fixed string w and any $t \geqslant 1$,

$$\begin{aligned}&\mathbf{P}[W_t\Pi_{t+1} = w \mid X, Y, P_{\leqslant t-1}]\\ &\qquad = \sum_{v \preceq w} \mathbf{P}[W_t = v \mid X, Y, P_{\leqslant t-1}]\,\mathbf{P}[v\Pi_{t+1} = w \mid W_t = v, X, Y, P_{\leqslant t-1}]\\ &\qquad = \sum_{v \preceq w} \mathbf{P}[W_t = v \mid X, Y, P_{\leqslant t-1}]\,\mathbf{P}[\Pi_t = w \mid \Pi_t \succeq v, X, Y, P_{\leqslant t-1}]\\ &\qquad = \mathbf{P}[\Pi_t = w \mid X, Y, P_{\leqslant t-1}] \sum_{v \preceq w} \frac{\mathbf{P}[W_t = v \mid X, Y, P_{\leqslant t-1}]}{\mathbf{P}[\Pi_t \succeq v \mid X, Y, P_{\leqslant t-1}]}\\ &\qquad \leqslant (1+\delta)\,\mathbf{P}[\Pi_t = w \mid X, Y, P_{\leqslant t-1}],\end{aligned} \tag{5.3}$$

where the second step follows from the definition of Π_{t+1} as the transcript of π_{t+1} on input X, Y, with π_{t+1} in turn obtained from π_t by restricting to the protocol subtree rooted at W_t; and the final step uses Theorem 4.1.

The newly derived bound implies that the random walk (5.2) behaves much like the original communication protocol π. Indeed, rewrite (5.3) to obtain

$$\begin{aligned}&\mathbf{P}[W_1W_2\ldots W_t\Pi_{t+1} = w \mid X, Y, P_{\leqslant t-1}]\\ &\qquad\qquad \leqslant (1+\delta)\,\mathbf{P}[W_1W_2\ldots W_{t-1}\Pi_t = w \mid X, Y, P_{\leqslant t-1}].\end{aligned}$$

Passing to expectations with respect to $P_{\leqslant t-1}$,

$$\begin{aligned}&\mathbf{P}[W_1W_2\ldots W_t\Pi_{t+1} = w \mid X, Y]\\ &\qquad\qquad \leqslant (1+\delta)\,\mathbf{P}[W_1W_2\ldots W_{t-1}\Pi_t = w \mid X, Y],\end{aligned}$$

whence by induction

$$\mathbf{P}[W_1W_2\ldots W_t\Pi_{t+1} = w \mid X, Y] \leqslant (1+\delta)^t\,\mathbf{P}[\Pi = w \mid X, Y]. \tag{5.4}$$

This result has a statistical distance interpretation in view of Fact 2.6:

$$\begin{aligned}\mathrm{TV}((X,Y,\Pi),(X,Y,W_1W_2\ldots W_t\Pi_{t+1})) &\leqslant 1-\frac{1}{(1+\delta)^t}\\ &\leqslant 1-(1-\delta)^t\\ &\leqslant t\delta. \qquad (5.5)\end{aligned}$$

Here, $W_1W_2\ldots W_t\Pi_{t+1}$ refers to the concatenation of $W_1,W_2,\ldots,W_t,\Pi_{t+1}$ rather than to the composite random variable $(W_1,W_2,\ldots,W_t,\Pi_{t+1})$. This distinction is essential from the point of view of information-theoretic distance.

5.3. Expected information gain. We will now obtain an upper bound on the progress measure $\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec W_t)$, which plays a critical role in relating the communication requirements of the stochastic process to the information cost of the original protocol π. Since π_{t+1} is by definition the protocol corresponding to the subtree of π_t rooted at W_t, we have

$$\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec W_t)=\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec(W_t\Pi_{t+1}))-\mathbf{E}\,\mathbb{D}^{\pi_{t+1},\mu_t|W_t}_{X,Y}(\prec\Pi_{t+1}), \qquad (5.6)$$

where the shorthand $\mu_t \mid v$ for a string $v\in\{0,1\}^*$ refers to the posterior probability distribution on $\mathscr{X}\times\mathscr{Y}$ obtained from μ_t by conditioning on $\Pi_t\succeq v$. Understanding the two expectations on the right-hand side requires subtle conditioning. As a consequence of (5.3),

$$\mathbf{E}[\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec(W_t\Pi_{t+1}))\mid X,Y,P_{\leqslant t-1}]\leqslant(1+\delta)\,\mathbf{E}[\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec\Pi_t)\mid X,Y,P_{\leqslant t-1}]$$

and hence

$$\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec(W_t\Pi_{t+1}))\leqslant(1+\delta)\,\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec\Pi_t). \qquad (5.7)$$

We now examine the other expectation on the right-hand side of (5.6). We claim that

$$\mathbf{E}[\mathbb{D}^{\pi_{t+1},\mu_t|W_t}_{X,Y}(\prec\Pi_{t+1})\mid P_{\leqslant t}]\geqslant\mathbf{E}[\mathbb{D}^{\pi_{t+1},\mu_{t+1}}_{X,Y}(\prec\Pi_{t+1})\mid P_{\leqslant t}]. \qquad (5.8)$$

Conditioning on $P_{\leqslant t}$ fixes $\pi_t,\mu_t,W_t,\mu_{t+1},\pi_{t+1}$, among other things, which means that the expectation on both sides of this inequality is with respect to random input X,Y and the resulting transcript Π_{t+1} in protocol π_{t+1}. But by definition, the posterior probability distribution of X,Y conditioned on $P_{\leqslant t}$ is μ_{t+1}. The claimed inequality (5.8) now follows from Theorem 2.12. Passing to expectations with respect to $P_{\leqslant t}$, we conclude that

$$\mathbf{E}\,\mathbb{D}^{\pi_{t+1},\mu_t|W_t}_{X,Y}(\prec\Pi_{t+1})\geqslant\mathbf{E}\,\mathbb{D}^{\pi_{t+1},\mu_{t+1}}_{X,Y}(\prec\Pi_{t+1}), \qquad (5.9)$$

which along with (5.6) and (5.7) leads to our sought upper bound on the progress measure in the t-th step of the stochastic process:

$$\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec W_t)\leqslant(1+\delta)\,\mathbf{E}\,\mathbb{D}^{\pi_t,\mu_t}_{X,Y}(\prec\Pi_t)-\mathbf{E}\,\mathbb{D}^{\pi_{t+1},\mu_{t+1}}_{X,Y}(\prec\Pi_{t+1}). \qquad (5.10)$$

As a result,

$$\begin{aligned}
\sum_{i=1}^{t} \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i) &\leqslant \sum_{i=1}^{t} (1+\delta)^{t-i}\, \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i) \\
&\leqslant \sum_{i=1}^{t} (1+\delta)^{t-i+1}\, \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec \Pi_i) \\
&\qquad - \sum_{i=1}^{t} (1+\delta)^{t-i}\, \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_{i+1},\mu_{i+1}}(\prec \Pi_{i+1}) \\
&= (1+\delta)^{t}\, \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_1,\mu_1}(\prec \Pi_1) - \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_{t+1},\mu_{t+1}}(\prec \Pi_{t+1}) \\
&\leqslant (1+\delta)^{t}\, \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_1,\mu_1}(\prec \Pi_1) \\
&= (1+\delta)^{t}\, \mathrm{IC}^{*}_{\mu_1}(\pi_1) \\
&= (1+\delta)^{t}\, \mathrm{IC}_{\mu_1}(\pi_1) \\
&= (1+\delta)^{t} I, \qquad (5.11)
\end{aligned}$$

where the second, fifth, and sixth steps use (5.10), Theorem 2.11, and Theorem 2.10, respectively. An analogous calculation involving a telescoping sum shows that

$$\sum_{i=1}^{t} \left(\mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i) + \delta\, \mathbf{E}\, \mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec \Pi_i)\right) \leqslant (1+2\delta)^{t} I. \qquad (5.12)$$

5.4. Expected time to leaf and communication cost. Using the new upper bound (5.11) on the sum of progress terms, we now show that the random walk (5.2) reaches a leaf reasonably quickly and with high probability has small communication cost. The first t stages of the stochastic process fail to reach a leaf with probability given by

$$\begin{aligned}
&\mathbf{P}[W_t \notin \mathscr{L}(\pi_t)]\\
&\qquad = \mathbf{P}[W_i \notin \mathscr{L}(\pi_i) \text{ for } i = 1, 2, \ldots, t]\\
&\qquad = \prod_{i=1}^{t} \mathbf{P}[W_i \notin \mathscr{L}(\pi_i) \mid W_{i-1} \notin \mathscr{L}(\pi_{i-1})]\\
&\qquad \leqslant \left(\frac{1}{t}\sum_{i=1}^{t} \mathbf{P}[W_i \notin \mathscr{L}(\pi_i) \mid W_{i-1} \notin \mathscr{L}(\pi_{i-1})]\right)^{t}\\
&\qquad \leqslant \left(1 - \frac{1}{c} + \frac{\log(1/\delta)}{t}\sum_{i=1}^{t} \mathbf{E}[\mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i) \mid W_{i-1} \notin \mathscr{L}(\pi_{i-1})]\right)^{t}\\
&\qquad = \left(1 - \frac{1}{c} + \frac{\log(1/\delta)}{t}\sum_{i=1}^{t} \frac{\mathbf{E}\,\mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i)}{\mathbf{P}[W_{i-1} \notin \mathscr{L}(\pi_{i-1})]}\right)^{t}\\
&\qquad \leqslant \left(1 - \frac{1}{c} + \frac{\log(1/\delta)}{t\,\mathbf{P}[W_t \notin \mathscr{L}(\pi_t)]}\sum_{i=1}^{t} \mathbf{E}\,\mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i)\right)^{t}\\
&\qquad \leqslant \left(1 - \frac{1}{c} + \frac{\log(1/\delta)}{t\,\mathbf{P}[W_t \notin \mathscr{L}(\pi_t)]} \cdot (1+\delta)^t I\right)^{t},
\end{aligned}$$

where the third, fourth, and last steps use convexity, Theorem 4.1, and (5.11), respectively, $c > 1$ being the absolute constant from Theorem 4.1. Switching to the statistical distance view, we have shown that

$$\begin{aligned}
&\mathrm{TV}((X, Y, W_1W_2\ldots W_t\Pi_{t+1}), (X, Y, W_1W_2\ldots W_t))\\
&\qquad \leqslant \mathbf{P}[W_t \notin \mathscr{L}(\pi_t)]\\
&\qquad \leqslant \min_{0\leqslant p\leqslant 1}\left\{\left(1 - \frac{1}{c} + \frac{\log(1/\delta)}{tp} \cdot (1+\delta)^t I\right)^{t} + p\right\}\\
&\qquad \leqslant \left(1 - \frac{1}{c} + \frac{3\log(1/\delta)}{t\epsilon} \cdot (1+\delta)^t I\right)^{t} + \frac{\epsilon}{3},
\end{aligned}$$

which along with (5.5) gives

$$\begin{aligned}
&\mathrm{TV}((X, Y, \Pi), (X, Y, W_1W_2\ldots W_t))\\
&\qquad \leqslant \left(1 - \frac{1}{c} + \frac{3\log(1/\delta)}{t\epsilon} \cdot (1+\delta)^t I\right)^{t} + \frac{\epsilon}{3} + t\delta. \qquad (5.13)
\end{aligned}$$

We now examine the communication requirements. By Theorem 4.1, stages $1, 2, \ldots, t$ of the stochastic process have communication cost

$$\sum_{i=1}^{t} |M_i| \leqslant \sum_{i=1}^{t} C_i' + \sum_{i=1}^{t} C_i'' + ct\log\frac{1}{\delta},$$

where the nonnegative random variables C_i', C_i'' obey

$$\begin{aligned}\mathbf{E}\left[\sum_{i=1}^{t} C_i'\right] &\leqslant c\sum_{i=1}^{t}(\mathbf{E}\,\mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec W_i) + \delta\,\mathbf{E}\,\mathbb{D}_{X,Y}^{\pi_i,\mu_i}(\prec \Pi_i)) \\ &\leqslant c(1+2\delta)^t I && \text{by (5.12)}\end{aligned}$$

and

$$\mathbf{P}\left[\sum_{i=1}^{t} C_i'' > 0\right] \leqslant t\delta.$$

By Markov's inequality,

$$\begin{aligned}\mathbf{P}&\left[\sum_{i=1}^{t} |M_i| \geqslant \frac{3}{\epsilon}\cdot c(1+2\delta)^t I + ct\log\frac{1}{\delta}\right] \\ &\leqslant \mathbf{P}\left[\sum_{i=1}^{t} C_i' \geqslant \frac{3}{\epsilon}\cdot c(1+2\delta)^t I\right] + \mathbf{P}\left[\sum_{i=1}^{t} C_i'' > 0\right] \\ &\leqslant \frac{\epsilon}{3} + t\delta. \end{aligned}\tag{5.14}$$

5.5. Final communication protocol. Sections 5.1–5.4 suggest a natural communication protocol π' for simulating π. Specifically, Alice and Bob simulate the stochastic process on their given inputs, terminating the simulation as soon as they have completed T stages or exchanged

$$\frac{3}{\epsilon}\cdot c(1+2\delta)^T I + cT\log\frac{1}{\delta}\tag{5.15}$$

bits of communication (whichever occurs first). The communication transcript $(R_1, R_2, R_3, \ldots, M_1, M_2, M_3, \ldots)$ of this simulation fully determines all the other random variables in (5.1), which are never explicitly communicated. Let E be the event that during the first T stages of the stochastic process, the communication cost exceeds (5.15). Then π' simulates π with respect to μ with error

$$\begin{aligned}&\mathrm{TV}((X,Y,\Pi),(X,Y,W_1W_2\ldots W_T)) + \mathbf{P}[E] \\ &\qquad\leqslant \left(1 - \frac{1}{c} + \frac{3\log(1/\delta)}{T\epsilon}\cdot(1+\delta)^T I\right)^T + \frac{2\epsilon}{3} + 2T\delta\end{aligned}\tag{5.16}$$

by (5.13) and (5.14). The communication cost (5.15) and the simulation error (5.16) are bounded by $O(\frac{I}{\epsilon}\log^2\frac{I}{\epsilon})$ and ϵ, respectively, for $\delta = \Theta(\frac{\epsilon}{I})^3$ and $T = \Theta(\frac{I}{\epsilon}\log\frac{I}{\epsilon})$. This completes the proof of Theorem 5.1.

Acknowledgments

The author is thankful to Pei Wu for stimulating discussions, and to Mark Braverman, Gillat Kol, Anup Rao, and Pei Wu for their feedback on an earlier version on this manuscript.

References

[1] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, 3rd edition, 2008.

[2] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. Information theory methods in communication complexity. In *Proceedings of the Seventeenth Annual IEEE Conference on Computational Complexity* (CCC), pages 93–102, 2002.

[3] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[4] B. Barak, M. Braverman, X. Chen, and A. Rao. How to compress interactive communication. *SIAM J. Comput.*, 42(3):1327–1363, 2013.

[5] B. Bauer, S. Moran, and A. Yehudayoff. Internal compression of protocols to entropy. In *Proceedings of the Nineteenth International Workshop on Randomization and Computation* (RANDOM), pages 481–496, 2015.

[6] M. Braverman. Interactive information complexity. *SIAM J. Comput.*, 44(6):1698–1739, 2015.

[7] M. Braverman and A. Rao. Information equals amortized communication. *IEEE Trans. Information Theory*, 60(10):6058–6069, 2014.

[8] M. Braverman, A. Rao, O. Weinstein, and A. Yehudayoff. Direct products in communication complexity. In *Proceedings of the Fifty-Fourth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), pages 746–755, 2013.

[9] M. Braverman and O. Weinstein. A discrepancy lower bound for information complexity. In *Proceedings of the Sixteenth International Workshop on Randomization and Computation* (RANDOM), pages 459–470, 2012.

[10] J. Brody, H. Buhrman, M. Koucký, B. Loff, F. Speelman, and N. K. Vereshchagin. Towards a reverse Newman's theorem in interactive information complexity. In *Proceedings of the Twenty-Eighth Annual IEEE Conference on Computational Complexity* (CCC), pages 24–33, 2013.

[11] A. Chakrabarti, Y. Shi, A. Wirth, and A. C.-C. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proceedings of the Forty-Second Annual IEEE Symposium on Foundations of Computer Science* (FOCS), pages 270–278, 2001.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley–Interscience, New York, 2nd edition, 2006.

[13] A. Ganor, G. Kol, and R. Raz. Exponential separation of information and communication. In *Proceedings of the Fifty-Fifth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), pages 176–185, 2014.

[14] A. Ganor, G. Kol, and R. Raz. Exponential separation of information and communication for Boolean functions. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing* (STOC), pages 557–566, 2015.

[15] A. Ganor, G. Kol, and R. Raz. Exponential separation of communication and external information. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (STOC), 2016. To appear.

[16] M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed. *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer-Verlag Berlin Heidelberg, 1998.

[17] P. Harsha, R. Jain, D. A. McAllester, and J. Radhakrishnan. The communication complexity of correlation. *IEEE Trans. Information Theory*, 56(1):438–449, 2010.

[18] T. Holenstein. Parallel repetition: Simplification and the no-signaling case. *Theory of Computing*, 5(1):141–172, 2009.

[19] R. Jain, J. Radhakrishnan, and P. Sen. A direct sum theorem in communication complexity via message compression. In *Proc. of the 30th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 300–315, 2003.

[20] R. Jain, P. Sen, and J. Radhakrishnan. Optimal direct sum and privacy trade-off results for quantum and classical communication complexity. Available at http://arxiv.org/abs/0807.1267, 2008.

[21] G. Kol. Interactive compression for product distributions. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (STOC), 2016. To appear.

[22] S. N. Ramamoorthy and A. Rao. How to compress asymmetric communication. In *Proceedings of the Thirtieth Annual IEEE Conference on Computational Complexity* (CCC), pages 102–123, 2015.

[23] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[24] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, October 1948.

[25] A. C.-C. Yao. Some complexity questions related to distributive computing. In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing* (STOC), pages 209–213, 1979.

Appendix A. A concentration bound

The purpose of this appendix is to prove Theorem 2.8, which states that two probability distributions on the leaves of a binary tree are multiplicatively close if the Kullback–Leibler divergence on any root-to-leaf path is small. This result is a minor adaptation of a theorem due to Barak et al. [4], and we closely follow their treatment. We start with some preparatory work.

Proposition A.1. *Let X be a bounded real random variable with $\mathbf{E}\, X = 0$. Then*

$$\mathbf{E}\exp(X) \leqslant \exp\left(\frac{\sup(X^2)}{2}\right).$$

Proposition A.1 is frequently used in the analysis of martingales. Its proof is widely available in the literature, e.g., Alon and Spencer [1, Theorem A.1.17], and is provided below for the reader's convenience.

Proof. Let $a = \sup|X|$. Then

$$\begin{aligned}\exp(x) &= \exp\left(\frac{a-x}{2a}\cdot(-a) + \frac{a+x}{2a}\cdot a\right)\\ &\leqslant \frac{a-x}{2a}\cdot\exp(-a) + \frac{a+x}{2a}\cdot\exp(a), \qquad\qquad -a \leqslant x \leqslant a,\end{aligned}$$

where the second step follows by convexity. Geometrically, this inequality expresses the fact that the exponential function $\exp(x)$ lies at or below the line segment joining the points $(-a, \exp(-a))$ and $(a, \exp(a))$. Passing to expectations,

$$\begin{aligned}\mathbf{E}\exp(X) &\leqslant \mathbf{E}\left[\frac{a-X}{2a}\cdot\exp(-a) + \frac{a+X}{2a}\cdot\exp(a)\right]\\ &= \frac{1}{2}\exp(-a) + \frac{1}{2}\exp(a)\\ &\leqslant \exp\left(\frac{a^2}{2}\right),\end{aligned}$$

where the final step can be verified using

$$\exp(x) = \sum_{i=0}^{\infty}\frac{x^i}{i!}. \qquad \square$$

Following Barak et al., we now recall a martingale-type inequality due to Habib et al. [16]. For the sake of completeness, we include its short proof.

LEMMA A.2 (Habib et al.). *Let $V_1, V_2, \ldots, V_N \in \{0,1\}$ be random variables. Fix $\theta > 0$ and consider a function $\phi\colon \{0,1\}^+ \to \mathbb{R}$ such that*

$$\begin{aligned} &\mathbf{E}[\phi(V_1V_2\ldots V_i) \mid V_1, V_2, \ldots, V_{i-1}] = 0, && i = 1,2,3,\ldots,N,\\ &\sum_{i=0}^{N-1} \max\{\phi(v_1v_2\ldots v_i0)^2, \phi(v_1v_2\ldots v_i1)^2\} \leqslant \theta, && v \in \{0,1\}^N. \end{aligned}$$

Then for every $c \geqslant 0$,

$$\mathbf{P}\left[\sum_{i=1}^{N} \phi(V_1V_2\ldots V_i) \geqslant c\right] \leqslant \exp\left(-\frac{c^2}{2\theta}\right).$$

Proof. Let $a > 0$ be a parameter to be chosen later. We claim that

$$\mathbf{E}\exp\left(a\sum_{i=1}^{N}\phi(V_1V_2\ldots V_i)\right) \leqslant \exp\left(\frac{a^2\theta}{2}\right).$$

The proof is by induction on N. The base case $N = 1$ follows from Proposition A.1. For $N \geqslant 2$,

$$\begin{aligned} &\mathbf{E}\exp\left(a\sum_{i=1}^{N}\phi(V_1V_2\ldots V_i)\right)\\ &\quad= \mathop{\mathbf{E}}_{V_1}\left[\exp(a\phi(V_1)) \mathop{\mathbf{E}}_{V_2,V_3,\ldots,V_N}\left[\exp\left(a\sum_{i=2}^{N}\phi(V_1V_2\ldots V_i)\right)\,\middle|\, V_1\right]\right]\\ &\quad\leqslant \max_{v_1\in\{0,1\}}\left\{\mathop{\mathbf{E}}_{V_2,V_3,\ldots,V_N}\left[\exp\left(a\sum_{i=2}^{N}\phi(V_1V_2\ldots V_i)\right)\,\middle|\, V_1 = v_1\right]\right\}\\ &\qquad\qquad\times \mathop{\mathbf{E}}_{V_1}\exp(a\phi(V_1))\\ &\quad\leqslant \exp\left(\frac{a^2(\theta - \max\{\phi(0)^2,\phi(1)^2\})}{2}\right)\mathop{\mathbf{E}}_{V_1}\exp(a\phi(V_1))\\ &\quad\leqslant \exp\left(\frac{a^2(\theta - \max\{\phi(0)^2,\phi(1)^2\})}{2}\right)\exp\left(\frac{a^2\max\{\phi(0)^2,\phi(1)^2\}}{2}\right)\\ &\quad= \exp\left(\frac{a^2\theta}{2}\right), \end{aligned}$$

where the third and fourth steps follow from the inductive hypothesis. We conclude that

$$\begin{aligned}\mathbf{P}\left[\sum_{i=1}^{N}\phi(V_1V_2\ldots V_i)\geqslant c\right] &= \mathbf{P}\left[\exp\left(a\sum_{i=1}^{N}\phi(V_1V_2\ldots V_i)\right)\geqslant \exp(ac)\right]\\ &\leqslant \mathbf{E}\left[\exp\left(a\sum_{i=1}^{N}\phi(V_1V_2\ldots V_i)\right)\right]\exp(-ac)\\ &\leqslant \exp\left(\frac{a^2\theta}{2}-ac\right),\end{aligned}$$

where the second step uses Markov's inequality. Setting $a = c/\theta$ completes the proof. ∎

One final ingredient that we will need is an inequality that involves the Kullback–Leibler divergence.

Proposition A.3. *For all $p, q \in (0,1)$,*

$$\left(\log\frac{p}{q}\right)^2 \leqslant \frac{\min\{\mathrm{KL}(p\parallel q),\mathrm{KL}(q\parallel p)\}}{(2\ln 2)\min\{p^2,q^2\}}.$$

Proof. Since $\log(p/q) = -\log(q/p)$, it is sufficient to consider the case $p \geqslant q$. We have:

$$\left(\log\frac{p}{q}\right)^2 \leqslant \frac{1}{\ln^2 2}\left(\frac{p}{q}-1\right)^2 = \frac{(p-q)^2}{q^2\ln^2 2} \leqslant \frac{\min\{\mathrm{KL}(p\parallel q),\mathrm{KL}(q\parallel p)\}}{(2\ln 2)q^2},$$

where the first step follows from basic calculus and the final step uses Pinsker's inequality (Fact 2.5). ∎

We are now in a position to prove the desired result, stated earlier as Theorem 2.8. Fix a binary tree T and let μ be a probability distribution on the leaves of T. Recall that we identify the vertices of T with binary strings in the usual manner: the root corresponds to the empty string ε, and inductively the left child and right child of a vertex v correspond to $v0$ and $v1$, respectively. For a vertex v of the tree, which can be either a leaf or an internal vertex, we let $\mu(v)$ stand for the probability of reaching a leaf in the subtree of v. Similarly, $\mu(v \mid u)$ denotes the probability of reaching a leaf in the subtree of v conditioned on reaching a leaf in the subtree of u.

Theorem (restatement of Theorem 2.8). *Let μ and $\tilde{\mu}$ be probability distributions on the leaves of a binary tree. For an internal vertex v, abbreviate*

$$\mathbb{D}(v) = \mathrm{KL}(\mu(v0\mid v)\parallel\tilde{\mu}(v0\mid v)).$$

Assume that:

(i) $\mu(v0\mid v), \tilde{\mu}(v0\mid v) \in [1/3, 2/3]$ *for every internal vertex v;*

(ii) $\sum_{u:u\prec v}\mathbb{D}(u) \leqslant \theta$ *for every leaf v.*

Then:

$$\mathop{\mathbf{P}}_{V\sim\mu}\left[\mu(V)\geqslant 2^{c+\theta}\tilde\mu(V)\right]\leqslant\exp\left(-\frac{c^2}{52\theta}\right),\qquad c\geqslant 0,\tag{A.1}$$

$$\mathop{\mathbf{P}}_{V\sim\tilde\mu}\left[\tilde\mu(V)\geqslant 2^{c+(21/20)\theta}\mu(V)\right]\leqslant\exp\left(-\frac{c^2}{55\theta}\right),\qquad c\geqslant 0.\tag{A.2}$$

Proof. Without loss of generality, we may assume that T is a full binary tree of height N. In what follows, the random variable $V=V_1V_2\dots V_N$ stands for a tree leaf distributed according to μ. By hypothesis,

$$\sum_{i=0}^{N-1}\mathbb{D}(v_1\dots v_i)\leqslant\theta\tag{A.3}$$

for every leaf $v\in\{0,1\}^N$. Define $\phi\colon\{0,1\}^+\to\mathbb{R}$ by

$$\phi(v_1v_2\dots v_i)=\log\frac{\mu(v_1\dots v_i\mid v_1\dots v_{i-1})}{\tilde\mu(v_1\dots v_i\mid v_1\dots v_{i-1})}-\mathbb{D}(v_1\dots v_{i-1})$$

for $v_1,v_2,\dots,v_i\in\{0,1\}$. We proceed to verify the zero expectation and boundedness properties that are required for an appeal to Lemma A.2. For $i=1,2,\dots,N$, we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{V\sim\mu}[\phi(V_1\dots V_i)\mid V_1\dots V_{i-1}]\\&\quad=\mathop{\mathbf{E}}_{V\sim\mu}\left[\log\frac{\mu(V_1\dots V_i\mid V_1\dots V_{i-1})}{\tilde\mu(V_1\dots V_i\mid V_1\dots V_{i-1})}\;\middle|\;V_1\dots V_{i-1}\right]-\mathbb{D}(V_1\dots V_{i-1})\\&\quad=\mathrm{KL}(\mu(V_1\dots V_{i-1}0\mid V_1\dots V_{i-1})\,\|\,\tilde\mu(V_1\dots V_{i-1}0\mid V_1\dots V_{i-1}))\\&\qquad\qquad-\mathbb{D}(V_1\dots V_{i-1})\\&\quad=0.\end{aligned}\tag{A.4}$$

For all $v_1, v_2, \ldots, v_N \in \{0,1\}$,

$$\begin{aligned}
&\sum_{i=0}^{N-1} \max\{\phi(v_1 \ldots v_i 0)^2, \phi(v_1 \ldots v_i 1)^2\} \\
&\quad = \sum_{i=0}^{N-1} \max_{b\in\{0,1\}} \left(\log \frac{\mu(v_1 \ldots v_i b \mid v_1 \ldots v_i)}{\tilde{\mu}(v_1 \ldots v_i b \mid v_1 \ldots v_i)} - \mathbb{D}(v_1 \ldots v_i)\right)^2 \\
&\quad \leqslant 4 \sum_{i=0}^{N-1} \max_{b\in\{0,1\}} \left(\log \frac{\mu(v_1 \ldots v_i b \mid v_1 \ldots v_i)}{\tilde{\mu}(v_1 \ldots v_i b \mid v_1 \ldots v_i)}\right)^2 \\
&\quad \leqslant \frac{18}{\ln 2} \sum_{i=0}^{N-1} \max_{b\in\{0,1\}} \mathrm{KL}(\mu(v_1 \ldots v_i b \mid v_1 \ldots v_i) \,\|\, \tilde{\mu}(v_1 \ldots v_i b \mid v_1 \ldots v_i)) \\
&\quad = \frac{18}{\ln 2} \sum_{i=0}^{N-1} \mathrm{KL}(\mu(v_1 \ldots v_i 0 \mid v_1 \ldots v_i) \,\|\, \tilde{\mu}(v_1 \ldots v_i 0 \mid v_1 \ldots v_i)) \\
&\quad \leqslant \frac{18\theta}{\ln 2},
\end{aligned} \tag{A.5}$$

where the first inequality uses the fact that $\sup((X - \mathbf{E}\, X)^2) \leqslant 4 \sup(X^2)$ for any real random variable X; the second inequality follows from Proposition A.3; and the final inequality is immediate by part (ii) of the theorem hypothesis. By (A.4), (A.5), and Lemma A.2,

$$\mathop{\mathbf{P}}_{V\sim\mu}\left[\sum_{i=1}^{N} \phi(V_1 V_2 \ldots V_i) \geqslant c\right] \leqslant \exp\left(-\frac{c^2 \ln 2}{36\theta}\right) \tag{A.6}$$

for $c \geqslant 0$. Hence,

$$\begin{aligned}
\mathop{\mathbf{P}}_{V\sim\mu}\left[\mu(V) \geqslant 2^{c+\theta}\tilde{\mu}(V)\right] &= \mathop{\mathbf{P}}_{V\sim\mu}\left[\log \frac{\mu(V)}{\tilde{\mu}(V)} \geqslant c + \theta\right] \\
&= \mathop{\mathbf{P}}_{V\sim\mu}\left[\sum_{i=1}^{N} (\phi(V_1 \ldots V_i) + \mathbb{D}(V_1 \ldots V_{i-1})) \geqslant c + \theta\right] \\
&\leqslant \mathop{\mathbf{P}}_{V\sim\mu}\left[\sum_{i=1}^{N} \phi(V_1 \ldots V_i) \geqslant c\right] \\
&\leqslant \exp\left(-\frac{c^2 \ln 2}{36\theta}\right),
\end{aligned}$$

where the last two steps use part (ii) of the theorem hypothesis and (A.6), respectively. Now (A.1) follows directly, whereas (A.2) follows by interchanging the roles of μ and $\tilde{\mu}$ and using (2.2). ∎

Appendix B. Protocol balancing

The purpose of this appendix is to present a proof of Theorem 2.15 on protocol balancing, due to Barak et al. [4] and Kol [21]. We start with some information-theoretic preliminaries. The *entropy* of a random variable X supported on a finite set $\mathscr{X}$ is given by

$$H(X)=\sum_{x\in\mathscr{X}}\mathbf{P}[X=x]\log\frac{1}{\mathbf{P}[X=x]}.$$

The entropy $H(X)$ is by definition nonnegative and measures the amount of uncertainty in X, with its maximum value $\log|\mathscr{X}|$ achieved when X is distributed uniformly in $\mathscr{X}$. Recall that we identify a real number $0\leqslant p\leqslant 1$ with the Bernoulli distribution $(p,1-p)$, resulting in the shorthand

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

For random variables X and Y governed by some joint probability distribution on the Cartesian product $\mathscr{X}\times\mathscr{Y}$ of their respective domains, the *conditional entropy of X given Y* is defined to be

$$H(X\mid Y)=\sum_{y\in\mathscr{Y}}\mathbf{P}[Y=y]H(X\mid Y=y).$$

The conditional entropy $H(X\mid Y)$ never exceeds the original entropy $H(X)$. Moreover, the drop in the entropy of X as a result of conditioning on Y is always the same as the drop in the entropy of Y as a result of conditioning on X, and it is equal to the mutual information of X and Y:

$$\begin{aligned}I(X;Y)&=H(X)-H(X\mid Y)\\&=H(Y)-H(Y\mid X).\end{aligned}$$

In particular,

$$\begin{aligned}&I(X;Y)\leqslant H(X),\\&I(X;Y)\leqslant H(Y).\end{aligned}$$

We are now ready to present a proof of Theorem 2.15, adapted from [4, 21].

Theorem (restatement of Theorem 2.15). *Let π be a private-coin protocol with input space $\mathscr{X}\times\mathscr{Y}$. Let μ be a probability distribution on $\mathscr{X}\times\mathscr{Y}$. Then for every $\beta>0$ and $\epsilon>0$, there exists a private-coin β-balanced protocol π' such that*

$$\pi'\hookrightarrow_{\mu,\epsilon}\pi, \tag{B.1}$$

$$\mathrm{IC}_\mu(\pi')\leqslant\mathrm{IC}_\mu(\pi)+\epsilon, \tag{B.2}$$

$$\mathrm{IC}^*_\mu(\pi')\leqslant\mathrm{IC}^*_\mu(\pi)+\epsilon. \tag{B.3}$$

Proof (adapted from [4, 21]). For inputs $x\in\mathscr{X}$ and $y\in\mathscr{Y}$ and an internal vertex v of the protocol tree for π, define

$$\pi(v,x,y)=\begin{cases}\pi(v,x) & \text{if } v\in\mathscr{A},\\ \pi(v,y) & \text{if } v\in\mathscr{B}.\end{cases}$$

Recall that execution of π on input x, y corresponds to a random walk from the root to a leaf of the protocol tree. When the walk reaches an internal vertex v, the owner of that vertex directs the walk to the left subtree with probability $\pi(v, x, y)$ and to the right subtree with probability $1 - \pi(v, x, y)$, by sending 0 and 1, respectively. In the new protocol π', Alice and Bob simulate this random walk one edge at a time. On reaching an internal vertex v at level i, the owner of that vertex sends the message

$$(\Pi_i \oplus R_{v,1}, \Pi_i \oplus R_{v,2}, \ldots, \Pi_i \oplus R_{v,k}), \tag{B.4}$$

where $k = k(|\pi|, \epsilon, \beta) \gg 1$ is a large enough integer and $\Pi_i, R_{v,1}, R_{v,2}, \ldots, R_{v,k}$ are independent Bernoulli variables distributed according to

$$\Pi_i = \begin{cases} 0 & \text{with probability } \pi(v, x, y), \\ 1 & \text{with probability } 1 - \pi(v, x, y) \end{cases}$$

and

$$R_{v,j} = \begin{cases} 0 & \text{with probability } \frac{1}{2} + \beta, \\ 1 & \text{with probability } \frac{1}{2} - \beta. \end{cases}$$

If the majority of the bits in (B.4) are 0, Alice and Bob move to the left child of v. Otherwise, they move to the right child of v.

The described protocol is clearly β-balanced. To verify (B.1), define $G_v = \mathrm{MAJ}(R_{v,1}, R_{v,2}, \ldots, R_{v,k})$ and $G = \bigvee_{v \in \mathscr{V}(\pi)} G_v$. Taking k sufficiently large ensures that

$$\mathbf{P}[G = 0] \geqslant 1 - \epsilon.$$

Conditioned on the event that $G = 0$, the simulated random walk on the protocol tree of π has the same distribution on any given input as the original random walk, settling (B.1).

We now examine the information cost. Let X and Y be random inputs to the protocol with joint distribution μ, and let $\Pi_i, R_{v,j}, G$ be as defined above. Abbreviate $R = (\ldots, R_{v,j}, \ldots) \in \{0,1\}^{|\mathscr{V}(\pi)| \cdot k}$. Then

$$\begin{aligned} \mathrm{IC}^*_\mu(\pi') &\leqslant I(XY; GR\Pi_0\Pi_1\Pi_2\ldots) \\ &\leqslant H(G) + \mathbf{P}[G=0]\, I(XY; R\Pi_0\Pi_1\Pi_2\ldots \mid G = 0) \\ &\qquad\quad + \mathbf{P}[G=1]\, I(XY; R\Pi_0\Pi_1\Pi_2\ldots \mid G = 1) \\ &\leqslant H(G) + \mathbf{P}[G=0]\, I(XY; R\Pi_0\Pi_1\Pi_2\ldots \mid G = 0) \\ &\qquad\quad + \mathbf{P}[G=1]\, H(XY). \end{aligned} \tag{B.5}$$

Conditioning on $G = 0$ has the following consequences: (i) the string $\Pi_0\Pi_1\Pi_2\ldots$ becomes distributed identically to the transcript of π on input X and Y; and (ii) the random string R becomes independent of $X, Y, \Pi_0\Pi_1\Pi_2\ldots$. Therefore,

$$\begin{aligned}
I(XY; &R\Pi_0\Pi_1\Pi_2\ldots \mid G = 0) \\
&= I(XY; \Pi_0\Pi_1\Pi_2\ldots \mid G = 0) + I(XY; R \mid G = 0, \Pi_0\Pi_1\Pi_2\ldots) \\
&\leqslant I(XY; \Pi_0\Pi_1\Pi_2\ldots \mid G = 0) + I(XY\Pi_0\Pi_1\Pi_2\ldots; R \mid G = 0) \\
&= I(XY; \Pi_0\Pi_1\Pi_2\ldots \mid G = 0) \\
&= \mathrm{IC}^*_\mu(\pi). \qquad \text{(B.6)}
\end{aligned}$$

We conclude from (B.5) and (B.6) that

$$\mathrm{IC}^*_\mu(\pi') \leqslant H(G) + \mathbf{P}[G = 0]\,\mathrm{IC}^*_\mu(\pi) + \mathbf{P}[G = 1]H(XY).$$

The quantities $\mathbf{P}[G = 1]$ and $H(G)$ can be made arbitrarily small by taking k sufficiently large. This completes the proof of (B.3). The proof of (B.2) is closely analogous. ∎