The Complexity of Agreement*

Scott Aaronson†

ABSTRACT

A celebrated 1976 theorem of Aumann asserts that Bayesian agents with common priors can never "agree to disagree": if their opinions about any topic are common knowledge, then those opinions must be equal. But two key questions went unaddressed: first, can the agents reach agreement after a conversation of reasonable length? Second, can the computations needed for that conversation be performed efficiently? This paper answers both questions in the affirmative, thereby strengthening Aumann's original conclusion.

We show that for two agents with a common prior to agree within ε about the expectation of a $[0,1]$ variable with high probability over their prior, it suffices for them to exchange $O\left(1/\varepsilon^2\right)$ bits. This bound is completely independent of the number of bits n of relevant knowledge that the agents have. We also extend the bound to three or more agents; and we give an example where the "standard protocol" (which consists of repeatedly announcing one's current expectation) nearly saturates the bound, while a new "attenuated protocol" does better. Finally, we give a protocol that would cause two Bayesians to agree within ε after exchanging $O\left(1/\varepsilon^2\right)$ messages, and that can be *simulated* by agents with limited computational resources. By this we mean that, after examining the agents' knowledge and a transcript of their conversation, no one would be able to distinguish the agents from perfect Bayesians. The time used by the simulation procedure is exponential in $1/\varepsilon^6$ but not in n.

Categories and Subject Descriptors

F.0 [**Theory of Computation**]: General

General Terms

Theory, Economics

Keywords

communication, bounded rationality, agreement, common priors, Bayesian agents, random walks

*Extended abstract. For the full version, please visit www.scottaaronson.com/papers.

†Institute for Advanced Study, Princeton. Email: aaronson@ias.edu. Most of this work was done while the author was a graduate student at UC Berkeley.

STOC'05, May 22-24, 2005, Baltimore, Maryland, USA.

1. INTRODUCTION

> "Not only do people disagree; they often consider their disagreements to be about what is objectively true, rather than about how they each feel or use words. Furthermore, people often consider their disagreements to be honest, meaning that the disputants respect each other's relevant abilities, and consider each person's stated opinion to be his best estimate of the truth, given his information and effort. Yet according to well-known theory, such honest disagreement is impossible." —Cowen and Hanson [6, p.3]

Suppose Alice and Bob are Bayesian agents, who have the same prior probability distribution over a finite set of possible states of the world. They are interested in estimating the population of New York City, and they have gathered different pieces of evidence related to that variable. Let E_A be Alice's expectation of the population conditioned on her evidence, and let E_B be Bob's expectation conditioned on his evidence. Of course E_A might differ from E_B. However, suppose E_A and E_B are *common knowledge* between Alice and Bob—meaning that both agents know the values of E_A and E_B, both know that they both know the values, both know that they both know that they both know the values, and so on. Then a 1976 theorem of Aumann asserts that $E_A = E_B$. In other words, Alice and Bob cannot "agree to disagree"; common knowledge of each other's opinions implies that their opinions are equal.

This counterintuitive theorem has already had an indirect impact on computer science, for it inspired the fields of interactive epistemology and reasoning about knowledge [12]. However, Aumann's theorem itself has so far mostly been studied in the economics and philosophy literature. Numerous papers have been written generalizing the theorem—for example, to three or more agents [16], to certain types of non-Bayesian agents [5], to approximate common knowledge [14], and to infinite state spaces [4]. Other papers have examined the justification for the Common Prior Assumption

[2, 6, 8, 11], the assumption that Alice and Bob start out with the same prior before they receive different evidence.

However, if we wish to draw a "real-world" moral from Aumann's theorem—for example, that human beings are not well approximated by Bayesian agents with common priors—then there is a crucial further issue to consider, namely *complexity*. To achieve common knowledge, it is not enough for Alice and Bob simply to exchange E_A and E_B. For Alice's expectation might change when she learns E_B, and Bob's expectation might change when he learns E_A. Then Alice and Bob would have to exchange their new expectations, and so on indefinitely. Provided the probability space is finite, Geanakoplos and Polemarchakis [7] showed that the resulting protocol must terminate after a finite number of messages, with Alice and Bob having the same expectation $E_A = E_B$. But even setting aside the fact that the messages in this protocol are unbounded-precision real numbers, no nontrivial upper bound was known on the number of messages needed to reach agreement or approximate agreement.

Of course, if communication cost is not an issue, then Alice and Bob might as well just exchange all of their evidence. Since they have the same prior probabilities, this will immediately cause them to agree about everything (and indeed, it will be common knowledge between them that they agree). The trouble is that "evidence" might include every piece of information to which the agents were ever exposed, in which case exchanging the evidence would take a prohibitively long time.

When we began studying this topic, our conjecture was that regardless of what protocol they used, Alice and Bob would in general need to exchange $\Omega(n)$ bits to approximately agree with high probability, where n is the total number of bits that they know. The intuition for this conjecture came from ordinary communication complexity, where it is easy to construct functions (such as the Boolean inner product function) for which a linear amount of communication is necessary. Indeed, communication cost seemed like it would effectively nullify Aumann's theorem for agents subject to realistic constraints. By analogy, several counterintuitive results in game theory (for example, the all-defect equilibrium in the Iterated Prisoners' Dilemma) can be nullified by assuming the players' resources are bounded [15].

Independently of communication cost, it seemed obvious that *computation* cost would provide another barrier to agreement. For just to calculate her opinions, Alice might need to take expectations over sets of size $2^{\Omega(n)}$ that are constantly being updated in response to messages from Bob. This seems hopeless, even if we assume Alice has the ability to sample efficiently from the common prior.

1.1 Our Results

This paper initiates the study of the communication complexity and computational complexity of agreement protocols. Its surprising conclusion is that, in spite of the above arguments, complexity is *not* a fundamental barrier to agreement. In our view, this conclusion closes a major gap between Aumann's theorem and its informal interpretation, by showing that agreeing to disagree is problematic not merely "in the limit" of common knowledge, but even for agents subject to realistic constraints on communication and computation.

From a computer science perspective, the main novelty of the paper will be that, when we analyze the communication complexity of a function f, we care only about how long it takes some set of agents to agree *among themselves* about the expectation of f. Whether the agents' expectations agree with the true value of f is irrelevant.

After introducing notation and definitions in Section 2, in Section 3 we study the "standard protocol," introduced by Geanakoplos and Polemarchakis [7] and alluded to earlier. In that protocol, Alice and Bob repeatedly announce their current expectations E_A and E_B, conditioned on all previous announcements. A curious feature of this protocol is that the agents only exchange their opinions, not the evidence on which their opinions are based. Yet an opinion, coming from an honest Bayesian agent, turns out to serve as a powerful summary of everything that agent knows.

The question we ask is how many messages are needed before the agents' expectations agree within ε with probability at least $1-\delta$ over their prior, given parameters ε and δ. We show that $1/\left(\delta\varepsilon^2\right)$ messages suffice. We then show that $O\left(1/\left(\delta\varepsilon^2\right)\right)$ messages still suffice, if instead of sending their whole expectations (which are real numbers), the agents send "summary" messages consisting of only 2 bits each. What makes these upper bounds surprising is that they are completely independent of n, the number of bits needed to represent the agents' knowledge. By contrast, in ordinary communication complexity (see [13]), it is easy to show that given a random function $f : \{0,1\}^n \times \{0,1\}^n \to [0,1]$, Alice and Bob would need to exchange $\Omega(n)$ bits to approximate f to within (say) $1/10$ with high probability.

Intuitively, the key point is that Alice's and Bob's expectations follow an unbiased random walk [9], which has absorbing barriers at 0 and 1 since the range of f is $[0,1]$. Furthermore, the step size of this walk is proportional to the amount by which Alice and Bob disagree. So for example, if the agents disagreed by ε with certainty at every time step, then the walk would hit one of the absorbing barriers after an expected number of steps $O\left(1/\varepsilon^2\right)$. The actual proof will formalize this idea by defining a progress measure based on the L_2-norm, and then analyzing the rate at which this measure increases from 0 to 1.

Given the results of Section 3, several questions demand our attention. Is the upper bound of $1/\left(\delta\varepsilon^2\right)$ bits tight, or can it be improved even further? Also, is the standard protocol always optimal, or do other protocols sometimes need even less communication? Section 3.2 addresses these questions. Though we are unable to show any lower bound better than $\Omega(\log 1/\varepsilon)$ that applies to *all* protocols, we do give examples where both the continuous and discretized standard protocols need $\Omega\left(1/\varepsilon^2\right)$ messages. We also show that the standard protocol is not optimal: for all ε there exists a scenario where the standard protocol needs $\Omega\left(1/\varepsilon^2\right)$ bits, while a new protocol (which we call the *attenuated protocol*) uses fewer bits.

In earlier work, Parikh and Krasucki [16] extended the Geanakoplos-Polemarchakis protocol to three or more agents, who send messages along the edges of a directed graph. Thus, it is natural to ask whether our complexity results extend to this setting as well. Section 3.1 shows that they do: given N agents with a common prior, who send messages along a strongly connected graph of diameter d, order $Nd^2/\left(\delta\varepsilon^2\right)$ messages suffice for every pair of agents to agree within ε about the expectation of a $[0,1]$ random variable with probability at least $1-\delta$ over their prior.

In Section 4 we shift attention to the *computational* complexity of agreement, the subject of our deepest technical result. What we want to show is that, even if two agents are computationally bounded, after a conversation of reasonable length they can still probably approximately agree about the expectation of a $[0,1]$ random variable. A large part of the problem is to say what this even means. After all, if the agents both ignored their evidence and estimated (say) $1/2$, then they would agree before exchanging even a single message. So agreement is only interesting if the agents have made some sort of "good-faith effort" to emulate Bayesian rationality.

Though it is unclear exactly what sort of effort is necessary, we do propose a criterion that we think is *sufficient*. This is that the agents be able to *simulate* a Bayesian agreement protocol, in such a way that a computationally-unbounded referee, given the agents' knowledge together with a transcript of their conversation, be unable to decide (with non-negligible bias) whether the agents are computationally bounded or not. By analogy to the Turing test for intelligence, we would argue that a statistically perfect simulation of Bayesian rationality *is* Bayesian rationality.

But what do we mean by computationally-bounded agents? We discuss this question in detail in Section 4, but the basic point is that we assume two "subroutines": one that computes the $[0,1]$ variable of interest, given a state of the world ω; and another that samples a state ω from any set in either agent's initial knowledge partition. The complexity of the simulation procedure is then expressed in terms of the number of calls to these subroutines.

Unfortunately, there is no way to simulate the standard protocol—even our discretized version of it—using a small number of subroutine calls. The reason is that Alice's ideal estimate p might lie on a "knife-edge" between the set of estimates that would cause her to send message m_1 to Bob, and the set that would cause her to send a different message m_2. In that case, it does not suffice for her to approximate p using random sampling; she needs to determine it exactly. Our solution, which we develop in Section 4.1, is to have the agents "smooth" their messages by adding random noise to them. By hiding small errors in the agents' estimates, such noise makes the knife-edge problem disappear. On the other hand, in the computationally-unbounded case, the noise does not prevent the agents from agreeing within ε with probability $1-\delta$ after $O\left(1/\left(\delta\varepsilon^2\right)\right)$ messages. Our main result is that the smoothed standard protocol can be simulated using a number of subroutine calls that depends only on ε and δ, not on n. The dependence, unfortunately, is exponential in $1/\left(\delta^3\varepsilon^6\right)$, but we expect that both the procedure and its analysis could be improved.

We conclude in Section 5 with some open problems.

2. PRELIMINARIES

Let Ω be a set of possible states of the world. Throughout this paper, Ω will be finite for simplicity of presentation. Let $\mathcal{D}$ be a prior probability distribution over Ω that is shared by some set of agents. We assume $\mathcal{D}$ assigns nonzero probability to every $\omega \in \Omega$, for if not, we simply remove the probability-0 states from Ω. We can identify any subset $S \subseteq \Omega$ with the event (or proposition) that $\omega \in S$. Whenever we talk about a probability or expectation over a subset $S \subseteq \Omega$, unless otherwise indicated we mean that we start from $\mathcal{D}$ and then conditionalize on S.

Let $\mathcal{I}$ be a set of agents; except in Section 3.1, $\mathcal{I}$ will consist of exactly two agents, Alice (A) and Bob (B). We will consider protocols in which agents $i \in \mathcal{I}$ send messages to each other in some order. Let $\Pi_i^t(\omega)$ be the set of states that agent i considers possible immediately after the t^{th} message has been sent, given that the true state of the world is ω.[1] Then $\omega \in \Pi_i^t(\omega) \subseteq \Omega$, and indeed the set $\left\{\Pi_i^t(\omega)\right\}_{\omega\in\Omega}$ (where we might have $\Pi_i^t(\omega) = \Pi_i^t(\omega')$ for distinct ω,ω') forms a partition of Ω. Furthermore, since the agents never forget messages, we have $\Pi_i^t(\omega) \subseteq \Pi_i^{t-1}(\omega)$. Thus we say that the partition $\left\{\Pi_i^t(\omega)\right\}_{\omega\in\Omega}$ refines $\left\{\Pi_i^{t-1}(\omega)\right\}_{\omega\in\Omega}$, or equivalently that $\left\{\Pi_i^{t-1}\right\}_{\omega\in\Omega}$ coarsens $\left\{\Pi_i^t\right\}_{\omega\in\Omega}$. Notice also that if the t^{th} message is not sent to i, then $\Pi_i^t(\omega) = \Pi_i^{t-1}(\omega)$. As a convention, we will freely omit arguments of ω when doing so will cause no confusion.

As is standard in this field, we assume that the agents know each other's initial partitions $\left\{\Pi_i^0\right\}_{\omega\in\Omega}$. The usual justification (see [1, 3]) is that a state of the world $\omega \in \Omega$ includes the agents' knowledge as part of it. From that assumption one can show that every agent must have a uniquely defined partition known to every other agent.

Let $f : \Omega \to [0,1]$ be a real-valued function that the agents are interested in estimating. The assumption $f \in [0,1]$ is without loss of generality—for since Ω is finite, any function from Ω to $\mathbb{R}$ has a bounded range, which we can take to be $[0,1]$ by rescaling. We can think of $f(\omega)$ as the probability of some future event conditioned on ω, but this is not necessary. By a *scenario*, we will mean a 5-tuple $\left(\Omega, \mathcal{D}, \mathcal{I}, f, \left\{\Pi_i^0(\omega)\right\}_{\omega\in\Omega, i\in\mathcal{I}}\right)$.

Given a subset $S \subseteq \Omega$, let

$$\operatorname*{EX}_S[f] = \frac{1}{\Pr_{\mathcal{D}}[S]} \sum_{\omega\in S} \Pr_{\mathcal{D}}[\omega] f(\omega)$$

be the expectation of f over S. Then let

$$E_i^t(\omega) = \operatorname*{EX}_{\Pi_i^t(\omega)}[f]$$

be agent i's expectation of f at step t, given that the true state of the world is ω. Note that E_i^t will *always* mean E_i at step t, not E_i to the t^{th} power. Also, let

$$\Theta_i^t(\omega) = \left\{\omega' : E_i^t(\omega') = E_i^t(\omega)\right\}$$

be the set of states for which i's expectation of f equals $E_i^t(\omega)$. Then the partition $\left\{\Theta_i^t\right\}_{\omega\in\Omega}$ coarsens $\left\{\Pi_i^t\right\}_{\omega\in\Omega}$, and $E_i^t(\omega) = \operatorname{EX}_{\Theta_i^t(\omega)}[f]$.

To build intuition, let us state some simple but important observations due to Hanson [9]. First, since $\left\{\Pi_i^t\right\}_{\omega\in\Omega}$ coarsens $\left\{\Pi_i^{t+1}\right\}_{\omega\in\Omega}$, we have

$$\operatorname*{EX}_{\Pi_i^t(\omega)}\left[E_i^{t+1}\right] = \operatorname*{EX}_{\omega'\in\Pi_i^t(\omega)}\left[\operatorname*{EX}_{\Pi_i^{t+1}(\omega')}[f]\right] = \operatorname*{EX}_{\Pi_i^t(\omega)}[f] = E_i^t(\omega).$$

This says that i's expectation of its own future expectation of f always equals its current expectation. Second, if i has just communicated its expectation of f to j, then $\left\{\Theta_i^t\right\}_{\omega\in\Omega}$

[1] We assume for now that messages are "noise-free"; that is, they partition the state space sharply. Later we will remove this assumption.

coarsens $\left\{\Pi_j^t\right\}_{\omega\in\Omega}$, and therefore

$$\underset{\Theta_i^t(\omega)}{\mathrm{EX}}\left[E_j^t\right] = \underset{\omega'\in\Theta_i^t(\omega)}{\mathrm{EX}}\left[\underset{\Pi_j^t(\omega')}{\mathrm{EX}}[f]\right] = \underset{\Theta_i^t(\omega)}{\mathrm{EX}}[f] = E_i^t(\omega).$$

Consider a outsider who has the same prior as i and j, and who sees their messages but not their inputs. Then the above equation says that, after i sends the message $E_i^t(\omega)$ to j, the outsider's expectation of j's new expectation of f is simply $E_i^t(\omega)$ itself.

2.1 (ε,δ)-Agreement

The basic goal of an agreement protocol is to cause Alice and Bob to agree about the expectation of f, meaning that $E_A^t = E_B^t$. However, it turns out that in order to prove anything nontrivial, we will need to relax the success condition to *probabilistic* and *approximate* agreement. More formally, we say that Alice and Bob (ε,δ)-*agree* after the t^{th} message if

$$\Pr_{\omega\in\mathcal{D}}\left[\left|E_A^t(\omega) - E_B^t(\omega)\right| > \varepsilon\right] \leq \delta.$$

In other words, the agents' expectations of f should agree to within ε, with probability at least $1-\delta$ over ω drawn from the common prior. We will be interested in the minimum t such that Alice and Bob (ε,δ)-agree after the t^{th} message.

Let us first remark that (ε,δ)-agreement is arguably a more fundamental notion than exact agreement. For since f can take arbitrary values in $[0,1]$, in general Alice and Bob need to exchange $2n$ bits (i.e. everything they know) to ensure that $E_A^t = E_B^t$.[2] But this trivial lower bound has less to do with agreement than with the fact that real numbers form a continuum. The real question is, how *closely* can Alice and Bob agree after a short conversation? Also, we should consider a protocol successful if it succeeds with probability $1-\delta$ for any $\delta > 0$, using resources that scale reasonably in $1/\delta$.

But why do we take the probability over $\mathcal{D}$, as opposed to some other distribution? That is, what if Alice's and Bob's priors agree with each other, but not with external reality? Unfortunately, it seems hard to prove anything in that situation, since the "true" prior could be concentrated on a few states that the agents consider vanishingly unlikely. Indeed, we conjecture that there exists a scenario such that for all agreement protocols, Alice and Bob must exchange $\Omega(n)$ bits to agree within ε on every ω (that is, $(\varepsilon,0)$-agree). In any case, it already seems counterintuitive that Alice and Bob can both enter their conversation *expecting* to agree after a short amount of time!

2.2 Communication Complexity

In an *agreement protocol*, Alice and Bob take turns sending messages to each other. Any such protocol is characterized by a sequence of functions $m_1, m_2, \ldots : 2^\Omega \to \mathcal{M}$, known to both agents, which map subsets of Ω to elements of a message space $\mathcal{M}$. Possibilities for $\mathcal{M}$ include $[0,1]$ in a continuous protocol, or $\{0,1\}$ in a discretized protocol. In all protocols considered in this paper, the m_t's will be extremely simple; for example, we might have $m_t(S) = \mathrm{EX}_S[f]$ be the agent's current expectation of f.

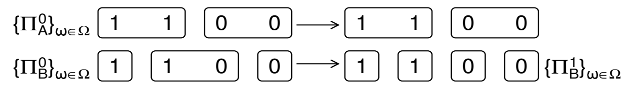

Figure 1: After Alice tells Bob whether E_A^0 is 1 or 0, Bob's partition $\left\{\Pi_B^0\right\}_{\omega\in\Omega}$ is refined to $\left\{\Pi_B^1\right\}_{\omega\in\Omega}$.

The protocol proceeds as follows: first Alice computes $m_1\left(\Pi_A^0(\omega)\right)$ and sends it to Bob. After seeing Alice's message, and assuming the true state of the world is ω, Bob's new set of possible states becomes

$$\Pi_B^1(\omega) = \left\{\omega' \in \Pi_B^0(\omega) : m_1\left(\Pi_A^0(\omega')\right) = m_1\left(\Pi_A^0(\omega)\right)\right\}$$

as in Figure 1. Then Bob computes $m_2\left(\Pi_B^1(\omega)\right)$ and sends it to Alice, whereupon Alice's set of possible states becomes

$$\Pi_A^2(\omega) = \left\{\omega' \in \Pi_A^0(\omega) : m_2\left(\Pi_B^1(\omega')\right) = m_2\left(\Pi_B^1(\omega)\right)\right\}.$$

Then Alice computes $m_3\left(\Pi_A^2(\omega)\right)$ and sends it to Bob, and so on.

Our ending condition is simply that the agents (ε,δ)-agree at *some* step t. We do not require them to fix t independently of the scenario. The reason is that for any t, there might exist perverse scenarios such that the agents nearly agree for the first $t-1$ steps, then disagree violently at the t^{th} step. However, it seems unfair to penalize the agents in such cases.

The following is the best lower bound we are able to show on agreement complexity.

Proposition 1. *For all n, there exists a scenario with $|\Omega| = 2^{2n}$ such that for all $\varepsilon \geq 2^{-n}$ and $\delta \geq 0$, Alice must send $\Omega\left(\log\frac{1-\delta}{\varepsilon}\right)$ bits to Bob and Bob must send $\Omega\left(\log\frac{1-\delta}{\varepsilon}\right)$ bits to Alice before the agents (ε,δ)-agree. In particular, if δ is bounded away from 1 by a constant, then $\Omega(\log 1/\varepsilon)$ bits are needed.*

Proof. Let $\Omega = \{1,\ldots,2^n\}^2$, let $\mathcal{D}$ be uniform over Ω, and let $f((x,y)) = (x+y)/2^{n+1}$ for all $(x,y)\in\Omega$. Thus if $\widehat{x}$ is Bob's expectation of x at step t and $\widehat{y}$ is Alice's expectation of y, then $E_A^t = (x+\widehat{y})/2^{n+1}$ and $E_B^t = (\widehat{x}+y)/2^{n+1}$. Suppose one agent, say Alice, has sent only $t < \log_2\left(\frac{1-\delta}{\varepsilon}\right) - 3$ bits to Bob. For each $i \in \left\{1,\ldots,2^t\right\}$, let p_i be the probability of the i^{th} message sequence from Alice. Then conditioned on i, there are exactly $2^n p_i$ values of x still possible from Bob's point of view, of which at most $2\varepsilon\left(2^{n+1}\right)+1$ could lead to $\left|E_A^t - E_B^t\right| \leq \varepsilon$. So regardless of E_B^t, the probability of $\left|E_A^t - E_B^t\right| \leq \varepsilon$ can be at most

$$\frac{2\varepsilon\left(2^{n+1}\right)+1}{2^n p_i} = \frac{4\varepsilon + 2^{-n}}{p_i} \leq \frac{5\varepsilon}{p_i}$$

since $\varepsilon \geq 2^{-n}$. Therefore the agents agree within ε with total probability at most

$$\sum_{i=1}^{2^t} p_i\left(\frac{5\varepsilon}{p_i}\right) = 5\varepsilon 2^t < 5\varepsilon\left(\frac{1-\delta}{8\varepsilon}\right) < 1-\delta.$$

$\square$

[2] Note the contrast with ordinary communication complexity, where n bits always suffice. Indeed, even to produce approximate agreement, two-way communication is necessary in general, as shown by the example $f((x,y)) = (2x+y)/3$, where $x,y\in\{0,1\}$ are uniformly distributed.

3. CONVERGENCE OF THE STANDARD PROTOCOL

The two-party "standard protocol" is simply the following: first Alice sends E_A^0, her current expectation of f, to Bob. Then Bob sends his expectation E_B^1 to Alice, then Alice sends E_A^2 to Bob, and so on. Geanakoplos and Polemarchakis [7] observed that if the agents use the standard protocol, then after a finite number of messages T, they will $(0,0)$-agree—that is, have $E_A^T(\omega) = E_B^T(\omega)$ for all $\omega \in \Omega$. The reason is simply that until the agents agree, there will always be a message that nontrivially refines one of their partitions, and the partitions can only be refined a finite number of times. Unfortunately, this argument does not yield any upper bound except $O(|\Omega|)$ on the number of messages.

In this section we ask how many messages are needed before the agents (ε,δ)-agree. The surprising and unexpected answer, in Theorem 3, is that $1/(\delta\varepsilon^2)$ messages always suffice, independently of $|\Omega|$ and all other details of a scenario. One might guess that, since the expectations $E_A^0, E_B^1, \ldots$ are real numbers, the cost of communication must be hidden in the length of the messages. However, in Theorem 4 we show that even if the agents send only 2-bit "summaries" of their expectations, $O(1/(\delta\varepsilon^2))$ messages still suffice for (ε,δ)-agreement.

Given any function $F : \Omega \to [0,1]$, let $\|F\|_2^2 = \mathrm{EX}_\Omega[F^2]$. The following proposition will greatly simplify the proofs of Theorems 3 and 4.

Proposition 2. *Suppose the partition $\{\Pi_i^t\}_{\omega\in\Omega}$ refines $\{\Theta_j^u\}_{\omega\in\Omega}$. Then*

$$\|E_i^t\|_2^2 - \|E_j^u\|_2^2 = \|E_i^t - E_j^u\|_2^2$$

so in particular, $\|E_i^t\|_2^2 \geq \|E_j^u\|_2^2$. A special case is that $\|E_i^{t+1}\|_2^2 \geq \|E_i^t\|_2^2$ for all i, t.

Proof. Since $\{\Pi_i^t\}_{\omega\in\Omega}$ refines $\{\Theta_j^u\}_{\omega\in\Omega}$, we have

$$\begin{aligned}\underset{\Omega}{\mathrm{EX}}\left[E_j^u E_i^t\right] &= \underset{\omega\in\Omega}{\mathrm{EX}}\left[E_j^u(\omega) \underset{\omega'\in\Theta_j^u(\omega)}{\mathrm{EX}}\left[E_i^t(\omega')\right]\right] \\ &= \underset{\Omega}{\mathrm{EX}}\left[E_j^u E_j^u\right] = \|E_j^u\|_2^2\end{aligned}$$

by the observations in Section 2, and therefore

$$\begin{aligned}\|E_i^t - E_j^u\|_2^2 &= \|E_i^t\|_2^2 + \|E_j^u\|_2^2 - 2\underset{\Omega}{\mathrm{EX}}\left[E_j^u E_i^t\right] \\ &= \|E_i^t\|_2^2 - \|E_j^u\|_2^2.\end{aligned}$$

□

We can now prove an upper bound on the number of messages needed for agreement.

Theorem 3. *The standard protocol causes Alice and Bob to (ε,δ)-agree after at most $1/(\delta\varepsilon^2)$ messages.*

Proof. We need only track the expectation, not of E_A^t and E_B^t, but of $(E_A^t)^2$ and $(E_B^t)^2$. Suppose Alice sends the t^{th} message. Then Bob's partition $\{\Pi_B^t\}_{\omega\in\Omega}$ refines $\{\Theta_A^{t-1}\}_{\omega\in\Omega}$. It follows by Proposition 2 that

$$\|E_B^t\|_2^2 - \|E_A^{t-1}\|_2^2 = \|E_B^t - E_A^{t-1}\|_2^2.$$

Assuming $\Pr\left[\left|E_B^t - E_A^{t-1}\right| > \varepsilon\right] \geq \delta$, this implies that

$$\|E_B^t\|_2^2 = \|E_A^{t-1}\|_2^2 + \|E_B^t - E_A^{t-1}\|_2^2 > \|E_A^{t-1}\|_2^2 + \delta\varepsilon^2.$$

Similarly, after Bob sends Alice the $(t+1)^{st}$ message, we have $\|E_A^{t+1}\|_2^2 > \|E_B^t\|_2^2 + \delta\varepsilon^2$. So until the agents (ε,δ)-agree, each message increases $\max\left\{\|E_A^t\|_2^2, \|E_B^t\|_2^2\right\}$ by more than $\delta\varepsilon^2$. But the maximum can never exceed 1 (since $E_A^t, E_B^t \in [0,1]$), which yields an upper bound of $1/(\delta\varepsilon^2)$ on the number of messages. □

As mentioned previously, the trouble with the standard protocol is that sending one's expectation might require too many bits. A simple way to discretize the protocol is as follows. Imagine a "monkey in the middle," Charlie, who has the same prior distribution $\mathcal{D}$ as Alice and Bob and who sees all messages between them, but who does not know either of their inputs. In other words, letting $\Pi_C^t(\omega)$ be the set of states that Charlie considers possible after the first t messages, we have $\Pi_C^0(\omega) = \Omega$ for all ω. Then the partition $\{\Pi_C^t\}_{\omega\in\Omega}$ coarsens both $\{\Pi_A^t\}_{\omega\in\Omega}$ and $\{\Pi_B^t\}_{\omega\in\Omega}$; therefore both Alice and Bob can compute Charlie's expectation $E_C^t(\omega) = \mathrm{EX}_{\Pi_C^t(\omega)}[f]$ of f.

Now whenever it is her turn to send a message to Bob, Alice sends the message "high" if $E_A^t > E_C^t + \varepsilon/4$, "low" if $E_A^t < E_C^t - \varepsilon/4$, and "medium" otherwise. This requires 2 bits. Likewise, Bob sends "high" if $E_B^t > E_C^t + \varepsilon/4$, "low" if $E_B^t < E_C^t - \varepsilon/4$, and "medium" otherwise.

Theorem 4. *The discretized protocol described above causes Alice and Bob to (ε,δ)-agree after $O(1/(\delta\varepsilon^2))$ messages.*

Proof. The plan is to show that either $\|E_A^t\|_2^2$, $\|E_B^t\|_2^2$, or $\|E_C^t\|_2^2$ increases by at least $\delta\varepsilon^2/512$ with every message of Alice's, until Alice and Bob (ε,δ)-agree. Since $\|E_i^t\|_2^2 \leq 1$ for all i, this will imply an upper bound of $3072/(\delta\varepsilon^2)$ on the number of messages (of course, we did not optimize the constant). Assume that $\Pr\left[\left|E_A^t - E_B^t\right| > \varepsilon\right] \geq \delta$ and it is Alice's turn to send the $(t+1)^{st}$ message. By the triangle inequality, either

$$\Pr\left[\left|E_A^t - E_C^t\right| > \frac{\varepsilon}{2}\right] \geq \frac{\delta}{2}$$

or

$$\Pr\left[\left|E_B^t - E_C^t\right| > \frac{\varepsilon}{2}\right] \geq \frac{\delta}{2}.$$

We analyze these two cases separately. In the first case, with probability at least $\delta/2$ Alice's message is either "high" or "low." If the message is "high," then E_C^{t+1} becomes an average of numbers each greater than $E_C^t + \varepsilon/4$, so $E_C^{t+1} > E_C^t + \varepsilon/4$. If the message is "low," then likewise $E_C^{t+1} < E_C^t - \varepsilon/4$. Since $\{\Pi_C^{t+1}\}_{\omega\in\Omega}$ refines $\{\Pi_C^t\}_{\omega\in\Omega}$, Proposition 2 thereby gives

$$\|E_C^{t+1}\|_2^2 - \|E_C^t\|_2^2 = \|E_C^{t+1} - E_C^t\|_2^2 > \frac{\delta}{2}\left(\frac{\varepsilon}{4}\right)^2.$$

Now for the second case. If, after Alice sends the $(t+1)^{st}$ message, we still have

$$\Pr\left[\left|E_B^{t+1} - E_C^{t+1}\right| > \frac{\varepsilon}{4}\right] \geq \frac{\delta}{4},$$

then the previous argument applied to Bob implies that

$$\|E_C^{t+2}\|_2^2 - \|E_C^{t+1}\|_2^2 > \frac{\delta}{4}\left(\frac{\varepsilon}{4}\right)^2$$

and we are done. So suppose otherwise. Then the difference between Bob's and Charlie's expectations must have changed significantly:

$$\Pr\left[\left|E_B^t - E_C^t\right| - \left|E_B^{t+1} - E_C^{t+1}\right| > \frac{\varepsilon}{4}\right] > \frac{\delta}{4}.$$

Hence by the triangle inequality (again), either

$$\Pr\left[\left|E_B^{t+1} - E_B^t\right| > \frac{\varepsilon}{8}\right] > \frac{\delta}{8}$$

or

$$\Pr\left[\left|E_C^{t+1} - E_C^t\right| > \frac{\varepsilon}{8}\right] > \frac{\delta}{8}.$$

In the former case, Proposition 2 yields

$$\left\|E_B^{t+1}\right\|_2^2 - \left\|E_B^t\right\|_2^2 = \left\|E_B^{t+1} - E_B^t\right\|_2^2 > \frac{\delta}{8}\left(\frac{\varepsilon}{8}\right)^2,$$

while in the latter case,

$$\left\|E_C^{t+1}\right\|_2^2 - \left\|E_C^t\right\|_2^2 > \frac{\delta}{8}\left(\frac{\varepsilon}{8}\right)^2.$$

□

3.1 N Agents

What if there are three or more agents, each of whom talks only to its 'neighbors'? Will the agents still reach agreement, and if so, after how long? The answer is not obvious, even given the results of Section 3. For we could imagine that the sole intermediary between Alice and Bob is a weak-willed agent who agrees with Alice after talking to Alice, then agrees with Bob after talking to Bob, and so on, but never brings Alice and Bob into agreement with each other.

Formally, let G be a directed graph with vertices $1, \ldots, N$, each representing an agent. Suppose messages can only be sent from agent i to agent j if (i,j) is an edge in G. We need to assume G is strongly connected, since otherwise reaching agreement could be impossible for trivial reasons. In this setting, a *standard protocol* consists of a sequence of edges $(i_1, j_1), \ldots, (i_t, j_t), \ldots$ of G. At the t^{th} step, agent i_t sends its current expectation $E_{i_t}^{t-1}$ of f to agent j_t, whereupon j_t updates its expectation accordingly. Call the protocol *fair* if every edge occurs infinitely often in the sequence. Parikh and Krasucki [16] proved the following important theorem.

Theorem 5 (Parikh and Krasucki). *Any fair protocol will cause all the agents' expectations to agree after a finite number of messages T. Indeed, it will be common knowledge among the agents that $E_1^T = \cdots = E_N^T$.*

Our goal is to cause every pair of agents to (ε, δ)-agree,[3] after a number of steps polynomial in N, $1/\delta$, and $1/\varepsilon$. We can achieve this via the following "spanning-tree protocol." Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two spanning trees of G of minimum diameter, with $\mathcal{T}_1$ pointing outward from agent 1 to the other $N-1$ agents, and $\mathcal{T}_2$ pointing inward back to agent 1. Let $\mathcal{O}_1$ be an ordering of the edges in $\mathcal{T}_1$, in which every edge originating at i is preceded by an edge terminating at i, unless $i = 1$. Likewise let $\mathcal{O}_2$ be an ordering of the edges in $\mathcal{T}_2$, in which every edge originating at i is preceded by an edge terminating at i, unless i is a leaf of $\mathcal{T}_2$. Then the protocol is simply for agents to send their current expectations along edges of G, in the following order: first all edges in $\mathcal{O}_1$, then all edges in $\mathcal{O}_2$, then all edges in $\mathcal{O}_1$, and so on alternately.

Theorem 6. *The spanning-tree protocol causes every pair of agents to (ε, δ)-agree after $O\left(\frac{Nd^2}{\delta\varepsilon^2}\right)$ messages, where d is the diameter of G.*

Proof. Like all subsequent proofs in this paper, the proof of Theorem 6 is deferred to the full version. □

Let us make three remarks about Theorem 6. First, naturally one can combine Theorems 6 and 4, to obtain an N-agent protocol in which the messages are discrete. Second, all we really need about the *order* of messages is that information gets propagated from any agent in G to any other in a reasonable number of steps. The spanning-tree construction is designed to guarantee this, but sending messages in a random order (for example) would also work. Third, it seems fair to assume that many agents send messages in parallel; if so, the complexity bound can certainly be improved.

3.2 Limitations of the Standard Protocol

We have seen that two agents, using the standard protocol, will always (ε, δ)-agree after exchanging only $O\left(1/\left(\delta\varepsilon^2\right)\right)$ messages. This result immediately raises three questions. First, is there a scenario where the standard protocol *needs* $\Omega\left(1/\varepsilon^2\right)$ messages to produce (ε, δ)-agreement? Second, is the standard protocol always optimal, or do other protocols sometimes outperform it? And third, is there a scenario where *any* agreement protocol needs a number of communication bits polynomial in $1/\varepsilon$? Although we leave the third question open, we were able to resolve the first and second questions, as follows.

Theorem 7. *For all ε, δ, there exists a scenario such that using the continuous or discretized standard protocols, Alice and Bob must exchange $\Omega\left(\frac{1/\varepsilon^2}{\log(2/(1-\delta))}\right)$ messages before they (ε, δ)-agree. On the other hand, there exists a different protocol for this particular scenario that requires only 2 messages, both consisting of $\Theta(\log 2/\delta)$ bits. In particular, if $\delta = 1/2$, then the standard protocol uses $\Theta\left(1/\varepsilon^2\right)$ bits while the new protocol uses only $\Theta(1)$ bits.*

The key to Theorem 7 is to construct a scenario that forces the agents' expectations to follow a random walk. Thus, there is an initial disagreement by $\sim 2\varepsilon$, which can only be resolved by Alice sending a message to Bob. But then that message causes a new disagreement by $\sim 2\varepsilon$ that can only be resolved by Bob sending a message to Alice, and so on. Assuming the agents use the standard protocol, each of these messages moves its recipient's expectation either up by $\sim 4\varepsilon$ or down by $\sim 4\varepsilon$, with equal probability from the recipient's point of view (see Figure 2). The magnitude of disagreement only falls below ε after E_A^t and E_B^t get close to one of the "absorbing barriers" 0 or 1, and we can lower-bound the expected number of steps until that happens using standard results about random walks on the line.

In our new protocol, Alice and Bob exchange the same sequence of messages as in the standard protocol, but they gradually "attenuate" their messages by adding more and more random noise to them.[4] Surprisingly, such noise would

[3] If we want every pair of agents to agree within ε with *global* probability $1-\delta$, then we want every pair to $\left(\varepsilon, \delta/N^2\right)$-agree.

[4] Also, since the messages turn out to be nonadaptive, they can all be concatenated into one message from Alice and one message from Bob.

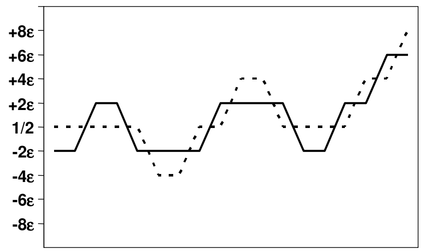

Figure 2: Alice's expectation E_A^t (solid line) and Bob's expectation E_B^t (dashed line) follow coupled random walks, in such a way that they continually differ by $\sim 2\varepsilon$.

actually help! For intuitively, the noise replaces a single large disagreement by many smaller disagreements that are likely to cancel each other out. Note that this strategy takes advantage of special properties of our random walk scenario, and we do not know how general it is.

The main defect of Theorem 7 is that the scenario had to be tailored to a particular choice of ε. We can give another theorem that fixes this defect, although the advantage of the attenuated protocol becomes smaller.

THEOREM 8. *For all n and all constants $\gamma \in (0,2)$, there exists a scenario with $|\Omega| = 2^{2n}$ such that for all ε greater than $1/n^{1/(2-\gamma)}$, Alice and Bob must exchange $\Omega\left(1/\varepsilon^{2-\gamma}\right)$ messages using the continuous or discretized standard protocols before they $(\varepsilon, 1/2)$-agree.*[5] *On the other hand, there exists a different protocol for this scenario that requires only 2 messages, both consisting of $\Theta(1/\varepsilon)$ bits.*

Further details about Theorems 7 and 8 are deferred to the full version.

4. COMPUTATIONAL COMPLEXITY

The previous sections have weakened the idea that communication cost is a fundamental barrier to agreement. However, we have glossed over the issue of *computational* cost entirely. A protocol that requires only $O\left(1/\left(\delta\varepsilon^2\right)\right)$ messages has little relevance if it would take Alice and Bob billions of years to calculate the messages! Moreover, all protocols we have discussed seem to have that problem, since the number of possible states $|\Omega|$ could be exponential in the length n of the agents' inputs.

Recognizing this issue, Hanson [10] introduced the notion of a "Bayesian wannabe": a computationally bounded agent that can still make sense of what its expectations would be if it had enough computational power to be a Bayesian. He then showed that under certain assumptions, if two Bayesian wannabes agree to disagree about the expectation of a function f, then they must also disagree about some variable that is independent of the state of the world $\omega \in \Omega$. However, this result does not suggest a *protocol* by which two Bayesian wannabes who agree about all state-independent variables could come to agree about f as well.

Admittedly, if the two wannabes have *very* limited abilities, it might be trivial to get them to agree. For example, if Alice and Bob both ignore all their evidence and estimate $f = 1/3$, then they agree before exchanging even a single message. But this example seems contrived: after all, if one of the agents (with equal justification) estimated $f = 2/3$, then no sequence of messages would ever cause them to agree within $\varepsilon < 1/3$. So informally, what we really want to know is whether two wannabes will always agree, having put in a "good-faith effort" to emulate Bayesian rationality.

We are thus led to the following question: is there an agreement protocol that

(i) would cause two computationally-unbounded Bayesians to (ε, δ)-agree after exchanging a small number of bits, and

(ii) can be simulated using a small amount of computation?

We will say shortly what we mean by a "small amount of computation." By "simulate," we mean that a computationally-unbounded referee, given the state $\omega \in \Omega$ together with a transcript $M = (m_1, \ldots, m_R)$ of all messages exchanged during the protocol, should be unable to decide (with non-negligible bias) whether Alice and Bob were Bayesians following the protocol exactly, or Bayesian wannabes merely simulating it. More formally, let $\mathcal{B}(\omega)$ be the probability distribution over message transcripts, assuming Alice and Bob are Bayesians and the state of the world is ω. Likewise, let $\mathcal{W}(\omega)$ be the distribution assuming Alice and Bob are wannabes. Then we require that for all Boolean functions $\Phi(\omega, M)$,

$$\left| \begin{array}{c} \Pr_{\omega \in \mathcal{D}, M \in \mathcal{B}(\omega)}\left[\Phi(\omega, M) = 1\right] - \\ \Pr_{\omega \in \mathcal{D}, M \in \mathcal{W}(\omega)}\left[\Phi(\omega, M) = 1\right] \end{array} \right| \leq \zeta \qquad (*)$$

where ζ is a parameter that can be made as small as we like.

A consequence of the requirement (*) is that even if Alice is computationally unbounded, she cannot decide with bias greater than ζ whether Bob is also unbounded, judging only from the messages he sends to her. For if Alice could decide, then so could our hypothetical referee, who learns at least as much about Bob as Alice does. Though a little harder to see, another consequence is that if Alice is unbounded, but knows Bob to be bounded and *takes his algorithm into account* when computing her expectations, her messages will still be statistically indistinguishable from what they would have been had she believed that Bob was unbounded. Indeed, no beliefs, beliefs about beliefs, etc., about whether either agent is bounded or not can significantly affect the sequence of messages, since the truth or falsehood of those beliefs is almost irrelevant to predicting the agents' future messages.

Because of these considerations, we claim that, while simulating a Bayesian agreement protocol might not be the *only* way for two Bayesian wannabes to reach an "honest" agreement, it is certainly a *sufficient* way. Therefore, if we can show how to meet even the stringent requirement (*), this will provide strong evidence that computation time is not a fundamental barrier to agreement.

But what do we mean by computation time? We assume the state space Ω is a subset of $\{0,1\}^n \times \{0,1\}^n$, so that Alice's initial knowledge is an n-bit string x, and Bob's is an n-bit string y. Given the prior distribution $\mathcal{D}$ over (x, y)

[5] Any constant $\delta \in (0,1)$ would work equally well here. For simplicity, we omit the asymptotic dependence on $1/\delta$ and $1/(1-\delta)$.

pairs, let $\mathcal{D}_A^x$ be Alice's posterior distribution over y conditioned on x, and let $\mathcal{D}_B^y$ be Bob's posterior distribution over x conditioned on y. The following two computational assumptions are the only ones that we make:

(1) Alice and Bob can both evaluate $f(\omega)$ for any $\omega \in \Omega$.

(2) Alice and Bob can both sample from $\mathcal{D}_A^x$ for any $x \in \{0,1\}^n$, and from $\mathcal{D}_B^y$ for any $y \in \{0,1\}^n$.

Our simulation procedure will *not* have access to descriptions of f or $\mathcal{D}$; it can learn about them only by calling subroutines for (1) and (2) respectively. The complexity of the procedure will then be expressed in terms of the number of subroutine calls, other computations adding only a negligible amount of time. Thus, we might stipulate that both subroutines should run in time polynomial in n. On the other hand, n could be extremely large—otherwise the agents would simply exchange their entire inputs and be done. So we might want to be even stricter, and stipulate that the subroutines should use time (say) *logarithmic* in n, albeit with many parallel processors. In any case, the simulation procedure will treat the subroutines purely as black boxes, so decisions about their implementation will not affect our results.

The justification for assumptions (1) and (2) is simply that without them, it is hard to see how the agents could estimate their expectations even before they started talking to each other. In other words, we have to assume the agents enter the conversation with minimal tools for reasoning about their universe of discourse. We do *not* assume that those tools extend to reasoning about each other's expectations, expectations of expectations, etc., conditioned on a sequence of messages exchanged. That the tools do extend in this way is what we intend to prove.

It might seem unreasonable that Alice can sample from Bob's distribution $\mathcal{D}_B^y$, and Bob can sample from Alice's distribution $\mathcal{D}_A^y$. On reflection, however, this is just the computational analogue of the standard assumption that the partitions themselves are known to both agents, and can justified using the same arguments (see Section 2).

Finally, let us note that assumptions (1) and (2) can both be relaxed. In particular, it is enough to approximate $f(\omega)$ to within an additive factor η with probability at least $1-\eta$, in time that increases polynomially in $1/\eta$. It is also enough to sample from a distribution whose variation distance from $\mathcal{D}_A^x$ or $\mathcal{D}_B^y$ is at most η, in time polynomial in $1/\eta$. Indeed, since the probabilities and f-values are real numbers, we will generally *need* to approximate in order to represent them with finite precision. For ease of presentation, though, we assume exact algorithms in what follows.

4.1 Smoothed Standard Protocol

Naïvely, requirement (*) seems impossible to satisfy. All of the agreement protocols discussed earlier in this paper—for example, that of Theorem 4—are easy to distinguish from any efficient simulation of them. For consider Alice's first message to Bob. If Alice's expectation E_A^0 is below some threshold c, she sends one message, whereas if $E_A^0 \geq c$, she sends a different message. Even if we fix f, and limit probabilities and f-values to (say) n bits of precision, we can arrange things so that $E_A^0(\omega)$ is exponentially close to c, sometimes greater and sometimes less, with high probability over ω. Then to decide which message to send, Alice needs to evaluate f exponentially many times.

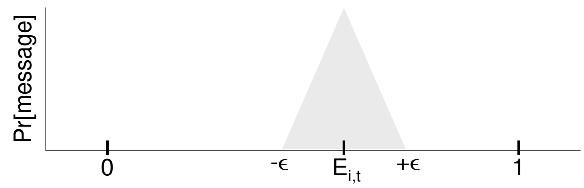

Figure 3: Agent i "smoothes" its expectation E_i^t with triangular noise before sending it.

We resolve this issue by having the agents add random noise to their messages ("smoothing" them), even if they are unbounded Bayesians. This noise does not prevent the agents from reaching (ε, δ)-agreement. On the other hand, it makes their messages easier to simulate. For unlike real numbers $a \neq b$, which are perfectly distinguishable no matter how close they are, two probability distributions with close means may be hard to distinguish, like wavepackets in quantum mechanics.

In the *smoothed standard protocol*, Alice generates her messages to Bob as follows. Let $b \geq \log_2(200/\varepsilon)$ be a positive integer to be specified later. Then let ϵ be an integer multiple of 2^{-b} between $\varepsilon/50$ and $\varepsilon/40$, and let $L = 2^b\epsilon$. First Alice rounds her current expectation E_A^t of f to the nearest multiple of 2^{-b}. Denote the result by round$\left(E_A^t\right)$. She then draws an integer $r \in \{-L, \ldots, L\}$, according to a *triangular distribution* in which $r = j$ with probability $(L - |j|)/L^2$ (see Figure 3). The message she sends Bob is $m_{t+1} = \text{round}\left(E_A^t\right) + 2^{-b}r$. Observe that since $m_{t+1} \in [-\epsilon, 1+\epsilon]$, there are at most $2^b(1+2\epsilon)+1$ possible values of m_{t+1}—meaning Alice's message takes only $b+1$ bits to specify. After receiving the message, Bob updates his expectation of f using Bayes' rule, then draws an integer $r \in \{-L, \ldots, L\}$ according to the same triangular distribution and sends Alice $m_{t+2} = \text{round}\left(E_B^{t+1}\right) + 2^{-b}r$. The two agents continue to send messages in this way.

The reader might be wondering why we chose triangular noise, and whether other types of noise would work equally well. The answer is that we want the message distribution to have three basic properties. First, it should be concentrated about a mean of E_i^t with variance at most $\sim \epsilon^2$. Second, shifting the mean by $\eta \leq \epsilon$ should shift the distribution by at most $\sim \eta/\epsilon$ in variation distance. And third, the derivative of the probably density function should never exceed $\sim \eta/\epsilon^2$ in absolute value. Thus, Gaussian noise would also work, though it is somewhat harder to analyze than triangular noise. However, noise that is uniform over $[-\epsilon, \epsilon]$ would *not* work (so far as we could tell), since it violates the third property.

Let $M_t = (m_1, \ldots, m_t)$ consist of the first t messages that Alice and Bob exchange. Since messages are now probabilistic, the agents' expectations of f at step t depend not only on the initial state of the world ω, but also on M_t. When we want to emphasize this, we denote the agents' expectations by $E_A^t(\omega, M_t)$ and $E_B^t(\omega, M_t)$ respectively. Another important consequence of messages being probabilistic is that after an agent has received a message, its posterior distribution over Ω is no longer obtainable by restricting the prior distribution $\mathcal{D}$ to a subset of possible states. Thus, we let $\Pi_i(\omega) = \Pi_i^0(\omega)$, since we will never refer to $\Pi_i^t(\omega)$ for $t > 0$.

Say the agents (ε, δ)-agree after the t^{th} message if

$$\Pr_{\omega \in \mathcal{D}, M_t}\left[\left|E_A^t(\omega, M_t) - E_B^t(\omega, M_t)\right| > \varepsilon\right] \leq \delta.$$

THEOREM 9. *The smoothed standard protocol causes Alice and Bob to (ε, δ)-agree after at most $2/\left(\delta\varepsilon^2\right)$ messages.*

4.2 Simulating the Smoothed Protocol

Having asserted that the smoothed standard protocol works, in this section we explain how Alice and Bob can simulate the protocol. In the ideal case—where the agents have unlimited computational power—they use the following recursive formulas. Let

$$\Delta\left(m_t, E_i^{t-1}\right) = \begin{cases} 1 - \left|m_t - \text{round}\left(E_i^{t-1}\right)\right| / \epsilon \\ \quad \text{if } \left|m_t - \text{round}\left(E_i^{t-1}\right)\right| \leq \epsilon, \\ \text{or } 0 \text{ otherwise} \end{cases}$$

be proportional to the probability that agent i sends message m_t, given that its expectation is E_i^{t-1}. Also, let $q_t(\omega, M_t)$ be proportional to the joint probability of messages $m_1, \ldots, m_t$ assuming the true state of the world is ω. Then assuming t is even and suppressing dependencies on M_t, for all X, Y we have

$$q_t(Y) = q_{t-2}(Y)\,\Delta\left(m_t, E_B^{t-1}(Y)\right),$$
$$q_{t-1}(X) = q_{t-3}(X)\,\Delta\left(m_{t-1}, E_A^{t-2}(X)\right),$$
$$E_A^t(X) = \frac{\text{EX}_{Y \in \Pi_A(X)}\left[q_t(Y) f(Y)\right]}{\text{EX}_{Y \in \Pi_A(X)}\left[q_t(Y)\right]},$$
$$E_B^{t-1}(Y) = \frac{\text{EX}_{X \in \Pi_B(Y)}\left[q_{t-1}(X) f(X)\right]}{\text{EX}_{X \in \Pi_B(Y)}\left[q_{t-1}(X)\right]}$$

with the base cases $q_0(Y) = q_{-1}(X) = 1$ for all X, Y. The correctness of these formulas follows from simple Bayesian manipulations. Having computed $E_i^t(\omega)$ by the formulas above (note that this does not require knowledge of ω), all agent i needs to do is draw $r \in \{-L, \ldots, L\}$ from the triangular distribution, then send the message

$$m_{t+1} = \text{round}\left(E_i^t(\omega)\right) + 2^{-b} r.$$

In the real case, the agents are computationally bounded, and can no longer afford the luxury of taking expectations over the exponentially large sets Π_i. A natural idea is to compensate by somehow *sampling* those sets. But since we never assumed the ability to sample Π_i conditioned on messages $m_1, \ldots, m_t$, it is not obvious how that make that idea work. Our solution will consist of two phases: the construction of "sampling-trees," which involves no communication, followed by a message-by-message simulation of the ideal protocol. Let us describe these phases in turn.

(I) Sampling-Tree Construction. Alice creates a tree $\mathcal{T}_A$ with height R and branching factor K. Here $R < 2/\left(\delta\varepsilon^2\right)$ is the number of messages, and K is a parameter to be specified later. Let root_A be the root node of $\mathcal{T}_A$, and let $S(v)$ be the set of children of node v. Then Alice labels each of the K nodes $w \in S(\text{root}_A)$ by a sample $Y_w \in \Omega_A(\omega)$, drawn independently from her posterior distribution $\mathcal{D}_A^x$ (recall that by assumption, she can sample efficiently from $\Pi_A(\omega)$ for any $\omega \in \Omega$). Next, for each $w \in S(\text{root}_A)$, she labels each of the K nodes $v \in S(w)$ by a sample $X_v \in \Pi_B(Y_w)$, drawn independently from Bob's distribution $\mathcal{D}_B^y$ where $Y_w = (x, y)$. She continues recursively in this manner, labeling each v an even distance from the root with a sample $X_v \in \Pi_B(Y_w)$ where w is the parent of v, and each w an odd distance from the root with a sample $Y_w \in \Pi_A(X_v)$ where v is the parent of w. Thus her total number of samples is $K + K^2 + \cdots + K^R$. Similarly, Bob creates a tree $\mathcal{T}_B$ with height R and branching factor K. Let root_B be the root of $\mathcal{T}_B$; then Bob labels each $v \in S(\text{root}_B)$ by a sample $X_v \in \Pi_B(\omega)$, each child $w \in S(v)$ of each $v \in S(\text{root}_B)$ by a sample $Y_w \in \Pi_A(X_v)$, and so on, alternating between Π_B and Π_A at successive levels. As a side remark, if the agents share a random string, then there is no reason for them not to use the same set of samples. However, we cannot assume that such a string is available.

(II) Simulation. We now explain how the agents can use the samples from (I) to simulate the smoothed standard protocol. Let i be the agent that sends the t^{th} message (Alice if t is odd, or Bob if t is even). Then i's main task at step t is to compute an estimator $\left\langle E_i^{t-1}(\text{root}_i)\right\rangle_i$ for its current expectation $E_i^{t-1}(\omega) = E_i^t(\omega)$. To do so, it recursively computes estimators for all nodes in its sample tree and all earlier time steps: $\langle E_A^u(v)\rangle_i \approx E_A^u(X_v)$ and $\langle q_{u-1}(v)\rangle_i \approx q_{u-1}(X_v)$ for all "Alice" nodes $v \in \mathcal{T}_i$ and even $u \leq t$, and $\langle E_B^u(w)\rangle_i \approx E_B^u(Y_w)$ and $\langle q_{u-1}(w)\rangle_i \approx q_{u-1}(Y_w)$ for all "Bob" nodes $w \in \mathcal{T}_i$ and odd $u \leq t$. Assuming t is even, the requisite formulas are as follows:

$$\langle q_t(w)\rangle_i = \langle q_{t-2}(w)\rangle_i\,\Delta\left(m_t, \left\langle E_B^{t-1}(w)\right\rangle_i\right),$$
$$\langle q_{t-1}(v)\rangle_i = \langle q_{t-3}(v)\rangle_i\,\Delta\left(m_{t-1}, \left\langle E_A^{t-2}(v)\right\rangle_i\right),$$
$$\left\langle E_A^t(v)\right\rangle_i = \frac{\sum_{w \in S(v)} \langle q_t(w)\rangle_i f(Y_w)}{\sum_{w \in S(v)} \langle q_t(w)\rangle_i},$$
$$\left\langle E_B^{t-1}(w)\right\rangle_i = \frac{\sum_{v \in S(w)} \langle q_{t-1}(v)\rangle_i f(X_v)}{\sum_{v \in S(w)} \langle q_{t-1}(v)\rangle_i}.$$

Here the base cases are $\langle q_{-1}(v)\rangle_i = 1$ for all Alice nodes $v \in \mathcal{T}_i$, and $\langle q_0(w)\rangle_i = 1$ for all Bob nodes $w \in \mathcal{T}_i$. So for example, Alice's initial estimate $\left\langle E_A^0(\text{root}_A)\right\rangle_A \approx E_A^0(\omega)$ is simply the average of $f(Y_w)$ over all $w \in S(\text{root}_A)$:

$$\left\langle E_A^0(\text{root}_A)\right\rangle_A = \frac{\sum_{w \in S(\text{root}_A)} 1 \cdot f(Y_w)}{\sum_{w \in S(\text{root}_A)} 1} = \underset{w \in S(\text{root}_A)}{\text{EX}}\left[f(Y_w)\right].$$

Given its estimate $\left\langle E_i^{t-1}(\text{root}_i)\right\rangle_i$ at the root of $\mathcal{T}_i$, agent i generates its message in the obvious way: it first chooses an $r \in \{-L, \ldots, L\}$ uniformly at random, and then sends the message

$$m_t = \text{round}\left(\left\langle E_i^{t-1}(\text{root}_i)\right\rangle_i\right) + 2^{-b} r.$$

That completes the description of the simulation procedure. Its complexity is easily determined: summing over all R communication rounds, both agents need $O\left(K^R\right)$ calls to the subroutine that samples from $\mathcal{D}_A^x$ or $\mathcal{D}_B^y$, and $O\left(K^R\right)$ calls to the subroutine that evaluates f.

4.3 Analysis

The key result proved in the full version of the paper is that the simulation procedure works. In other words, for some "reasonable" precision b and sample size K, the distribution over message sequences in the simulated protocol is statistically indistinguishable from the distribution in the ideal protocol:

THEOREM 10. *By setting* $b = \lceil \log_2 (5R/\zeta\epsilon) \rceil$ *and* $K = O\left((11/\epsilon)^{R^2} / \zeta^2 \right)$, *we can achieve*

$$\left| \begin{array}{c} \Pr_{\omega \in \mathcal{D}, M \in \mathcal{B}(\omega)} \left[\Phi(\omega, M) = 1 \right] - \\ \Pr_{\omega \in \mathcal{D}, M \in \mathcal{W}(\omega)} \left[\Phi(\omega, M) = 1 \right] \end{array} \right| \leq \zeta$$

for all Boolean functions Φ.

More concretely, if we substitute $R \leq 2/\left(\delta\varepsilon^2\right)$ and $\epsilon \geq \varepsilon/50$, then the total number of bits communicated is

$$Rb = O\left(\frac{1}{\delta\varepsilon^2} \log \frac{1}{\zeta\delta\varepsilon^3} \right),$$

while the total number of subroutine calls is of order

$$\left(\frac{(11/\epsilon)^{R^2}}{\zeta^2} \right)^R \leq \exp\left(\frac{8 \ln (550/\varepsilon)}{\delta^3 \varepsilon^6} + \frac{4 \ln (1/\zeta)}{\delta\varepsilon^2} \right).$$

While no one would pretend that the second bound above is practical, note that it has no dependence on n, and that it grows "only" exponentially in $\mathrm{poly}(1/\varepsilon)$ and $\mathrm{poly}(1/\delta)$.

The proof of Theorem 10 is extremely involved; here we can only sketch the main ideas. One's first thought is that the proof should be a straightforward (if tedious) application of Chernoff bounds. The problem is that *a priori*, it is possible that a single large error anywhere in the sample tree $\mathcal{T}_i$ could propagate all the way up to the root, destroying the simulation. Of course, if the probability of such an error were small enough, then we would simply use the union bound to argue that almost certainly, no such error happens anywhere in the tree. Unfortunately, the probability is *not* small enough, for in the formulas for $\left\langle E_A^t(v) \right\rangle_i$ and $\left\langle E_B^{t-1}(w) \right\rangle_i$, whenever the denominators are zero or close to zero, the resulting estimates of $E_A^t(X_v)$ and $E_B^{t-1}(Y_w)$ are worthless. Intuitively, this means that if a message has low probability from its recipient's point of view, then the recipient needs many samples to find even a single input that would have caused the sender to produce that message. But as we increase the number of samples K to deal with this problem, the number of nodes $\sim K^R$ where something could go wrong increases at an even faster rate, so we need to increase K again, and so on *ad infinitum*! Our proof cuts off this infinite regress by showing that with high probability, the errors introduced by "bad nodes" are washed out by "good nodes" before they can propagate to the root. Ultimately, the problem reduces to one of evaluating some nasty integrals.

5. DISCUSSION

> "We publish this observation with some diffidence, since once one has the appropriate framework, it is mathematically trivial. Intuitively, though, it is not quite obvious..." —Aumann [1], on his original agreement theorem

This paper has studied agreement protocols from the quantitative perspective of theoretical computer science. If nothing else, we hope to have shown that adopting that perspective leads to rich mathematical questions. Here are a few of the more interesting open problems raised by our results.

(1) How tight is our $O\left(1/\left(\delta\varepsilon^2\right)\right)$ upper bound on agreement complexity? Recall that the best lower bound we currently know is $\Omega(\log 1/\varepsilon)$, from Proposition 1.

(2) Do our conclusions break down if the "true" distribution over Ω differs from the common prior $\mathcal{D}$? In particular, is there a scenario where Alice and Bob must exchange $\Omega(n)$ bits to $(\varepsilon, 0)$-agree for some $\varepsilon > 0$? (It is easy to construct a scenario where the discretized standard protocol needs $\Omega(n)$ bits to produce $(\varepsilon, 0)$-agreement.)

(3) Can the simulation procedure of Section 4.2 be made more efficient? In particular, can we reduce the number of subroutine calls to (say) $c^{1/\left(\delta\varepsilon^2\right)}$, or even to a polynomial in $1/\delta$ and $1/\varepsilon$? Alternatively, can we prove a lower bound showing that such reductions are impossible?

6. ACKNOWLEDGMENTS

I thank Robin Hanson for correspondence during the early stages of this work, and for his wonderfully disorienting papers which introduced me to the topic. I am also grateful to Robert Aumann, Eddie Dekel, Ran Raz, Alexander Razborov, Umesh Vazirani, Ronald de Wolf, and the anonymous reviewers for helpful discussions and comments.

7. REFERENCES

[1] R. J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[2] R. J. Aumann. Reply to Gul. *Econometrica*, 66(4):929–938, 1998.

[3] R. J. Aumann. Interactive epistemology I: knowledge. *International J. Game Theory*, 28:263–300, 1999.

[4] A. Brandenburger and E. Dekel. Common knowledge with probability 1. *Journal of Mathematical Economics*, 16:237–245, 1987.

[5] J. A. K. Cave. Learning to agree. *Economics Letters*, 12:147–152, 1983.

[6] T. Cowen and R. Hanson. Are disagreements honest? At http://hanson.gmu.edu/deceive.pdf, 2004.

[7] J. D. Geanakoplos and H. M. Polemarchakis. We can't disagree forever. *J. Economic Theory*, 28:192–200, 1982.

[8] F. Gul. A comment on Aumann's Bayesian view. *Econometrica*, 66(4):923–927, 1998.

[9] R. Hanson. Disagreement is unpredictable. *Economics Letters*, 77(3):365–369, 2002.

[10] R. Hanson. For savvy Bayesian wannabes, are disagreements not about information? *Theory and Decision*, 54(2):105–123, 2003.

[11] R. Hanson. Uncommon priors require origin disputes. To appear, 2005.

[12] Y. Moses J. Y. Halpern, R. Fagin and M. Y. Vardi. *Reasoning About Knowledge*. Cambridge, 1996.

[13] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge, 1996.

[14] D. Monderer and D. Samet. Approximating common knowledge with common beliefs. *Games and Economic Behavior*, 1:170–190, 1989.

[15] C. H. Papadimitriou and M. Yannakakis. On complexity as bounded rationality. In *Proc. ACM STOC*, pages 726–733, 1994.

[16] R. Parikh and P. Krasucki. Communication, consensus, and knowledge. *J. Economic Theory*, 52:178–189, 1990.