Agreement Implies Accuracy for Substitutable Signals

RAFAEL FRONGILLO, University of Colorado Boulder, USA
ERIC NEYMAN, Columbia University, USA
BO WAGGONER, University of Colorado Boulder, USA

Inspired by Aumann's agreement theorem, Aaronson [2005] studied the amount of communication necessary for two Bayesian experts to approximately agree on the expectation of a random variable. Aaronson showed that, remarkably, the number of bits does not depend on the amount of information available to each expert. However, in general the agreed-upon estimate may be inaccurate: far from the estimate they would settle on if they were to share all of their information. We show that if the experts' signals are *substitutes*—meaning the experts' information has diminishing marginal returns—then it is the case that if the experts are close to agreement then they are close to the truth. We prove this result for a broad class of agreement and accuracy measures that includes squared distance and KL divergence. Additionally, we show that although these measures capture fundamentally different kinds of agreement, Aaronson's agreement result generalizes to them as well.

CCS Concepts: • **Theory of computation → Communication complexity**; **Market equilibria**.

Additional Key Words and Phrases: Agreement, communication protocols, information structures, informational substitutes

ACM Reference Format:
Rafael Frongillo, Eric Neyman, and Bo Waggoner. 2023. Agreement Implies Accuracy for Substitutable Signals. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC '23), July 9–12, 2023, London, United Kingdom.* ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/3580507.3597679

1 Introduction

Suppose that Alice and Bob are honest, rational Bayesians who wish to estimate some quantity—say, the unemployment rate one year from now. Alice is an expert on historical macroeconomic trends, while Bob is an expert on contemporary monetary policy. They convene to discuss and share their knowledge with each other until they reach an agreement about the expected value of the future unemployment rate. Alice and Bob could reach agreement by sharing everything they had ever learned, at which point they would have the same information, but the process would take years. How then should they proceed?

In the seminal work "Agreeing to Disagree," Aumann [Aumann, 1976] observed that Alice and Bob can reach agreement simply by taking turns sharing their current expected value for the quantity. In addition to modeling communication between Bayesian agents, protocols similar to this one model financial markets: each trader shares partial information about their expected value on their turn (discussed in Section 5). A remarkable result by Scott Aaronson [Aaronson, 2005] shows that if Alice and Bob follow certain protocols of this form, they will agree to within ϵ with

EC '23, July 9–12, 2023, London, United Kingdom

ACM ISBN 979-8-4007-0104-7/23/07.
https://doi.org/10.1145/3580507.3597679

probability $1-\delta$ by communicating $O\left(\frac{1}{\delta\epsilon^2}\right)$ bits.[1] Notably, this bound only depends on the error Alice and Bob are willing to tolerate, and not on the amount of information available to them.

Absent from Aaronson's results, however, is *what Alice and Bob agree on*. In particular, there is no guarantee that Alice and Bob will be *accurate*, meaning their agreed-upon estimate will be close (in e.g. expected squared distance) to what they would believe if they shared all of their information. In fact, they might agree on something highly inaccurate: suppose that Alice and Bob have independent, uniformly random bits b_A, b_B, and wish to estimate the XOR $b_A \oplus b_B$. Alice and Bob agree from the onset, as from each of their perspectives, the expected value of $b_A \oplus b_B$ is $\frac{1}{2}$. Yet this expectation is far from the best estimate given their collective knowledge, which is either 0 or 1. So while agreement is fundamental to understanding communication between Bayesians—in Aumann's terms, they cannot "agree to disagree"—agreement is far from the whole story. An important open problem is therefore what assumptions guarantee that Alice and Bob are accurate once they agree.

We address this open problem by introducing a natural condition, called *rectangle substitutes*, under which agreement implies accuracy. Rectangle substitutes is a notion of *informational substitutes*: the property that additional information has diminishing marginal returns. The notion of substitutes is ubiquitous in optimization problems, and informational substitutes conditions have recently been used to analyze equilibria in markets [Chen and Waggoner, 2016]. In that context, Kong and Schoenebeck [2023] showed for conditionally independent signals convergence of the popular LMSR market implies full information aggregation, i.e. accuracy. We show that under the rectangle substitutes condition, *any* protocol leading to agreement will also lead to accuracy. We then extend these results beyond the case of squared error, to a broad family of measures of agreement and accuracy including KL divergence.

1.1 Overview of approach and results

In Aaronson [2005], Alice and Bob are said to *agree* if the squared distance between their estimates is small. Likewise, we can say that Alice and Bob are *accurate* if the squared distance between each of their estimates and the truth is small. In Section 3 we present our first main result: under these definitions, **if the information structure satisfies rectangle substitutes, then agreement implies accuracy**. In other words, under this assumption, when two Bayesians agree—regardless of how little information they have shared—they necessarily agree *on the truth*.

The proof involves carefully partitioning the space of posterior beliefs induced by the protocol. Agreement is used to show that Alice and Bob usually fall into the same partition element, which means that Bob would not learn much from learning the partition element of Alice's expectation. Then, the rectangle substitutes condition is used to show that if Bob were to learn Alice's partition element, then he would be very close to knowing the truth.

Aaronson measures agreement in terms of squared error, yet other measurements like KL divergence may be better suited for some settings. For example, if Alice and Bob estimate the probability of a catastrophic event as 10^{-10} and 10^{-2}, respectively, then under squared error they are said to agree closely, but arguably they disagree strongly, as reflected by their large KL divergence. Motivated these different ways to measure agreement, we next ask:

(1) Can Aaronson's protocols be generalized to other notions of agreement, such that the number of bits communicated is independent of the amount of information available to Alice and Bob?
(2) Do other notions of agreement necessarily imply accuracy under rectangle substitutes?

[1] To ensure that each message is short, Alice and Bob share discretized versions of their estimates; we discuss this in Section 2.

In Section 4, we give our second and third main results: **the answer to both questions is yes.** Specifically, the positive results apply when when measuring agreement and accuracy using Bregman divergences, a class of error measures that includes both squared distance and KL divergence.[2]

Aaronson's proof of his agreement theorem turns out to be specific to squared distance. Our agreement theorem (Theorem 4.11) modifies Aaronson's protocol to depend on the particular Bregman divergence, i.e. the relevant error measure. It then proceeds in a manner inspired by Aaronson but using several new ideas. Our proof that agreement implies accuracy under rectangle substitutes for general Bregman divergences also involves some nontrivial changes to our proof for squared distance. In particular, the fact that the length of an interval cannot be inferred from the Bregman divergence between its endpoints necessitates a closer analysis of the partition of Alice's and Bob's beliefs.

We conclude in Section 5 with a discussion of connections between agreement protocols and information revelation in financial markets, and discuss an interesting potential avenue for future work.

1.2 Related Work

[Geanakoplos and Polemarchakis, 1982] discussed the distinction between agreement and full information revelation. One result shown is that under a natural probability measure on information structures, full agreement and information revelation occur in a single round of communication with probability one. However, conditions for accuracy and the concept of substitutes are not discussed.

Our setting is related to but distinct from communication complexity. In that field (e.g. [Rao and Yehudayoff, 2020]), the goal is for Alice and Bob to correctly compute a function of their inputs while communicating as few bits as possible and using any protocol necessary. By contrast, Aaronson [2005] considered a goal of agreement, not correctness, and focused on specific natural protocols, which he showed achieve this goal in a constant number of bits. Our work focuses on Aaronson's setting. We discuss how our results might be framed in terms of communication complexity in Appendix E.

Our introduction of the substitutes condition is inspired by its usefulness in prediction markets [Chen and Waggoner, 2016]. The "expectation-sharing" agreement protocols we study bear a strong similarity to dynamics of market prices. Ostrovsky [2012] introduced a condition under which convergence of prices in a market implies that all information is aggregated. This can be viewed as an "agreement implies accuracy" condition. Similarly, Kong and Schoenebeck [2023] presented a result that, for the logarithmic market scoring rule (LMSR) and conditionally independent signals, convergence of the market implies full information revelation. Our results are conceptually similar, although they are technically quite different as we rely on the novel condition of *rectangle substitutes*. In the context of the LMSR, the rectangle substitutes notion includes conditionally independent signals as a special case (see discussion in Section 4.1). We discuss the connection of our work to markets in Section 5. Another similar definition of informational substitutes is used by [Neyman and Roughgarden, 2021a] in the context of robust aggregation of forecasts.

Finally, we note that the "agreement protocols" we study are not related to key agreement protocols in cryptography, where the goal is for two communicating parties to jointly construct a shared string for cryptographic use.

[2]The third result holds under an "approximate triangle inequality" condition on the Bregman divergence, which is satisfied by most or all natural choices; indeed, it is nontrivial to construct a Bregman divergence that does not satisfy this property.

2 Preliminaries

2.1 Information Structures

We consider a set Ω of states of the world, with a probability distribution $\mathbb{P}$ over the world states. There are two experts, Alice and Bob. Alice learns the value of a random variable $\sigma : \Omega \to \mathcal{S}$; we call σ Alice's *signal* and $\mathcal{S}$ her *signal set*. Correspondingly, Bob learns the value of a random variable $\tau : \Omega \to \mathcal{T}$. These signals each convey partial information about the true state $\omega \in \Omega$. Alice and Bob are interested in a third random variable $Y : \Omega \to [0, 1]$. We use the term *information structure* to refer to the tuple $\mathcal{I} := (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$.

We denote by $\mu_{\sigma\tau} := \mathbb{E}[Y \mid \sigma, \tau]$ the random variable that is equal to the expected value of Y conditioned on both Alice's signal σ and Bob's signal τ.[3] We also define $\mu_\sigma := \mathbb{E}[Y \mid \sigma]$ and $\mu_\tau := \mathbb{E}[Y \mid \tau]$. For a measurable set $S \subseteq \mathcal{S}$, we define $\mu_S := \mathbb{E}[Y \mid \sigma \in S]$; we define μ_T analogously for $T \subseteq \mathcal{T}$. Additionally, for $T \subseteq \mathcal{T}$, we define $\mu_{\sigma T} := \mathbb{E}[Y \mid \tau \in T, \sigma]$, i.e. the expected value of Y conditioned on the particular value of σ and the knowledge that $\tau \in T$. If Alice knows that Bob's signal belongs to T (and nothing else about his signal), then the expected value of Y conditional on her information is $\mu_{\sigma T}$; we refer to this as Alice's *expectation*. Likewise, for $S \subseteq \mathcal{S}$, we define $\mu_{S\tau} := \mathbb{E}[Y \mid \sigma \in S, \tau]$. Finally, we define $\mu_{ST} := \mathbb{E}[Y \mid \sigma \in S, \tau \in T]$. This is the expectation of a third party who only knows that $\sigma \in S$ and $\tau \in T$.

In general we often wish to take expectations conditioned on $\sigma \in S, \tau \in T$ (for some $S \subseteq \mathcal{S}, T \subseteq \mathcal{T}$). We will use the shorthand $\mathbb{E}[\cdot \mid S, T]$ for $\mathbb{E}[\cdot \mid \sigma \in S, \tau \in T]$ in such cases.

2.2 Agreement Protocols

The notion of *agreement* between Alice and Bob is central to our work. We first define agreement in terms of squared error, and generalize to other error measures in Section 4.

Definition 2.1 (ϵ-agree). *Let a and b be Alice's and Bob's expectations, respectively (a and b are random variables on Ω). Alice and Bob* ϵ-agree *if* $\frac{1}{4}\mathbb{E}\left[(a-b)^2\right] \le \epsilon$.

The constant $\frac{1}{4}$ makes the left-hand side represent Alice's and Bob's distance to the average of their expectations.

Our setting follows [Aaronson, 2005], which examined communication protocols that cause Alice and Bob to agree. In a *(deterministic) communication protocol*, Alice and Bob take turns sending each other messages (strings of bits). On Alice's turns, Alice communicates a message that is a deterministic function of her input (i.e. her signal σ) and all previous communication, and likewise for Bob on his turns. A *rectangle* is a set of the form $S \times T$ where $S \subseteq \mathcal{S}$ and $T \subseteq \mathcal{T}$.

The *communication transcript* is the ordered tuple of all messages that have been sent. The transcript *at time step t* refers to the tuple consisting of the first t messages. The transcript at time step t partitions Ω into rectangles: for any given sequence of t messages, there are subsets $S_t \subseteq \mathcal{S}, T_t \subseteq \mathcal{T}$ such that the protocol transcript at time t is equal to this sequence if and only if $(\sigma, \tau) \in S_t \times T_t$.[4]

For a given communication protocol, we may think of S_t and T_t as random variables. Alice's expectation at time t (i.e. *after* the t-th message has been sent) is $\mu_{\sigma T_t}$ and Bob's expectation at time t is $\mu_{S_t \tau}$. Finally, the protocol terminates at a certain time (which need not be known in advance of the protocol). While typically in communication complexity a protocol is associated with a final

[3] The value of Y need not be determined by σ and τ, although for our purposes the case in which it is determined is essentially equivalent.

[4] We can see this inductively: suppose the transcript at time step $t-1$ partitions Ω into rectangles, and (without loss of generality) that the t-th turn is Alice's. Consider one of these rectangles. Alice's message can only depend on her input and the transcript so far, which means that her message can only partition this rectangle into sub-rectangles.

output, in this case we are interested in Alice's and Bob's expectations, so we do not require an output.

It will be convenient to hypothesize a third party observer, whom we call Charlie, who observes the protocol but has no other information. At time t, Charlie has expectation $\mu_{S_t T_t}$. Charlie's expectation can also be interpreted as the expectation of Y according to Alice and Bob's common knowledge. Note that Alice and Bob each know Charlie's expectation at any given time.

The following definition formalizes the relationship between communication protocols and agreement.

Definition 2.2 (ϵ-agreement protocol). *Given an information structure $\mathcal{I}$, a communication protocol* causes Alice and Bob to ϵ-agree *on $\mathcal{I}$ if Alice and Bob ϵ-agree at the end of the protocol, i.e., if $\frac{1}{4}\mathbb{E}\left[(\mu_{\sigma T_t} - \mu_{S_t \tau})^2\right] \leq \epsilon$, where the expected value is over Alice's and Bob's inputs. We say that a communication protocol is an* ϵ-agreement protocol *if the protocol causes Alice and Bob to ϵ-agree on every information structure.*

Aaronson defines and analyzes two ϵ-agreement protocols.[5] The first of these is the *standard protocol*, in which Alice and Bob take turns stating their expectations for a number of time steps that can be computed by Alice and Bob independently in advance of the protocol, and which is guaranteed to be at most $O(1/\epsilon)$.

The fact that exchanging their expectations for $O(1/\epsilon)$ time steps results in ϵ-agreement is profound and compelling. However, the standard protocol may require an unbounded number of bits of communication, since Alice and Bob are exchanging real numbers. To address this, Aaronson defines another agreement protocol that is truly polynomial-communication (which we slightly modify for our purposes):

Definition 2.3 (Discretized protocol, [Aaronson, 2005]). *Choose $\epsilon > 0$. In the* discretized protocol *with parameter ϵ, on her turn (at time t), Alice sends "low" if her expectation is smaller than Charlie's by more than $\epsilon/4$, i.e. if $\mu_{S_{t-1}\tau} < \mu_{S_{t-1}T_{t-1}} - \epsilon/4$; "high" if her expectation is larger than Charlie's by more than $\epsilon/4$; and "medium" otherwise. Bob acts analogously on his turn. At the start of the protocol, Alice and Bob use the information structure to independently compute the time $t_{end} \leq \frac{1000}{\epsilon}$ that minimizes $\mathbb{E}\left[(\mu_{\sigma T_{t_{end}}} - \mu_{S_{t_{end}}\tau})^2\right]$. The protocol ends at this time.*

Theorem 2.4 ([Aaronson, 2005, Theorem 4]). *The discretized protocol with parameter ϵ is an ϵ-agreement protocol with transcript length $O(1/\epsilon)$ bits.*

In general, we refer to Aaronson's standard and discretized protocols as examples of *expectation-sharing* protocols. We will define other examples in Section 4, similar to Aaronson's discretized protocol but with different cutoffs for low, medium, and high. We also interpret expectation-sharing protocols in the context of markets in Section 5.

2.3 Accuracy and Informational Substitutes

Most of our main results give conditions such that if Alice and Bob ϵ-agree, then Alice's and Bob's estimates are accurate. By *accurate*, we mean that Alice's and Bob's expectations are close to $\mu_{\sigma\tau}$, i.e., what they would believe if they knew each other's signals. (After all, they cannot hope to have a better estimate of Y than $\mu_{\sigma\tau}$; for this reason we sometimes refer to $\mu_{\sigma\tau}$ as the "truth.") Formally:

[5] A minor difference to our framing is that Aaronson [2005] focuses on *probable approximate agreement*: protocols that cause the absolute difference between Alice and Bob to be at most ϵ with probability all but δ. While the results as presented in this section are stronger than those in [Aaronson, 2005] (the original results follow from these as a consequence of Markov's inequality), these results follow from a straightforward modification of his proofs.

Definition 2.5 (ϵ-accurate). *Let a be Alice's expectation. Alice is* ϵ-accurate *if* $\mathbb{E}\left[(\mu_{\sigma\tau} - a)^2\right] \leq \epsilon$. *We define ϵ-accuracy analogously for Bob.*

One cannot hope for an unconditional result stating that if Alice and Bob agree, then they are accurate. Consider for instance the *XOR information structure* from the introduction: Alice and Bob each receive independent random bits as input, and Y is the XOR of these bits. Then from the start Alice and Bob agree that the expected value of Y is exactly $\frac{1}{2}$, but this value is far from $\mu_{\sigma\tau}$, which is either 0 or 1.

Intuitively, this situation arises because Alice's and Bob's signals are *informational complements*: each signal is not informative by itself, but they are informative when taken together. On the other hand, we say that signals are *informational substitutes* if learning one signal is less valuable if you already know the other signal. An extreme example is if $\sigma = \tau = X$ for any random variable X. Here σ becomes useless upon learning τ and vice versa. Chen and Waggoner [2016],[6] discuss formalizations of several notions of informational substitutes. All of these notions capture "diminishing marginal value," in the sense that, roughly speaking, the value of partial information is a submodular set function. The various definitions proposed by Chen and Waggoner [2016] only differ in how finely they allow decomposing σ and τ to obtain a marginal unit. Our definition has the same format, but uses a decomposition inspired by information rectangles in communication complexity. Recall that we write $\mid S, T$ as shorthand for $\mid \sigma \in S, \tau \in T$.

Definition 2.6. *An information structure $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ satisfies* rectangle substitutes *if for every $S \subseteq \mathcal{S}, T \subseteq \mathcal{T}$ such that $\mathbb{P}[\sigma \in S, \tau \in T] > 0$, we have*

$$\mathbb{E}\left[(Y - \mu_{S\tau})^2 \mid S, T\right] - \mathbb{E}\left[(Y - \mu_{\sigma\tau})^2 \mid S, T\right] \leq \mathbb{E}\left[(Y - \mu_{ST})^2 \mid S, T\right] - \mathbb{E}\left[(Y - \mu_{\sigma T})^2 \mid S, T\right]. \quad (1)$$

This definition is a strengthening of Chen and Waggoner's notion of weak substitutes for two agents: an information structure satisfies *weak substitutes* if Equation 1 holds for $S = \mathcal{S}$ and $T = \mathcal{T}$ [Chen and Waggoner, 2016].

We will show that under rectangle substitutes, if Alice and Bob approximately agree, then they are approximately accurate.

Interpreting substitutes. Both sides of Equation 1 represent the "value" of learning σ as measured by a decrease in error. The left-hand side gives the decrease if one already knows τ and that $\sigma \in S$; the right-hand side gives the decrease if one only knows that $\sigma \in S, \tau \in T$. Substitutes thus says: the marginal value of learning σ is smaller if one already knows τ than if one does not. This statement should hold for every sub-rectangle S, T. We remark that the inequality can be rearranged to focus instead on the marginal value of τ rather than σ. We also note that in the XOR information structure, the left-hand side of the inequality is $\frac{1}{4}$ while the right-hand side is zero: a large violation of the substitutes condition. In the example $\sigma = \tau = X$, the left side is always zero.

Chen and Waggoner [2016] discusses three interpretations of substitutes, which motivate it as a natural condition. (1) Each side of the inequality measures an improvement in *prediction error*, here the squared loss, due to learning σ. Under substitutes, the improvement is smaller if one already knows τ. (2) Each side measures a *decrease in uncertainty* (here, measured roughly by *variance*) due to learning σ. Under substitutes, σ provides less information about Y if one already knows τ.[7] (3) Each side measures the *decrease in distance* of a posterior expectation from the truth when learning σ. The distance to Y changes less if one already knows τ.

[6] We recommend the ArXiv version for the most up-to-date introduction to informational substitutes.

[7] Here, uncertainty is measured by variance of one's belief. Under the KL divergence analogue covered in Section 4.1, uncertainty is measured in bits via Shannon entropy.

Restrictiveness of substitutes. It is natural to ask about the strength of the rectangle substitutes assumption. In the case that $|\mathcal{S}| = |\mathcal{T}| = 2$, the condition reduces to the well-established "weak substitutes" condition of [Chen and Waggoner, 2016]. For larger signal sets, the set of information structures satisfying rectangle substitutes remains nontrivial. For example, it is satisfied by a positive fraction of information structures (for a natural choice of measure). We show this fact in Appendix A by exhibiting an information structure in which Equation 1 holds *strictly* for all S, T with $|S|, |T| \geq 2$ — meaning that all nearby information structures also satisfy rectangle substitutes. Finally, we note that although the rectangle substitutes condition is strong due to the quantification over sub-rectangles, in Section 3.2 we prove that our main results decay gracefully for information structures that are close to but do not quite satisfy the rectangle substitutes condition.

2.4 The Pythagorean Theorem

We will use the following fact throughout. We defer the proof to Appendix C, where we establish a more general version of this statement.

Proposition 2.7 (Pythagorean theorem). *Let A be a random variable, $B = \mathbb{E}[A \mid \mathcal{F}]$ where $\mathcal{F}$ is a sigma-algebra, and C be a random variable defined on $\mathcal{F}$. Then*

$$\mathbb{E}\left[(A-C)^2\right] = \mathbb{E}\left[(A-B)^2\right] + \mathbb{E}\left[(B-C)^2\right].$$

We use the phrase *Pythagorean theorem* in part because of its form, and in part because it is precisely the familiar Pythagorean theorem when the random variables are viewed as points in a Hilbert space[8] with inner product $\langle X, Y\rangle := \mathbb{E}[XY]$.

Informally, A is a random variable, B is the expected value of A conditional on some partial information, and C is a random variable that only depends on this information. So the theorem applies when B is a coarse estimate of A and C is at least as coarse as B, a scenario that often occurs in our setting.

One application of the Pythagorean theorem in our context takes $A = Y$, $B = \mu_{\sigma\tau}$ (the expected value of Y conditioned on the experts' signals), and $C = \mu_{\sigma T}$ (Alice's expected value, which only depends on her signal and thus on the signal pair). This particular application, along with the symmetric one taking $C = \mu_{S\tau}$, allows us to rewrite the rectangle substitutes condition in a form that we will find more convenient:

Remark 2.8. An information structure $\mathcal{I}$ satisfies rectangle substitutes if and only if

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2 \mid S, T\right] \leq \mathbb{E}\left[(\mu_{\sigma T} - \mu_{ST})^2 \mid S, T\right] \tag{2}$$

for all S, T such that $\mathbb{P}[\sigma \in S, \tau \in T] > 0$.

3 Results for Squared Distance

Our main results show that, under the rectangle substitutes condition, any communication protocol that causes Alice and Bob to agree also causes them to be accurate. We now show the first of these results, which is specific to the squared distance error measure that we have been discussing.

3.1 Agreement Implies Accuracy

Theorem 3.1. *Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies rectangle substitutes. For any communication protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$, Alice and Bob are $10\epsilon^{1/3}$-accurate after the protocol terminates.*

The crux of the argument is the following lemma.

[8] We do not make use of this abstraction in our work, but we refer the interested reader to [Šidák, 1957].

Lemma 3.2. *Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies rectangle substitutes. Let $\epsilon = \mathbb{E}\left[(\mu_\sigma - \mu_\tau)^2\right]$. Then*

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] \le 6\epsilon^{1/3}.$$

Let us first prove Theorem 3.1 assuming Lemma 3.2 is true.

Proof of Theorem 3.1. Consider any protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$. Let S be the set of possible signals of Alice at the end of the protocol which are consistent with the protocol transcript, and define T likewise for Bob. Intuitively, $S \times T$ is the set of plausible signal pairs (σ, τ) according to an external observer of the protocol. Observe that S and T are random variables, each a function of both σ and τ. We have

$$\begin{aligned}
\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2\right] &= \mathbb{E}_{S,T}\left[\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2 \mid S, T\right]\right] \\
&\le \mathbb{E}_{S,T}\left[6\left(\mathbb{E}\left[(\mu_{\sigma T} - \mu_{S\tau})^2 \mid S, T\right]\right)^{1/3}\right] \\
&\le 6\mathbb{E}_{S,T}\left[\mathbb{E}\left[(\mu_{\sigma T} - \mu_{S\tau})^2 \mid S, T\right]\right]^{1/3} \\
&= 6\mathbb{E}\left[(\mu_{\sigma T} - \mu_{S\tau})^2\right]^{1/3} \le 6(4\epsilon)^{1/3} \le 10\epsilon^{1/3}.
\end{aligned}$$

In the second step, we apply Lemma 3.2 to the information structure $\mathcal{I}$ restricted to $S \times T$ — that is, to $\mathcal{I}' = (\Omega', \mathbb{P}', S, T, Y)$, where $\Omega' = \{\omega \in \Omega : \sigma \in S, \tau \in T\}$ and $\mathbb{P}'[\omega] = \mathbb{P}[\omega \mid \sigma \in S, \tau \in T]$. (Note that we use the fact that if $\mathcal{I}$ satisfies rectangle substitutes, then so does $\mathcal{I}'$; this is because a rectangle of $\mathcal{I}'$ is also a rectangle of $\mathcal{I}$.) The third step follows by the concavity of $x^{1/3}$. Therefore, Bob is $10\epsilon^{1/3}$ accurate (and Alice is likewise by symmetry). □

The proof of Lemma 3.2 relies on the following claim. We defer the proof of Lemma 3.2 (and Claim 3.3) to Appendix B, and instead sketch the proofs here.

Claim 3.3. *In the setting of Lemma 3.2, for any $N \ge 1$, it is possible to partition $[0, 1]$ into N intervals $[0, x_1), [x_1, x_2), \ldots, [x_{N-1}, 1]$ in a way so that each interval has length at most $\frac{2}{N}$, and*

$$\mathbb{P}\left[k(\sigma) \ne k(\tau)\right] \le \sqrt{\epsilon}N,$$

where $k(\sigma)$ denotes the $k \in [N]$ such that $x_{k-1} \le \mu_\sigma < x_k$, and $k(\tau)$ is defined analogously.[9]

Intuitively, Claim 3.3 is true because if $\mathbb{E}\left[(\mu_\sigma - \mu_\tau)^2\right]$ is small, then μ_σ and μ_τ are likely to fall into the same interval.

We now sketch the proof of Lemma 3.2. To see why Claim 3.3 is relevant, recall that we wish to upper bound the expectation of $(\mu_{\sigma\tau} - \mu_\tau)^2$. Let $S^{(k)} := \{\sigma \in \mathcal{S} : x_{k-1} \le \mu_\sigma < x_k\}$. By the Pythagorean theorem, we have

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] = \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S^{(k(\sigma))}\tau})^2\right] + \mathbb{E}\left[(\mu_{S^{(k(\sigma))}\tau} - \mu_\tau)^2\right].$$

By using the rectangle substitutes condition for $S = S^{(k)}, T = \mathcal{T}$ for every k, we find that

$$\mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k(\sigma))}})^2\right] \ge \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S^{(k(\sigma))}\tau})^2\right]. \tag{3}$$

Therefore, we have

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] \le \mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k(\sigma))}})^2\right] + \mathbb{E}\left[(\mu_{S^{(k(\sigma))}\tau} - \mu_\tau)^2\right]. \tag{4}$$

Claim 3.3 lets us argue that the first of these two terms is small (because μ_σ and $\mu_{S^{(k(\sigma))}}$ are always within $\frac{2}{N}$ of each other) and that the second term is also small (because conditioned on τ, $k(\sigma)$ is known with high probability). We find that choosing $N = \epsilon^{-1/6}$ gives us the bound in Lemma 3.2.

[9]For convenience we define $x_0 = 0$ and x_N to be some number greater than 1.

Theorem 3.1 is a general result about agreement protocols. Applying the result to Aaronson's discretized protocol gives us the following result.

Corollary 3.4. *Let $\mathcal{I}$ be any information structure that satisfies universal rectangle substitutes. For any $\epsilon > 0$, Alice and Bob will be ϵ-accurate after running Aaronson's discretized protocol with parameter $\epsilon^3/1000$ (and this takes $O(1/\epsilon^3)$ bits of communication).*

Remark 3.5. The discretized protocol is not always the most efficient agreement protocol. For example, Proposition B.1 shows that if the rectangle substitutes condition holds, agreement (and therefore accuracy) can be reached with just $O(\log(1/\epsilon))$ bits, an improvement on Corollary 3.4. We discuss communication complexity further in Appendix E. Even if more efficient protocols are sometimes possible, expectation-sharing protocols are of interest because they model naturally-occurring communication processes. For example, they capture the dynamics of prices in markets, which we also discuss in Section 5. More generally, we find it remarkable that Alice and Bob become accurate by running the agreement protocol (indeed *any* agreement protocol), despite such protocols being designed with only agreement in mind.

Finally, we observe the following important consequence of Theorem 3.1: once Alice and Bob agree, they continue to agree.

Corollary 3.6. *Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies rectangle substitutes. Consider a communication protocol with the property that Alice and Bob ϵ-agree after round t. Then Alice and Bob $10\epsilon^{1/3}$-agree on all subsequent time steps.*

Proof. If Alice and Bob ϵ-agree then they are $10\epsilon^{1/3}$-accurate, so in particular $\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{\sigma T_t})^2\right] \leq 10\epsilon^{1/3}$. Note that $\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{\sigma T_s})^2\right]$ is a decreasing function of s, since for any $s_1 \leq s_2$ we have

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{\sigma T_{s_1}})^2\right] = \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{\sigma T_{s_2}})^2\right] + \mathbb{E}\left[(\mu_{\sigma T_{s_2}} - \mu_{\sigma T_{s_1}})^2\right]$$

by the Pythagorean theorem. Therefore, for any $t' > t$, we have $\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{\sigma T_{t'}})^2\right] \leq 10\epsilon^{1/3}$. Symmetrically, we have $\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S_{t'}\tau})^2\right] \leq 10\epsilon^{1/3}$. Therefore, $\mathbb{E}\left[(\mu_{\sigma T_{t'}} - \mu_{S_{t'}\tau})^2\right] \leq 40\epsilon^{1/3}$, which means that after round t', Alice and Bob $10\epsilon^{1/3}$-agree. □

Corollary 3.6 stands in contrast to the more general case, in which it is possible that Alice and Bob "nearly agree for the first $t-1$ time steps, then disagree violently at the t-th step" [Aaronson, 2005, §2.2]. Thus, while the main purpose of Theorem 3.1 is a property about *accuracy*, an *agreement* property falls out naturally: under the rectangle substitutes condition, once Alice and Bob are close to agreement, they will remain in relatively close agreement into the future.

3.2 Graceful Decay Under Closeness to Rectangle Substitutes

In a sense, the rectangle substitutes condition is quite strong: it requires that the weak substitutes condition be satisfied on *every* sub-rectangle. One might hope for a result that generalizes Theorem 3.1 to information structures that almost satisfy the rectangle substitutes condition but do not quite. Let us formally define a notion of closeness to rectangle substitutes.

Definition 3.7. *An information structure $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ satisfies* δ-approximate rectangle substitutes *if for every partition of $\mathcal{S} \times \mathcal{T}$ into rectangles,[10] the rectangle substitutes condition holds in expectation over the partition, up to an additive constant of δ, i.e., if we have*

$$\mathbb{E}_{\sigma,\tau}\left[(\mu_{\sigma\tau} - \mu_{S_{\sigma,\tau}\tau})^2\right] \leq \mathbb{E}_{\sigma,\tau}\left[(\mu_{\sigma T_{\sigma,\tau}} - \mu_{S_{\sigma,\tau}T_{\sigma,\tau}})^2\right] + \delta, \tag{5}$$

[10]There are partitions into rectangles that cannot arise from a communication protocol. Our results would apply equally if this condition were instead defined for every partition that could arise from a communication protocol, but we state this condition more generally so that it could be applicable in a broader context than the analysis of communication protocols.

where $S_{\sigma,\tau} \times T_{\sigma,\tau}$ is the rectangle containing (σ, τ).

Remark 3.8. The δ-approximate rectangle substitutes property is a relaxation of the rectangle substitutes property, in the sense that the two are equivalent if $\delta = 0$. To see this, first observe that if $\mathcal{I}$ satisfies rectangle substitutes, then it satisfies Equation 5 with $\delta = 0$ pointwise across all $S_{\sigma,\tau}, T_{\sigma,\tau}$, and thus in expectation. In the other direction, suppose that $\mathcal{I}$ satisfies 0-approximate rectangle substitutes. Let $S \subseteq \mathcal{S}, T \subseteq \mathcal{T}$ and consider the partition of $\mathcal{I}$ into rectangles that contains $S \times T$ and, separately, every other signal pair (σ, τ) in its own rectangle. For this partition, Equation 5 reduces precisely to Equation 2 (the rectangle substitutes condition for S and T).

Theorem 3.1 generalizes to approximate rectangle substitutes as follows.

Theorem 3.9. *Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies δ-approximate rectangle substitutes. For any communication protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$, Alice and Bob are $(10\epsilon^{1/3} + \delta)$-accurate after the protocol terminates.*

Proof. We first observe that Lemma 3.2 can be modified as follows.

Lemma 3.10. *Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies δ-approximate rectangle substitutes. Let $\epsilon = \mathbb{E}\left[(\mu_\sigma - \mu_\tau)^2\right]$. Then*

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] \le 6\epsilon^{1/3} + \delta.$$

The proof of Lemma 3.10 is exactly the same as that of Lemma 3.2, except that Equation 3 (Equation 7 in the full proof) includes an additive δ term on the left-hand side:

$$\mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k(\sigma))}})^2\right] + \delta \ge \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S^{(k(\sigma))}\tau})^2\right].$$

This modified inequality follows immediately from the δ-approximate rectangle substitutes condition, noting that one partition of $\mathcal{S} \times \mathcal{T}$ into rectangles is $\{S_1 \times \mathcal{T}, \ldots, S_N \times \mathcal{T}\}$. The extra δ term produces the δ term in the lemma statement.

To prove the theorem, let S be the set of possible signals of Alice at the end of the protocol which are consistent with the protocol transcript, and define T likewise for Bob. Let δ_{ST} be the minimum δ such that $S \times T$ satisfies δ-approximate rectangle substitutes. Note that $\mathbb{E}_{S,T}[\delta_{ST}] \le \delta$: otherwise, by taking the union over the worst-case partitions for each S, T we would exhibit a partition of $\mathcal{S} \times \mathcal{T}$ into rectangles that would violate the δ-approximate rectangle substitutes property. Therefore we have

$$\begin{aligned}
\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2\right] &= \mathbb{E}_{S,T}\left[\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2 \mid S, T\right]\right] \\
&\le \mathbb{E}_{S,T}\left[6\left(\mathbb{E}\left[(\mu_{\sigma T} - \mu_{S\tau})^2 \mid S, T\right]\right)^{1/3} + \delta_{ST}\right] \\
&\le 6\mathbb{E}_{S,T}\left[\mathbb{E}\left[(\mu_{\sigma T} - \mu_{S\tau})^2 \mid S, T\right]\right]^{1/3} + \delta \\
&= 6\mathbb{E}\left[(\mu_{\sigma T} - \mu_{S\tau})^2\right]^{1/3} + \delta = 6(4\epsilon)^{1/3} + \delta \le 10\epsilon^{1/3} + \delta.
\end{aligned}$$

As in the proof of Theorem 3.1, the second step follows by applying Lemma 3.2 to the information structure $\mathcal{I}$ restricted to $S \times T$. □

4 Results for Other Divergence Measures

Squared distance is a compelling error measure because it elicits the mean. That is, if you wish to estimate a random variable Y and will be penalized according to the squared distance between Y and your estimate, the strategy that minimizes your expected penalty is to report the expected value of Y (conditional on the information you have). This is in contrast to e.g. absolute distance as

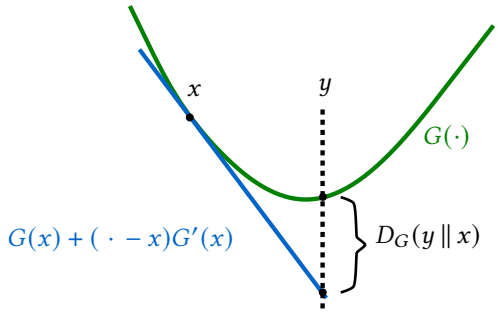

Fig. 1. The Bregman divergence $D_G(y \parallel x)$ is the vertical distance at y between G and the tangent line to G at x.

an error measure, which would instead elicit the median of your distribution. The class of error measures that elicit the mean is precisely the class of *Bregman divergences* (defined below).

In this section, our main result is a **generalization of Theorem 3.1 to (almost) arbitrary Bregman divergences** (see e.g. Theorem 4.14). Additionally, we provide a **generalization of Aaronson's discretized protocol to arbitrary Bregman divergences** (Theorem 4.11).

4.1 Preliminaries on Bregman Divergences

Definition 4.1. *Given a differentiable,*[11] *strictly convex function* $G : [0,1] \to \mathbb{R}$*, and* $x, y \in [0,1]$*, the* Bregman divergence *from* y *to* x *is*

$$D_G(y \parallel x) := G(y) - G(x) - (y - x)G'(x).$$

Proposition 4.2 ([Banerjee et al., 2005]). *Given a random variable* Y*, the quantity* $\mathbb{E}\left[D_G(Y \parallel x)\right]$ *is minimized by* $x = \mathbb{E}[Y]$.

An intuitive formulation of Bregman divergence is that $D_G(y \parallel x)$ can be found by drawing the line tangent to G at x and computing how far below the point $(y, G(y))$ this line passes. We illustrate this in Figure 1. Note that the Bregman divergence is not in general symmetric in its arguments; indeed, $G(x) = x^2$ is the only G for which it is.

The Bregman divergence with respect to $G(x) = x^2$ is precisely the squared distance. Another common Bregman divergence is the KL divergence, which corresponds to $G(x) = x \log x + (1 - x)\log(1 - x)$, the negative of Shannon entropy.

We generalize relevant notions such as agreement and accuracy to arbitrary Bregman divergences as follows. In the definitions below, $G : [0,1] \to \mathbb{R}$ is a differentiable, strictly convex function.

Definition 4.3. *Let* a *be Alice's expectation. Alice is* ϵ-accurate *if* $\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel a)\right] \leq \epsilon$*, and likewise for Bob.*

We discuss our choice of the order of these two arguments (i.e. why we do not instead consider the expectation of $D_G(a \parallel \mu_{\sigma\tau})$) in Appendix D. We now define ϵ-agreement, and to do so we first define the *Jensen-Bregman divergence.*

Definition 4.4. *For* $a, b \in [0,1]$*, the* Jensen-Bregman divergence *between* a *and* b *with respect to* G *is*

$$\mathcal{JB}_G(a,b) := \frac{1}{2}\left(D_G\left(a \parallel \frac{a+b}{2}\right) + D_G\left(b \parallel \frac{a+b}{2}\right)\right) = \frac{G(a)+G(b)}{2} - G\left(\frac{a+b}{2}\right).$$

[11] When we say "differentiable," we mean differentiable on the interior of the interval on which G is defined.

The validity of the second equality can be easily derived from the definition of Bregman divergence. Note that the Jensen-Bregman divergence, unlike the Bregman divergence, is symmetric in its arguments. The Jensen-Bregman divergence is a lower bound on the average Bregman divergence from Alice and Bob to any other point (see Proposition C.1 (i)).

Definition 4.5. *Let a and b be Alice's and Bob's expectations, respectively. Alice and Bob* ϵ-agree *with respect to G if $\mathcal{JB}_G(a, b) \leq \epsilon$.*

In Appendix D we discuss alternative definitions of agreement and accuracy. The upshot of this discussion is that our definition of agreement is the *weakest* reasonable one, and our definition of accuracy is the *strongest* reasonable one. This means that the main result of this section—that under a wide class of Bregman divergence, agreement implies accuracy—is quite powerful: it starts with a weak premise and proves a strong conclusion.

Definition 4.6. *Given an information structure $\mathcal{I}$, a communication protocol* causes Alice and Bob to ϵ-agree *on $\mathcal{I}$ with respect to G if Alice and Bob ϵ-agree with respect to G at the end of the protocol. A communication protocol is an* ϵ-agreement protocol *with respect to G if the protocol causes Alice and Bob to ϵ-agree with respect to G on every information structure.*

We also generalize the notion of rectangle substitutes to this domain, following [Chen and Waggoner, 2016], which explored notions of substitutes for arbitrary Bregman divergences.

Definition 4.7. *Let $G : [0, 1] \to \mathbb{R}$ be a differentiable, strictly convex function. An information structure $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ satisfies rectangle substitutes with respect to G if for every $S \subseteq \mathcal{S}, T \subseteq \mathcal{T}$, we have*

$$\begin{aligned}&\mathbb{E}\left[D_G(Y \parallel \mu_{S\tau}) \mid S, T\right] - \mathbb{E}\left[D_G(Y \parallel \mu_{\sigma\tau}) \mid S, T\right] \\ &\leq \mathbb{E}\left[D_G(Y \parallel \mu_{ST}) \mid S, T\right] - \mathbb{E}\left[D_G(Y \parallel \mu_{\sigma T}) \mid S, T\right].\end{aligned}$$

Chen and Waggoner [2016] explore the notion of weak substitutes with respect to arbitrary G's as well; just as before, $\mathcal{I}$ is said to satisfy the weak substitutes condition if the above inequality holds for $S = \mathcal{S}$ and $T = \mathcal{T}$. The authors additionally explore in detail the weak substitutes condition with respect to negative entropy, i.e. for D_G equal to the KL divergence. They show that if Alice and Bob have independent signals conditioned on Y, then the information structure satisfies weak substitutes with respect to this G. In fact, any such information structure also satisfies rectangle substitutes, because an information structure with conditionally independent signals retains the conditional independence when restricted to any sub-rectangle. The rectangle substitutes condition thus covers the specific case of conditionally independent signals under which Kong and Schoenebeck [2023] prove their accuracy result. On the other hand, the greater generality of our setting necessitates a different proof strategy.

The Pythagorean theorem (Proposition 2.7) generalizes to arbitrary Bregman divergences:

Proposition 4.8. *Let A be a random variable, $B = \mathbb{E}\left[A \mid \mathcal{F}\right]$ where $\mathcal{F}$ is a sigma-algebra, and C be a random variable defined on $\mathcal{F}$. Then*

$$\mathbb{E}\left[D_G(A \parallel C)\right] = \mathbb{E}\left[D_G(A \parallel B)\right] + \mathbb{E}\left[D_G(B \parallel C)\right].$$

Although the proof of this observation is fairly straightforward, to our knowledge Proposition 4.8 is original to this work. We provide a proof in Appendix C. Just as we did with squared error, this general Pythagorean theorem allows us to rewrite the rectangle substitutes condition for Bregman divergences.

Remark 4.9. An information structure $\mathcal{I}$ satisfies rectangle substitutes with respect to G if and only if for all $S \subseteq \mathcal{S}, T \subseteq \mathcal{T}$ we have

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_{S\tau}) \mid S, T\right] \leq \mathbb{E}\left[D_G(\mu_{\sigma T} \parallel \mu_{ST}) \mid S, T\right]. \tag{6}$$

Given the interpretation of Bregman divergences as measures of error, we can interpret the left side as Bob's expected error in predicting the truth while the right side is Charlie's expected error when predicting Alice's expectation (with Charlie as defined in Section 2.2). Both sides measure a prediction error due to not having Alice's signal, but from different starting points.

4.2 Generalizing the Discretized Protocol

Later in this work, we will show that under some weak conditions, protocols that cause Alice and Bob to agree with respect to G also cause Alice and Bob to be accurate with respect to G. However, this raises an interesting question: are there protocols that cause Alice and Bob to agree with respect to G? In particular, we are interested in natural expectation-sharing protocols. Aaronson's discretized protocol is specific to $G(x) = x^2$, and it is not immediately obvious how to generalize it. We present the following generalization.

Definition 4.10. *Let G be a differentiable, strictly convex function, and let $M := \max_x G(x) - \min_x G(x)$. Choose $\epsilon > 0$. In the* discretized protocol with respect to G *with parameter ϵ, on her turn (at time t), Alice sends "medium" if $D_G(\mu_{\sigma T_{t-1}} \parallel \mu_{S_{t-1}T_{t-1}}) < \frac{\epsilon}{2}$, and otherwise either "low" or "high", depending on whether $\mu_{\sigma T_t}$ is smaller or larger (respectively) than $\mu_{S_t T_t}$. Bob acts analogously on his turn. At the start of the protocol, Alice and Bob use the information structure to independently compute the time $t_{end} \leq \frac{24M(4M+\epsilon)}{\epsilon^2}$ that minimizes $\mathbb{E}\left[D_G(\mu_{\sigma T_{t_{end}}} \parallel \mu_{S_{t_{end}}\tau})\right]$. The protocol ends at this time.*

Theorem 4.11. *The discretized protocol with respect to G with parameter ϵ is an ϵ-agreement protocol that involves $O\left(\frac{M(M+\epsilon)}{\epsilon^2}\right)$ bits of communication.*

Our proof draws inspiration from Aaronson's proof of the discretized protocol, but has significant differences. The key idea is to keep track of the monovariant $\mathbb{E}\left[D_G(Y \parallel \mu_{S_t T_t})\right]$. This is Charlie's expected error (as measured by the Bregman divergence from the correct answer Y) after time step t—recall that Charlie is our name for a third-party observer of the protocol. Note that this quantity is at most M and at least 0. Hence, if we show that the quantity decreases by at least some value β every time Alice and Bob do not ϵ-agree, then we will have shown that Alice and Bob must ϵ-agree within $\frac{\beta}{M}$ time steps. We defer the proof to Appendix C.

4.3 Approximate Triangle Inequality

Our results will hold for a class of Jensen-Bregman divergences that satisfy an approximate version of the triangle inequality. Specifically, we will require JB_G to satisfy the following *c-approximate triangle inequality* for some $c > 0$.

Definition 4.12. *Given a differentiable, strictly convex function $G : [0,1] \to \mathbb{R}$ and a positive number c, we say that $\mathcal{JB}_G(\cdot,\cdot)$ satisfies the* c-approximate triangle inequality *if for all $a, b, x \in [0,1]$ we have*

$$\mathcal{JB}_G(a,x) + \mathcal{JB}_G(x,b) \geq c\,\mathcal{JB}_G(a,b).$$

It is possible to construct functions G such that there is no positive c for which JB_G satisfies the c-approximate triangle inequality. However, JB_G satisfies the c-approximate triangle inequality for some positive c for essentially all natural choices of G.

Proposition 4.13. *Let $G : [0,1] \to \mathbb{R}$ be a differentiable, strictly convex function.*

(i) If $\sqrt{\mathcal{JB}_G(\cdot,\cdot)}$ satisfies the triangle inequality, then $\mathcal{JB}_G$ satisfies the $\frac{1}{2}$-approximate triangle inequality.

(ii) If $G(x) = x^2$ (i.e. D_G is squared distance) or if $G(x) = x\log x + (1-x)\log(1-x)$ (i.e. D_G is KL divergence), then $\sqrt{\mathcal{JB}_G}$ satisfies the triangle inequality (and so $\mathcal{JB}_G$ satisfies the $\frac{1}{2}$-approximate triangle inequality).

Proof. Regarding Fact (i), suppose that $\sqrt{\mathrm{JB}_G}$ satisfies the triangle inequality. Then for all a, b, x we have $\sqrt{\mathrm{JB}_G(a,x)} + \sqrt{\mathrm{JB}_G(x,b)} \geq \sqrt{\mathrm{JB}_G(a,b)}$. Squaring both sides and observing that $\mathrm{JB}_G(a,x) + \mathrm{JB}_G(x,b) \geq 2\sqrt{\mathrm{JB}_G(a,x)\mathrm{JB}_G(x,b)}$ completes the proof.

Fact (ii) is trivial for $G(x) = x^2$, since $\sqrt{\mathrm{JB}_G}$ is the absolute distance metric (times a constant factor). As for $G(x) = x\log x + (1-x)\log(1-x)$, we refer the reader to [Endres and Schindelin, 2003]. □

The question of when $\sqrt{\mathrm{JB}_G}$ satisfies the triangle inequality has been explored in previous work; we refer the interested reader to [Acharyya et al., 2013] and [Chen et al., 2008].

4.4 Generalized Agreement Implies Generalized Accuracy

In all of the results in this subsection, we consider the following setting: G is a differentiable convex function; c is a positive real number such that JB_G satisfies the c-approximate triangle inequality; and $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ is an information structure that satisfies rectangle substitutes with respect to G.

We prove generalizations of Theorem 3.1, showing that under the rectangle substitutes condition, if a protocol ends with Alice and Bob in approximate agreement, then Alice and Bob are approximately accurate. The first result we state assumes that G is symmetric, but is otherwise quite general.

Theorem 4.14. *Assume that G is symmetric about the line $x = \frac{1}{2}$. For any communication protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$, and for any $\beta \geq \frac{2}{c}\epsilon$, Alice and Bob are*

$$\left(\frac{8}{c^2}\beta + 16\left(G(0) - G\left(\left(\frac{\epsilon}{\beta}\right)^{1/(1-\log_2 c)}\right)\right)\right)\text{-accurate}$$

after the protocol terminates.

This result is not our most general, as it assumes that G is symmetric, but this assumption likely holds for most use cases. To apply the result optimally, one must first optimize β as a function of G. For example, setting $\beta = \epsilon^{r/(r+1-\log_2 c)}$ (with r defined below) gives us the following corollary:[12]

Corollary 4.15. *Assume that $G(0) - G(x), G(1) - G(1-x) \leq O(x^r)$. For any communication protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$, Alice and Bob are $O\left(\epsilon^{r/(r+1-\log_2 c)}\right)$-accurate after the protocol terminates, where the constant hidden by $O(\cdot)$ depends on G.*

Remark 4.16. Concretely, if G' is bounded then we can choose $r = 1$, in which case our bound simplifies to $O\left(\epsilon^{1/(2-\log_2 c)}\right)$. If instead we assume that $c = \frac{1}{2}$ (as is the case if $\sqrt{\mathrm{JB}_G(\cdot,\cdot)}$ is a metric), then the bound is $O\left(\epsilon^{r/(r+2)}\right)$. If both of these are true, as is the case for $G(x) = x^2$, then the bound is $O(\epsilon^{1/3})$, which recovers our result in Theorem 3.1.

For G equal to the negative of Shannon entropy (i.e. the G for which D_G is KL divergence), setting $\beta = \epsilon^{1/3}(\log 1/\epsilon)^{2/3}$ in Theorem 4.14 gives us the following corollary.

[12] Corollary 4.15 as stated (without the symmetry assumption) is actually a corollary of Theorem 4.18.

Corollary 4.17. *If $G(x) = x\log x + (1-x)\log(1-x)$, then for any communication protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$, Alice and Bob are $O(\epsilon^{1/3}(\log 1/\epsilon)^{2/3})$-accurate after the protocol terminates.*

Theorem 4.14 follows from our most general result about agreement implying accuracy:

Theorem 4.18. *Let $\tilde{G}(x) := \max_{a,b:|a-b|\le x}(G(a) - G(b))$ be the maximum possible difference in G-values of two points that differ by at most x, and let $\tilde{G}^*(x)$ be the concave envelope of $\tilde{G}$, i.e.*

$$\tilde{G}^*(x) := \max_{0\le a,b,w\le 1: wa+(1-w)b=x} w\tilde{G}(a) + (1-w)\tilde{G}(b).$$

For any communication protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$, and for any $\beta > 0$, Alice and Bob are

$$\left(\frac{8}{c^2}\beta + 16\tilde{G}^*\left(\left(\frac{\epsilon}{\beta}\right)^{1/(1-\log_2 c)}\right)\right)\text{-accurate}$$

after the protocol terminates.

Proof. To prove Theorem 4.18, it suffices to prove the following lemma.

Lemma 4.19. *Let G be a differentiable convex function on $[0,1]$ and $c \in (0,1)$ be such that $\mathcal{JB}_G$ satisfies the c-approximate triangle inequality. Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies rectangle substitutes with respect to G. Let $\epsilon = \mathbb{E}\left[\mathcal{JB}_G(\mu_\sigma \parallel \mu_\tau)\right]$. Then for any $\beta > 0$, we have*

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_\tau)\right] \le \frac{8}{c^2}\beta + 16\tilde{G}^*\left(\left(\frac{\epsilon}{\beta}\right)^{1/(1-\log_2 c)}\right).$$

Let us first prove Theorem 4.18 assuming Lemma 4.19 is true.

Consider any protocol that causes Alice and Bob to ϵ-agree on $\mathcal{I}$. Let S be the set of possible signals of Alice at the end of the protocol which are consistent with the protocol transcript, and define T likewise for Bob.

Let $\epsilon_{ST} = \mathbb{E}\left[\mathrm{JB}_G(\mu_{\sigma T}, \mu_{S\tau}) \mid S, T\right]$. Note that

$$\mathbb{E}_{S,T}\left[\epsilon_{ST}\right] = \mathbb{E}_{S,T}\left[\mathbb{E}\left[\mathrm{JB}_G(\mu_{\sigma T}, \mu_{S\tau}) \mid S, T\right]\right] = \mathbb{E}\left[\mathrm{JB}_G(\mu_{\sigma T}, \mu_{S\tau})\right] \le \epsilon.$$

Therefore, for any $\beta > 0$ we have

$$\begin{aligned}
\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_{S\tau})\right] &\le \mathbb{E}_{S,T}\left[\frac{8}{c^2}\beta + 16\tilde{G}^*\left(\left(\frac{\epsilon_{ST}}{\beta}\right)^{1/(1-\log_2 c)}\right)\right] \\
&\le \frac{8}{c^2}\beta + 16\tilde{G}^*\left(\mathbb{E}_{S,T}\left[\left(\frac{\epsilon_{ST}}{\beta}\right)^{1/(1-\log_2 c)}\right]\right) \\
&\le \frac{8}{c^2}\beta + 16\tilde{G}^*\left(\left(\frac{\mathbb{E}_{S,T}\left[\epsilon_{ST}\right]}{\beta}\right)^{1/(1-\log_2 c)}\right) \\
&\le \frac{8}{c^2}\beta + 16\tilde{G}^*\left(\left(\frac{\epsilon}{\beta}\right)^{1/(1-\log_2 c)}\right).
\end{aligned}$$

In the first step, we apply Lemma 4.19 to the information structure $\mathcal{I}$ restricted to $S \times T$—that is, to $\mathcal{I}' = (\Omega', \mathbb{P}', S, T, Y)$, where $\Omega' = \{\omega \in \Omega : \sigma \in S, \tau \in T\}$ and $\mathbb{P}'[\omega] = \mathbb{P}\left[\omega \mid \sigma \in S, \tau \in T\right]$. The next two steps follow by the convexity of $\tilde{G}^*$ and $x^{1/(1-\log_2 c)}$, respectively. □

The basic outline of the proof of Lemma 4.19 is similar to that of Lemma 3.2. Once again, we partition $[0,1]$ into N intervals. Analogously to Equation 4, and with $S^{(k(\sigma))}$ defined analogously, we find that

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_\tau)\right] \le \mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k(\sigma))}})\right] + \mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right].$$

As before, we wish to upper bound each summand. However, the fact that the Bregman divergence is now arbitrary introduces complications. First, it is no longer the case that we can directly relate the length of an interval to the Bregman divergence between its endpoints. Second, we consider functions G that become infinitely steep near 0 and 1 (such as the negative of Shannon entropy), which makes matters more challenging. This means that we need to be more careful when partitioning $[0,1]$ into N intervals: see Algorithm C.3 for our new approach. Additionally, bounding the second summand involves reasoning carefully about the behavior of the function G, which is responsible for the introduction of $\tilde{G}^*$ into the lemma statement. We defer the full proof of Lemma 4.19 to Appendix C.

5 Connections to Markets

In this work, we established a natural condition on information structures, *rectangle substitutes*, under which any agreement protocol results in accurate beliefs. As we saw, a particularly natural class of agreement protocols are *expectation-sharing* protocols, where Alice and Bob take turns stating their current expected value, or discretizations thereof.

Expectation-sharing protocols have close connections to financial markets. In markets, the actions of traders reveal partial information about their believed value for an asset, i.e., their expectation. Specifically, a trader's decision about whether to buy or sell, and how much, can be viewed as revealing a discretization of this expectation. In many theoretical models of markets (see e.g. [Ostrovsky, 2012]) traders eventually reach agreement. The intuition behind this phenomenon is that a trader who disagrees with the price leaves money on the table by refusing to trade. Our work thus provides a lens into a well-studied question:[13] when are market prices accurate? Our results can be viewed as generalizing and conceptually supporting the result presented in Kong and Schoenebeck [2023], under which convergence in a popular prediction market design implies full information revelation in the prices.

An important caveat, however, is that traders behave strategically, and may not disclose their true expected value. For example, a trader may choose to withhold information until a later point when doing so would be more profitable. Therefore, to interpret the actions of traders as revealing discretized versions of their expected value, one first has to understand the Bayes-Nash equilibria of the market. Chen and Waggoner [2016] studies conditions under which traders are incentivized to reveal all of their information on their first trading opportunity. They call a market equilibrium *all-rush* if every trader is incentivized to reveal their information immediately. Their main result, roughly speaking, is that there is an all-rush equilibrium if and only if the information structure satisfies *strong substitutes*—a different strengthening of the weak substitutes condition. This result is specific to settings in which traders have the option to reveal all of their information on their turn—a setting that would be considered trivial from the standpoint of communication theory.

An exciting question for further study is therefore: under what information structure conditions and market settings is it a Bayes-Nash equilibrium to follow an agreement protocol leading to accurate beliefs? In other words, what conditions give not only that agreement implies accuracy,

[13] This is related to the *efficient market hypothesis*, the problem of when market prices reflect all available information, which traces back at least to Fama [1970] and Hayek [1945]. Modern models of financial markets are often based on Kyle [1985]; we refer the reader to e.g. [Ostrovsky, 2012] and references therein for further information.

but also that the market incentivizes participants to *follow* the protocol? Together with Chen and Waggoner [2016], our work suggests that certain substitutes-like conditions could suffice.

References

Scott Aaronson. 2005. The complexity of agreement. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, Harold N. Gabow and Ronald Fagin (Eds.). ACM, Baltimore, MD, USA, May 22-24, 2005, 634–643. https://doi.org/10.1145/1060590.1060686

Sreangsu Acharyya, Arindam Banerjee, and Daniel Boley. 2013. Bregman Divergences and Triangle Inequality. In *Proceedings of the 13th SIAM International Conference on Data Mining*. SIAM, May 2-4, 2013. Austin, Texas, USA, 476–484. https://doi.org/10.1137/1.9781611972832.53

Robert J. Aumann. 1976. Agreeing to Disagree. *The Annals of Statistics* 4, 6 (1976), 1236–1239. http://www.jstor.org/stable/2958591

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. 2005. Clustering with Bregman divergences. *Journal of Machine Learning Research* 6 (2005), 1705–1749.

P. Chen, Y. Chen, and M. Rao. 2008. Metrics defined by Bregman Divergences. *Communications in Mathematical Sciences* 6, 4 (2008), 915 – 926.

Yiling Chen and Bo Waggoner. 2016. Informational Substitutes. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016*, Irit Dinur (Ed.). IEEE Computer Society, Hyatt Regency, New Brunswick, New Jersey, USA, 239–247. https://doi.org/10.1109/FOCS.2016.33

Dominik Maria Endres and Johannes E. Schindelin. 2003. A new metric for probability distributions. *IEEE Trans. Inf. Theory* 49, 7 (2003), 1858–1860. https://doi.org/10.1109/TIT.2003.813506

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25, 2 (1970), 383–417.

John D Geanakoplos and Heraklis M Polemarchakis. 1982. We can't disagree forever. *Journal of Economic Theory* 28, 1 (1982), 192–200.

Friedrich August Hayek. 1945. The use of knowledge in society. *The American economic review* 35, 4 (1945), 519–530.

Yuqing Kong and Grant Schoenebeck. 2023. False Consensus, Information Theory, and Prediction Markets. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023 (LIPIcs, Vol. 251)*, Yael Tauman Kalai (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, MIT, Cambridge, Massachusetts, USA, 81:1–81:23. https://doi.org/10.4230/LIPIcs.ITCS.2023.81

Albert S. Kyle. 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society* 53, 6 (1985), 1315–1335.

Eric Neyman and Tim Roughgarden. 2021a. Are You Smarter Than a Random Expert? The Robust Aggregation of Substitutable Signals. (11 2021).

Eric Neyman and Tim Roughgarden. 2021b. From Proper Scoring Rules to Max-Min Optimal Forecast Aggregation. In *EC '21: The 22nd ACM Conference on Economics and Computation, July 18-23, 2021*, Péter Biró, Shuchi Chawla, and Federico Echenique (Eds.). ACM, Budapest, Hungary, 734. https://doi.org/10.1145/3465456.3467599

Michael Ostrovsky. 2012. Information aggregation in dynamic markets with strategic traders. *Econometrica* 80, 6 (2012), 2595–2647.

Anup Rao and Amir Yehudayoff. 2020. *Communication Complexity: and Applications.* Cambridge University Press. https://doi.org/10.1017/9781108671644

Z. Šidák. 1957. On Relations Between Strict-Sense and Wide-Sense Conditional Expectations. *Theory of Probability and Its Applications* 2 (1957), 267–272.

A Details Omitted From Section 2

Above we claimed that a positive measure of $n \times n$ information structures satisfy rectangle substitutes. To formalize this claim, we choose a natural measure over $n \times n$ information structures, specified via the following probability distribution over the values of Y and $\mathbb{P}[\sigma, \tau]$:

- Alice has signals labeled $\sigma_0, \ldots, \sigma_{n-1}$; Bob has signals labeled $\tau_0, \ldots, \tau_{n-1}$. Correspondingly, there are n^2 states which we identify with the pair (i, j). For each i, j, whenever $\sigma = \sigma_i$ and $\tau = \tau_j$, $Y = y(i, j)$ where $y(i, j)$ is uniformly random in $[0, 1]$.
- The probability distribution over states (i, j) is selected uniformly from the space of probability distributions over n^2 states.

Theorem A.1. *For every n, a positive measure of $n \times n$ information structures (per the above measure) satisfy the rectangle substitutes condition.*

Proof. The proof is conceptually quite simple. It suffices to exhibit an information structure in which the weak substitutes condition (i.e. Equation 1) holds *strictly* for every S, T such that $|S|, |T| \geq 2$. It then follows that for a sufficiently small δ, every information structure in the δ-ball around this one[14] also satisfies rectangle substitutes, completing the proof.[15]

The information structure $\mathcal{I}$ that we exhibit is as follows: choose any increasing, strictly concave function $f : [0, 2(n-1)] \to \mathbb{R}$ (for example, $f(x) = \sqrt{x}$). Let $y(i,j) = \frac{i+j}{2n}$, and let $\mathbb{P}[(i,j)]$ be proportional to $\epsilon^{f(i+j)}$.

For convenience, define the *substitutes slack* of an information structure to be the additive margin by which the information structure satisfies weak substitutes, i.e. the right-hand side of Equation 1 minus the left-hand side for $S = \mathcal{S}$ and $T = \mathcal{T}$.

Fix a particular S and T such that $|S|, |T| \geq 2$. We wish to show that for sufficiently small positive values of ϵ, Equation 1 holds strictly. We will show that the substitutes slack of $\mathcal{I}|_{S,T}$, i.e. $\mathcal{I}$ restricted to $S \times T$, is positive when ϵ is sufficiently small.

In order to prove this, we first consider the following (different) information structure for values $v, a, a', b, b', c, x, x', y, y'$ (obeying comparisons that we specify below). Each row corresponds to a possible signal value σ for Alice, and each column a possible signal value τ for Bob.

$$Y = \begin{bmatrix} v & v+b & b' \\ v+a & v+a+b & - \\ a' & - & - \end{bmatrix} \qquad \text{with probability proportional to} \qquad \begin{bmatrix} 1 & y & y' \\ x & cxy & 0 \\ x' & 0 & 0 \end{bmatrix}$$

In this information structure, suppose that $x' \leq x \ll 1$; $y' \leq y \ll 1$; and $1 \ll c \leq \frac{1}{x}, \frac{1}{y}$ (so $xy \ll cxy \leq x, y$). It can be verified (e.g. with a computer algebra system) that the substitutes slack of this information structure is $2abcxy + O(xy)$.

We will transform this information structure into $\mathcal{I}_{S,T}$ while (approximately) preserving substitutes slack. To foreshadow the correspondence, define i_S and i'_S be the smallest and second smallest values of i such that $\sigma_i \in S$, and define j_T and j'_T analogously. The rows of the information structure above will correspond to $\sigma = \sigma_{i_S}, \sigma_{i'_S}$, and all other values of $\sigma \in S$, in that order; the columns will correspond to $\tau = \tau_{j_T}, \tau_{j'_T}$, and all other values of $\tau \in T$, in that order.

Set $x := \epsilon^{f(i'_S+j_T)-f(i_S+j_T)}$, $y := \epsilon^{f(i_S+j'_T)-f(i_S+j_T)}$, and $c := \epsilon^{f(i_S+j_T)+f(i'_S+j'_T)-f(i_S+j'_T)-f(i'_S+j_T)}$, so that $cxy = \epsilon^{f(i'_S+j'_T)-f(i_S+j_T)}$. Note that these values satisfy the aforementioned inequalities involving x, y, and c. (The fact that $1 \ll c$ follows from the strict concavity of f.) Set $v := \frac{i_S}{2n}$, $a := \frac{i'_S - i_S}{2n}$, and $b := \frac{j'_T - j_T}{2n}$. Set x' so that $\mathbb{P}\left[i > i'_S \mid i \in S, j = j_T\right] = \frac{x'}{1+x+x'}$ and y' so that $\mathbb{P}\left[j > j'_T \mid i = i_S, j \in T\right] = \frac{y'}{1+y+y'}$. We set $a' := \mathbb{E}\left[y(i, j_T) \mid i > i'_S, i \in S\right]$ and $b' := \mathbb{E}\left[y(i_S, j) \mid j > j'_T, j \in T\right]$.

We now make the following transformation to this information structure: we replace the third row with $|S| - 2$ rows, each corresponding to a different $i > i'_S$. As before, each signal will only be possible in conjunction with Bob's first signal; the value of Y for the signal corresponding to σ_i in $\mathcal{I}$ will be $\frac{i+j_T}{2n}$, and the probability will be $\mathbb{P}[(i, j_T)]$. Note that this simply "splits" Alice's third signal into multiple (more informative) signals while preserving the total probability and

[14] We can for example define the distance between information structures $\mathcal{I}$ and $\mathcal{I}'$ as $\sum_{i,j} (y(i,j) - y'(i,j))^2 + (\mathbb{P}[(i,j)] - \mathbb{P}'[(i,j)])^2$.

[15] This uses the continuity of the terms in Equation 1. Note that the continuity of conditional expectations relies on the conditioning events having positive probability, as is the case in the information structure that we exhibit. Note also that we need not concern ourselves with cases in which $|S| = 1$ or $|T| = 1$, since in those cases the equation is necessarily an equality.

expectation (this is due to how we picked a', b', x', y' above). This does not affect the substitutes slack of the information structure, because the value of Bob's signal does not change as a result of the transformation (regardless of whether Alice's signal is known).

We make the same transformation but this time to Bob, replacing the third column with $|T| - 2$ columns. The transformation is otherwise analogous, and the substitutes slack again does not change.

Finally, in our last transformation we make this information structure match $\mathcal{I}$ exactly. Note that the information structures already match in the first row ($i = i_S$), and in the first column ($j = j_T$), and in the (second row, second column) entry ($(i, j) = (i'_S, j'_T)$). All other entries in $\mathcal{I}$ have probabilities that are $o(cxy)$ (recall that $cxy = \epsilon^{f(i'_S+j'_T)-f(i_S+j_T)}$). As a consequence, adding these entries to the information structure that we are transforming only changes the substitutes slack by $o(cxy)$.

Therefore, $\mathcal{I}$ has substitutes slack $2abcxy + o(cxy) \geq \frac{2}{n^2}\epsilon^{f(i'_S+j'_T)-f(i_S+j_T)}(1 + o(1))$. This is positive for ϵ sufficiently small, as desired.

We complete the proof by setting ϵ to be such that it is sufficiently small (in the above argument) for all S, T such that $|S|, |T| \geq 2$. □

B Details Omitted From Section 3

Proof of Claim 3.3. We claim that in fact we can choose the x_i's so that each x_i is in $\left[\frac{i}{N} - \frac{1}{2N}, \frac{i}{N} + \frac{1}{2N}\right]$. This ensures that each interval has length at most $\frac{2}{N}$.

For $x \in [0, 1]$, let $\rho(x)$ be the probability that x is between μ_σ and μ_τ, inclusive. Note that $\mathbb{P}[k(\sigma) \neq k(\tau)] \leq \sum_{i=1}^{N-1} \rho(x_i)$.

Observe that if x is selected uniformly from $[0, 1]$, the expected value of $\rho(x)$ is equal to $|\mu_\sigma - \mu_\tau|$, because both quantities are equal to the probability that x is between μ_σ and μ_τ. Therefore, if (σ, τ) is additionally chosen according to $\mathbb{P}$, we have

$$\mathbb{E}_{x\leftarrow[0,1]}[\rho(x)] = \mathbb{E}[|\mu_\sigma - \mu_\tau|] \leq \sqrt{\mathbb{E}[(\mu_\sigma - \mu_\tau)^2]} = \sqrt{\epsilon}.$$

This means that

$$\sum_{i=1}^{N-1} \mathbb{E}_{x\leftarrow\left[\frac{i}{N}-\frac{1}{2N}, \frac{i}{N}+\frac{1}{2N}\right]}[\rho(x)] = (N-1)\mathbb{E}_{x\leftarrow\left[\frac{1}{2N}, 1-\frac{1}{2N}\right]}[\rho(x)] \leq \sqrt{\epsilon}N.$$

Thus, if each x_i is selected uniformly at random from $\left[\frac{i}{N} - \frac{1}{2N}, \frac{i}{N} + \frac{1}{2N}\right]$, the expected value of $\mathbb{P}[k(\sigma) \neq k(\tau)]$ would be at most $\sqrt{\epsilon}N$. In particular this means that there exist choices of the x_i's such that $\mathbb{P}[k(\sigma) \neq k(\tau)] \leq \sqrt{\epsilon}N$. □

Proof of Lemma 3.2. Fix a large positive integer N (we will later find it optimal to set $N = \epsilon^{-1/6}$). Consider a partition of $[0, 1]$ into N intervals $[0, x_1), [x_1, x_2), \ldots, [x_{N-1}, 1]$ satisfying the conditions of Claim 3.3. Let $S^{(k)} := \{\sigma \in \mathcal{S} : x_{k-1} \leq \mu_\sigma < x_k\}$. Additionally, let $k(\sigma)$ and $k(\tau)$ be as defined in Claim 3.3.

Our goal is to upper bound the expectation of $(\mu_{\sigma\tau} - \mu_\tau)^2$. In pursuit of this goal, we observe that by the Pythagorean theorem, we have

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] = \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S^{(k(\sigma))}\tau})^2\right] + \mathbb{E}\left[(\mu_{S^{(k(\sigma))}\tau} - \mu_\tau)^2\right].$$

We now use the rectangle substitutes assumption: for any k, by applying Equation 2 to $S = S^{(k)}$ and $T = \mathcal{T}$, we know that

$$\mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k)}})^2 \mid \sigma \in S^{(k)}\right] \geq \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S^{(k)}\tau})^2 \mid \sigma \in S^{(k)}\right].$$

Taking the expectation over k (i.e. choosing each k with probability equal to $\mathbb{P}\left[\sigma \in S^{(k)}\right]$), we have that

$$\mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k(\sigma))}})^2\right] \geq \mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S^{(k(\sigma))}\tau})^2\right]. \tag{7}$$

Therefore, we have

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] \leq \mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k(\sigma))}})^2\right] + \mathbb{E}\left[(\mu_{S^{(k(\sigma))}\tau} - \mu_\tau)^2\right]. \tag{8}$$

We will use Claim 3.3 to argue that each of these two summands is small. The argument regarding the first summand is straightforward: for any σ, we have that $x_{k(\sigma)} \leq \mu_\sigma, \mu_{S^{(k(\sigma))}} < x_{k(\sigma)+1} \leq x_{k(\sigma)} + \frac{2}{N}$, which means that $\mathbb{E}\left[(\mu_\sigma - \mu_{S^{(k(\sigma))}})^2\right] \leq \frac{4}{N^2}$.

We now upper bound the second summand.[16] For any $\hat{\tau} \in \mathcal{T}$, let $p(\hat{\tau}) = \mathbb{P}\left[\tau = \hat{\tau}\right]$ and $q(\hat{\tau}) = \mathbb{P}\left[\tau = \hat{\tau}, k(\sigma) \neq k(\tau)\right]$. Then $\sum_{\hat{\tau}\in\mathcal{T}} p(\hat{\tau}) = 1$ and $\sum_{\hat{\tau}\in\mathcal{T}} q(\hat{\tau}) \leq \sqrt{\epsilon}N$. Observe that

$$\begin{aligned}
\mathbb{E}\left[(\mu_{S^{(k(\sigma))}\tau} - \mu_\tau)^2\right] &= \sum_{\hat{\tau}} p(\hat{\tau})\mathbb{E}\left[(\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\hat{\tau}})^2 \mid \tau = \hat{\tau}\right] \\
&= \sum_{\hat{\tau}} (p(\hat{\tau}) - q(\hat{\tau}))\mathbb{E}\left[(\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\hat{\tau}})^2 \mid \tau = \hat{\tau}, k(\sigma) = k(\hat{\tau})\right] \\
&\quad + q(\hat{\tau})\mathbb{E}\left[(\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\hat{\tau}})^2 \mid \tau = \hat{\tau}, k(\sigma) \neq k(\hat{\tau})\right].
\end{aligned} \tag{9}$$

To handle the first expectation, we note that if $k(\sigma) = k(\hat{\tau})$, then $\left|\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\hat{\tau}}\right| \leq \frac{q(\hat{\tau})}{p(\hat{\tau})}$. To see this, observe

$$p(\hat{\tau})\mu_{\hat{\tau}} = (p(\hat{\tau}) - q(\hat{\tau}))\mu_{S^{(k(\sigma))}\hat{\tau}} + q(\hat{\tau})\mu_{\mathcal{S}\setminus S^{(k(\sigma))}\hat{\tau}}.$$

Rerranging and taking absolute values, we conclude

$$p(\hat{\tau})\left|\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\hat{\tau}}\right| = q(\hat{\tau})\left|\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\mathcal{S}\setminus S^{(k(\sigma))}}\right| \leq q(\hat{\tau}).$$

Therefore, recalling $q(\hat{\tau}) \leq p(\hat{\tau})$, we have

$$(p(\hat{\tau}) - q(\hat{\tau}))\mathbb{E}\left[(\mu_{S^{(k(\sigma))}\hat{\tau}} - \mu_{\hat{\tau}})^2 \mid \tau = \hat{\tau}, k(\sigma) = k(\hat{\tau})\right] \leq (p(\hat{\tau}) - q(\hat{\tau}))\left(\frac{q(\hat{\tau})}{p(\hat{\tau})}\right)^2 \leq \frac{q(\hat{\tau})^2}{p(\hat{\tau})} \leq q(\hat{\tau}).$$

On the other hand, we can bound the second expectation in Equation 9 by 1. Therefore we have

$$\mathbb{E}\left[(\mu_{S^{(k(\sigma))}\tau} - \mu_\tau)^2\right] \leq \sum_{\hat{\tau}} (q(\hat{\tau}) + q(\hat{\tau})) = 2\sum_{\hat{\tau}} q(\hat{\tau}) \leq 2\sqrt{\epsilon}N.$$

To conclude, we now know that

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_\tau)^2\right] \leq \frac{4}{N^2} + 2\sqrt{\epsilon}N.$$

Setting $N = \epsilon^{-1/6}$ makes the right-hand side equal to $6\epsilon^{1/3}$, completing the proof. □

Proposition B.1. *Consider the following protocol, parametrized by $\epsilon > 0$. Alice and Bob send their initial expectations to each other, rounding to the nearest multiple of ϵ. This protocol entails communicating $O(\log 1/\epsilon)$ bits. At the end of the protocol, Alice and Bob $2\epsilon^2$-agree and are ϵ^2-accurate (with respect to $G(x) = x^2$).*

Proof. Let S be the set of possible signals of Alice at the end of the protocol which are consistent with the protocol transcript, and define T likewise for Bob. Recall that we use $\mathcal{S}$ and $\mathcal{T}$ to denote the sets of all of Alice's and Bob's possible signals, respectively. We have

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2\right] \leq \mathbb{E}\left[(\mu_\sigma - \mu_S)^2\right] \leq \epsilon^2,$$

[16] The proof below takes sums over $\hat{\tau} \in \mathcal{T}$ and thus implicitly assumes that $\mathcal{T}$ is finite, but the proof extends to infinite $\mathcal{T}$, with sums over τ replaced by integrals with respect to the probability measure over $\mathcal{T}$.

since μ_σ and μ_S) are guaranteed to be within ϵ of each other by construction. Thus, Bob is ϵ^2-accurate, and likewise for Alice. By the $\frac{1}{2}$-approximate triangle inequality for $G(x) = x^2$, it follows that Alice and Bob $2\epsilon^2$-agree. □

C Details Omitted From Section 4

Proof of Proposition 4.8. Let $g = G'$. We have

$$\begin{aligned}
&\mathbb{E}\left[D_G(A \parallel B)\right] + \mathbb{E}\left[D_G(B \parallel C)\right] - \mathbb{E}\left[D_G(A \parallel C)\right] \\
&= \mathbb{E}\left[G(A) - G(B) - (A-B)g(B) + G(B) - G(C) - (B-C)g(C) - G(A) + G(C) + (A-C)g(C)\right] \\
&= \mathbb{E}\left[(A-B)(g(C) - g(B))\right] = \mathbb{E}\left[\mathbb{E}\left[(A-B)(g(C) - g(B)) \mid \mathcal{F}\right]\right] \\
&= \mathbb{E}\left[(g(C) - g(B))\mathbb{E}\left[A - B \mid \mathcal{F}\right]\right] = \mathbb{E}\left[(g(C) - g(B))(\mathbb{E}\left[A \mid \mathcal{F}\right] - B)\right] = 0.
\end{aligned}$$

The third-to-last step follows from the fact that $g(C) - g(B)$ is $\mathcal{F}$-measurable (we are using the "pulling out known factors" property of conditional expectation). The last step follows from the fact that $\mathbb{E}\left[A \mid \mathcal{F}\right] = B$. □

Proposition C.1. *Let G be a differentiable convex function on the interval $[0,1]$. For all $0 \le a \le b \le 1$, we have*

(i) $\frac{1}{2}(D_G(a \parallel x) + D_G(b \parallel x)) \ge \mathcal{JB}_G(a,b)$ *for every* $x \in [0,1]$.
(ii) $\mathcal{JB}_G$ *satisfies the reverse triangle inequality: for every* $x \in [a,b]$, *we have* $\mathcal{JB}_G(a,x) + \mathcal{JB}_G(x,b) \le \mathcal{JB}_G(a,b)$.
(iii) For all $a \le a' \le b' \le b$, *we have* $\mathcal{JB}_G(a',b') \le \mathcal{JB}_G(a,b)$.
(iv) For a random variable X *supported on* $[a,b]$, *we have*

$$\mathbb{E}\left[D_G(X \parallel \mathbb{E}\left[X\right])\right] = \mathbb{E}\left[G(X)\right] - G(\mathbb{E}\left[X\right]) \le 2\mathcal{JB}_G(a,b).$$

Proof. Fact (i) follows from Proposition 4.2. Regarding Fact (ii), without loss of generality assume that $x \le \frac{a+b}{2}$ and that $G(x) = G\left(\frac{a+b}{2}\right)$ (uniformly adding a constant to the derivative of G does not change any Jensen-Bregman divergence, hence the second assumption). Then $G\left(\frac{a+x}{2}\right) \ge G(x)$, so $\mathrm{JB}_G(a,x) \le \frac{G(a)-G(x)}{2}$. Since $\frac{b+x}{2} \ge \frac{a+b}{2}$, we also have that $G\left(\frac{b+x}{2}\right) \ge G(x)$, so $\mathrm{JB}_G(b,x) \le \frac{G(b)-G(x)}{2}$. Thus, we have

$$\mathrm{JB}_G(a,x) + \mathrm{JB}_G(b,x) \le \frac{G(a)+G(b)}{2} - G(x) = \frac{G(a)+G(b)}{2} - G\left(\frac{a+b}{2}\right) = \mathrm{JB}_G(a,b).$$

Fact (iii) follows from Fact (ii): we have

$$\mathrm{JB}_G(a,b) = \mathrm{JB}_G(a,a') + \mathrm{JB}_G(a',b') + \mathrm{JB}_G(b',b) \ge \mathrm{JB}_G(a',b').$$

Regarding the equality in Fact (iv), we have

$$\begin{aligned}
\mathbb{E}\left[D_G(X \parallel \mathbb{E}\left[X\right])\right] &= \mathbb{E}\left[G(X) - G(\mathbb{E}\left[X\right]) - (X - \mathbb{E}\left[X\right])G'(\mathbb{E}\left[X\right])\right] \\
&= \mathbb{E}\left[G(X) - G(\mathbb{E}\left[X\right])\right] = \mathbb{E}\left[G(X)\right] - G(\mathbb{E}\left[X\right]),
\end{aligned}$$

where the first step follows from the fact that $\mathbb{E}\left[(X - \mathbb{E}\left[X\right])G'(\mathbb{E}\left[X\right])\right] = G'(\mathbb{E}\left[X\right])\mathbb{E}\left[X - \mathbb{E}\left[X\right]\right]$, and $\mathbb{E}\left[X - \mathbb{E}\left[X\right]\right] = 0$.

Regarding the inequality in Fact (iv), without loss of generality assume that $\mathbb{E}\left[X\right] \le \frac{a+b}{2}$. By convexity we have that

$$G\left(\frac{a+b}{2}\right) \le \frac{\frac{b-a}{2}}{b - \mathbb{E}\left[X\right]} G(\mathbb{E}\left[X\right]) + \frac{\frac{a+b}{2} - \mathbb{E}\left[X\right]}{b - \mathbb{E}\left[X\right]} G(b),$$

so

$$\begin{aligned}
\mathrm{JB}_G(a,b) &= G(a) + G(b) - 2G\left(\frac{a+b}{2}\right) \\
&\geq G(a) + G(b) - \frac{b-a}{b-\mathbb{E}[X]}G(\mathbb{E}[X]) - \frac{a+b-2\mathbb{E}[X]}{b-\mathbb{E}[X]}G(b) \\
&= G(a) + \frac{\mathbb{E}[X]-a}{b-\mathbb{E}[X]}G(b) - \frac{b-a}{b-\mathbb{E}[X]}G(\mathbb{E}[X]) \\
&= \frac{b-a}{b-\mathbb{E}[X]}\left(\frac{b-\mathbb{E}[X]}{b-a}G(a) + \frac{\mathbb{E}[X]-a}{b-a}G(b) - G(\mathbb{E}[X])\right) \\
&\geq \frac{b-\mathbb{E}[X]}{b-a}G(a) + \frac{\mathbb{E}[X]-a}{b-a}G(b) - G(\mathbb{E}[X]) \geq \mathbb{E}[G(X)] - G(\mathbb{E}[X]).
\end{aligned}$$

In the last step we use the fact that for a convex function f and a random variable X defined on an interval $[a,b]$ with mean μ, the maximum possible value of $\mathbb{E}[f(X)]$ is attained if X is either a or b with the appropriate probabilities. □

Proof of Theorem 4.11. Suppose that Alice and Bob do not ϵ-agree at time step t, and without loss of generality assume that the next turn (number $t+1$) is Alice's. We begin by observing that, by Proposition C.1 (i), we have

$$\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) + D_G(\mu_{S_t \tau} \parallel \mu_{S_t T_t})\right] \geq 2\mathbb{E}\left[\mathrm{JB}_G(\mu_{\sigma T_t}, \mu_{S_t \tau})\right] > 2\epsilon.$$

Therefore, either $\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t})\right] \geq \frac{2\epsilon}{3}$ or $\mathbb{E}\left[D_G(\mu_{S_t \tau} \parallel \mu_{S_t T_t})\right] \geq \frac{4\epsilon}{3}$.

Case 1: $\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t})\right] \geq \frac{2\epsilon}{3}$. Let us use "hi," "lo," and "md" to denote the events that Alice says "high," Alice says "low," and Alice says "medium," respectively. We have

$$\begin{aligned}
\frac{2\epsilon}{3} &\leq \mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t})\right] = \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \mid S_t, T_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{hi or lo}} \mid S_t, T_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{md}} \mid S_t, T_t\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{hi or lo}} \mid S_t, T_t\right]\right] + \frac{\epsilon}{2},
\end{aligned}$$

where "$\mid S_t, T_t$" is short for "$\mid \sigma \in S_t, \tau \in T_t$," a notation we use throughout the proof. We thus have

$$\mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{hi}} \mid S_t, T_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{lo}} \mid S_t, T_t\right]\right] \geq \frac{\epsilon}{6}. \tag{10}$$

We now make use of the following lemma.

Lemma C.2. *Suppose that turn $t+1$ is Alice's. Let "hi" denote the event that Alice says "high." Let $\alpha := \mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{hi} \mid S_t, T_t\right]$. Then*

$$\mathbb{E}\left[D_G(\mu_{S_{t+1} T_{t+1}} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{hi} \mid S_t, T_t\right] \geq \frac{\alpha\epsilon}{8M + 2\epsilon}.$$

The analogous statement is true if Alice says "low," and likewise if it is instead Bob's turn.

We assume Lemma C.2 and return to prove it afterward. This lemma translates Equation 10 into a statement about how much Charlie learns. Specifically, we have that

$$\begin{aligned}
&\mathbb{E}\left[D_G(\mu_{S_{t+1} T_{t+1}} \parallel \mu_{S_t T_t})\right] = \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{S_{t+1} T_{t+1}} \parallel \mu_{S_t T_t}) \mid S_t, T_t\right]\right] \\
&\geq \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{S_{t+1} T_{t+1}} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{hi}} \mid S_t, T_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{S_{t+1} T_{t+1}} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{lo}} \mid S_t, T_t\right]\right] \\
&\geq \frac{\epsilon}{8M + 2\epsilon}\left(\mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{hi}} \mid S_t, T_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[D_G(\mu_{\sigma T_t} \parallel \mu_{S_t T_t}) \cdot \mathbb{1}_{\text{lo}} \mid S_t, T_t\right]\right]\right)
\end{aligned}$$

$$\geq \frac{\epsilon^2}{6(8M+2\epsilon)}.$$

Case 2: $\mathbb{E}\left[D_G(\mu_{S_t\tau} \parallel \mu_{S_tT_t})\right] \geq \frac{4\epsilon}{3}$. Using the Pythagorean theorem to write the same Bregman divergence in two ways, we have that

$$\mathbb{E}\left[D_G(\mu_{S_{t+1}\tau} \parallel \mu_{S_{t+1}T_{t+1}})\right] + \mathbb{E}\left[D_G(\mu_{S_{t+1}T_{t+1}} \parallel \mu_{S_tT_t})\right] = \mathbb{E}\left[D_G(\mu_{S_{t+1}\tau} \parallel \mu_{S_tT_t})\right]$$
$$= \mathbb{E}\left[D_G(\mu_{S_{t+1}\tau} \parallel \mu_{S_t\tau})\right] + \mathbb{E}\left[D_G(\mu_{S_t\tau} \parallel \mu_{S_tT_t})\right] \geq \mathbb{E}\left[D_G(\mu_{S_t\tau} \parallel \mu_{S_tT_t})\right] \geq \frac{4\epsilon}{3}.$$

This means that one of the two summands on the left-hand side is at least $\frac{2\epsilon}{3}$.

Case 2a: $\mathbb{E}\left[D_G(\mu_{S_{t+1}\tau} \parallel \mu_{S_{t+1}T_{t+1}})\right] \geq \frac{2\epsilon}{3}$. In that case we have that

$$\mathbb{E}\left[D_G(\mu_{S_{t+2}T_{t+2}} \parallel \mu_{S_{t+1}T_{t+1}})\right] \geq \frac{\epsilon^2}{6(8M+2\epsilon)}$$

by the same logic as in Case 1.

Case 2b: $\mathbb{E}\left[D_G(\mu_{S_{t+1}T_{t+1}} \parallel \mu_{S_tT_t})\right] \geq \frac{2\epsilon}{3} \geq \frac{\epsilon^2}{12\epsilon} \geq \frac{\epsilon^2}{6(8M+2\epsilon)}$.

In each of our cases, we have that

$$\mathbb{E}\left[D_G(Y \parallel \mu_{S_tT_t}) - D_G(Y \parallel \mu_{S_{t+2}T_{t+2}})\right] = \mathbb{E}\left[D_G(\mu_{S_{t+2}T_{t+2}} \parallel \mu_{S_tT_t})\right]$$
$$= \mathbb{E}\left[D_G(\mu_{S_{t+2}T_{t+2}} \parallel \mu_{S_{t+1}T_{t+1}})\right] + \mathbb{E}\left[D_G(\mu_{S_{t+1}T_{t+1}} \parallel \mu_{S_tT_t})\right] \geq \frac{\epsilon^2}{6(8M+2\epsilon)}.$$

Therefore, the total number of steps until agreement is first reached cannot be more than

$$2 \cdot \frac{M}{\frac{\epsilon^2}{6(8M+2\epsilon)}} = \frac{24M(4M+\epsilon)}{\epsilon^2}.$$

This completes the proof. □

We now prove Lemma C.2.

Proof of Lemma C.2. We will restrict our probability space to outcomes where Charlie knows S_t, T_t at time t (and thus omit "$\mid S_t, T_t$" from here on). For convenience, we will let $A := \mu_{\sigma T_t}$ be Alice's expectation (a random variable) and $c := \mu_{S_tT_t}$ be Charlie's expectation (which is a particular number in $[0,1]$). We will let $\epsilon' := \frac{\epsilon}{2}$, so that if Alice says "high" then Charlie knows that $A > c$ and that $D_G(A \parallel c) \geq \epsilon'$.

Let $D(x) := D_G(x \parallel c) = G(x) - G(c) - G'(c)(x-c)$, and let $\hat{a}_h := \mathbb{E}[A \mid \text{hi}]$. Note that if Alice says "high" then $\mu_{S_{t+1}T_{t+1}} = \hat{a}_h$. In our new notation, we may write $\alpha = \mathbb{E}[D(A) \mid \text{hi}] \cdot \mathbb{P}[\text{hi}]$, and we wish to show that $D(\hat{a}_h) \cdot \mathbb{P}[\text{hi}] \geq \frac{\alpha\epsilon'}{2(M+\epsilon')}$. Put otherwise, our goal is to show that

$$\frac{D(\hat{a}_h)}{\mathbb{E}[D(A) \mid \text{hi}]} \geq \frac{\epsilon'}{2(M+\epsilon')}.$$

For convenience we will let B denote the quantity on the left-hand side.

Let a_{hmin} be the number larger than c such that $D(a) = \epsilon'$, so that $A \geq a_{\text{hmin}}$ whenever Alice says "high."[17] Observe that since D is convex (Bregman divergences are convex in their first argument),

[17] If $D(a) < \epsilon'$ for all $a > c$ then Alice never says "high" and the lemma statement is trivial.

for a fixed value of $\hat{a}_h$, the value of $\mathbb{E}[D(A) \mid \text{hi}]$ is maximized when A is either a_{hmin} or 1 (with probabilities $\frac{1-\hat{a}_h}{1-a_{\text{hmin}}}$ and $\frac{\hat{a}_h - a_{\text{hmin}}}{1-a_{\text{hmin}}}$, respectively). Therefore we have

$$B = \frac{D(\hat{a}_h)}{\mathbb{E}[D(A) \mid \text{hi}]} \geq \frac{D(\hat{a}_h)(1-a_{\text{hmin}})}{(1-\hat{a}_h)\epsilon' + (\hat{a}_h - a_{\text{hmin}})D(1)}. \tag{11}$$

Case 1: $(1-\hat{a}_h)\epsilon' \geq (\hat{a}_h - a_{\text{hmin}})D(1)$. In that case we have

$$B \geq \frac{D(\hat{a}_h)(1-a_{\text{hmin}})}{2(1-\hat{a}_h)\epsilon'} \geq \frac{\epsilon'(1-a_{\text{hmin}})}{2(1-\hat{a}_h)\epsilon'} \geq \frac{1}{2} \geq \frac{\epsilon'}{2(M+\epsilon')}.$$

Case 2: $(1-\hat{a}_h)\epsilon' \leq (\hat{a}_h - a_{\text{hmin}})D(1)$. In that case we have

$$B \geq \frac{D(\hat{a}_h)(1-a_{\text{hmin}})}{2(\hat{a}_h - a_{\text{hmin}})D(1)}. \tag{12}$$

Case 2a: $D(1) \leq \frac{1-c}{\hat{a}_h - c}(M+\epsilon')$. Then we have

$$B \geq \frac{D(\hat{a}_h)(1-a_{\text{hmin}})}{2(\hat{a}_h - a_{\text{hmin}}) \cdot \frac{1-c}{\hat{a}_h - c}(M+\epsilon')} \geq \frac{\epsilon'}{2(M+\epsilon')} \cdot \frac{(1-a_{\text{hmin}})(\hat{a}_h - c)}{(\hat{a}_h - a_{\text{hmin}})(1-c)}.$$

(In the last step we again use that $D(\hat{a}_h) \geq \epsilon$.) Now, it is easy to verify that the second fraction is at least 1 (this comes down to the fact that $a_{\text{hmin}} \geq c$), so we indeed have that $B \geq \frac{\epsilon'}{2(M+\epsilon')}$.

Case 2b: $D(1) \geq \frac{1-c}{\hat{a}_h - c}(M+\epsilon')$. We claim that for all $x \geq c$, we have that

$$D(x) \geq \frac{x-c}{1-c}D(1) - M. \tag{13}$$

To see this, suppose for contradiction that for some x we have $D(x) < \frac{x-c}{1-c}D(1) - M$. Then

$$\begin{aligned} G(x) - G(c) - G'(c)(x-c) &< \frac{x-c}{1-c}(G(1) - G(c) - G'(c)(1-c)) - M \\ (1-c)G(x) - (1-c)G(c) &< (x-c)G(1) - (x-c)G(c) - (1-c)M \\ G(x) + M &< \frac{(1-x)G(c) + (x-c)G(1)}{1-c}. \end{aligned}$$

On the other hand, we have that both $G(c)$ and $G(1)$ are less than or equal to $G(x)+M$, by definition of M. This means that

$$G(1), G(c) < \frac{(1-x)G(c) + (x-c)G(1)}{1-c}$$

but this implies that $G(1) < G(c)$ and that $G(c) < G(1)$, a contradiction.

Plugging in $x = \hat{a}_h$ into Equation 13, we find that

$$D(\hat{a}_h) \geq \frac{\hat{a}_h - c}{1-c}D(1) - M.$$

Plugging this bound into Equation 12, we get that

$$\begin{aligned} B &\geq \frac{\left(\frac{\hat{a}_h - c}{1-c}D(1) - M\right)(1-a_{\text{hmin}})}{2(\hat{a}_h - a_{\text{hmin}})D(1)} = \frac{1-a_{\text{hmin}}}{2(\hat{a}_h - a_{\text{hmin}})} \cdot \frac{\hat{a}_h - c}{1-c}\left(1 - \frac{M}{\frac{\hat{a}_h - c}{1-c}D(1)}\right) \\ &\geq \frac{1-a_{\text{hmin}}}{2(\hat{a}_h - a_{\text{hmin}})} \cdot \frac{\hat{a}_h - c}{1-c}\left(1 - \frac{M}{M+\epsilon'}\right) \geq \frac{\epsilon'}{2(M+\epsilon')}, \end{aligned}$$

where in the second-to-last step we use that $D(1) \geq \frac{1-c}{\hat{a}_h - c}(M+\epsilon')$ and in the last step we again use the fact that $\frac{(1-a_{\text{hmin}})(\hat{a}_h - c)}{(\hat{a}_h - a_{\text{hmin}})(1-c)} \geq 1$. □

Proof of Lemma 4.19.
We will partition $[0,1]$ into a number N of small intervals $I_1 = [x_0 = 0, x_1), I_2 = [x_1, x_2), I_3 = [x_2, x_3)$, $\ldots$, $I_N = [x_{N-1}, x_N = 1]$ with certain desirable properties (which we will describe below). For $k \in [N]$, we will let $S^{(k)} := \{\sigma \in \mathcal{S} : \mu_\sigma \in I_k\}$. For a given $\sigma \in \mathcal{S}$, we will let $k(\sigma)$ be the k such that $\sigma \in S^{(k)}$.

Our goal is to upper bound the expectation of $D_G(\mu_{\sigma\tau} \parallel \mu_\tau)$. In pursuit of this goal, we observe that by Proposition 4.8 we have

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_\tau)\right] = \mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_{S^{(k(\sigma))}\tau})\right] + \mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right]. \tag{14}$$

Now, for any k, by applying Equation 6 to $S = S^{(k)}$ and $T = \mathcal{T}$, we know that

$$\mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k)}}) \mid S^{(k)}\right] \geq \mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_{S^{(k)}\tau}) \mid S^{(k)}\right].$$

(Here, "$\mid S^{(k)}$" is short for "$\mid \sigma \in S^{(k)}$.") This is our only use of the rectangle substitutes assumption. Now, taking the expectation over k (i.e. choosing each k with probability equal to $\mathbb{P}\left[\sigma \in S^{(k)}\right]$), we have that

$$\mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k(\sigma))}})\right] \geq \mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_{S^{(k(\sigma))}\tau})\right].$$

Together with Equation 14, this tells us that

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_\tau)\right] \leq \mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k(\sigma))}})\right] + \mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right]. \tag{15}$$

Our goal will be to bound the two summands in Equation 15. We will specify the boundaries of the intervals $I_1, \ldots, I_N$ with this goal in mind.

On an intuitive level, we are hoping for two things to be true:

- In order for the first summand to be small, we want μ_σ and $\mu_{S^{(k(\sigma))}}$ to be similar in value. In other words, we want each interval is "short" (for a notion of shortness with respect to G that we are about to discuss).
- In order for the second summand to be small, we want $\mu_{S^{(k(\sigma))}\tau}$ and μ_τ to be similar in value. In other words, the estimate of a third party who knows τ shouldn't change much upon learning $k(\sigma)$. One way to ensure this is by creating the intervals in a way that makes the third party very confident about the value of $k(\sigma)$ before learning it. Intuitively this should be true because Alice and Bob approximately agree, so Alice's estimate is likely to be close to Bob's. However, we must be careful to strategically choose the boundaries of our intervals $x_1, \ldots, x_{N-1}$ so that Alice's and Bob's estimates are unlikely to be on opposite sides of a boundary.[18]

What, formally, do we need for the first summand to be small? For any k, we have $\mu_{S^{(k(\sigma))}} = \mathbb{E}\left[\mu_\sigma \mid \sigma \in S^{(k)}\right]$. We can apply Proposition C.1 (iv) to the random variable $X = \mu_\sigma$ on the probability subspace given by $\sigma \in S^{(k)}$. Since X takes on values in I_k, we have that

$$\mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k)}}) \mid S^{(k)}\right] \leq 2\mathrm{JB}_G(I_k), \tag{16}$$

where $\mathrm{JB}_G(I_k)$ is shorthand for the Jensen-Bregman divergence between the endpoints of I_k. Therefore, if $\mathrm{JB}_G(I_k)$ is small for all k, then the first summand (which is an expected value of $\mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k)}}) \mid S^{(k)}\right]$ over $k \in [N]$) is also small.

What about the second summand? As per the intuition above, we wish to choose our boundary points $x_1, \ldots, x_{N-1}$ so that Alice's and Bob's estimates are unlikely to be on opposite sides of any

[18] This limits how many intervals we can reasonably use, which is why we cannot make our intervals arbitrarily short to satisfy the first of our two criteria.

boundary. Let $\mu_- = \min(\mu_\sigma, \mu_\tau)$ be the smaller of the two estimates and $\mu_+ = \max(\mu_\sigma, \mu_\tau)$ be the larger one. We say that μ_-, μ_+ *thwart* a point $x \in (0,1)$ if $\mu_- \le x \le \mu_+$ and $\mu_- \ne \mu_+$. We define the *thwart density* of x to be

$$\rho(x) := \mathbb{P}\left[\mu_-, \mu_+ \text{ thwart } x\right].$$

Roughly speaking, we will choose $x_1, \ldots, x_{N-1}$ such that $\rho(x_k)$ is small on average.

We will approach this problem by first creating intervals to satisfy the first criterion (short intervals), without regard to the second, and then modifying them to satisfy the second without compromising the first. Formally, we choose our intervals according to the following algorithm.

Algorithm C.3 (Partitioning $[0,1]$ into intervals $I_1, \ldots, I_N$)**.**

(1) Choose points $0 < x'_1 < x'_2 < \cdots < x'_{N-2} < 1$ such that the $N-1$ intervals thus created all have Jensen-Bregman divergence between β and $\frac{2\beta}{c}$, inclusive, where β and c are as in the statement of Lemma 4.19. (N is not pre-determined; it is defined as one more than the number of intervals created.) (See footnote for why this is possible.[19])
(2) Let $x'_0 := 0, x'_{N-1} := 1$ for convenience. Define $I'_k := [x'_{k-1}, x'_k]$. For $k \in [N-1]$, let $\alpha_k := \inf_{x \in I'_k} \rho(x)$. Let $x_k \in I'_k$ be such[20] that $\rho(x_k) \le 2\alpha_k$.
(3) Return the intervals $I_1 = [0, x_1), I_2 = [x_1, x_2), \ldots, I_N = [x_{N-1}, 1]$.

We begin by observing that for any $k \in [N]$, we have

$$\mathrm{JB}_G(I_k) = \mathrm{JB}_G(x_{k-1}, x_k) \le \mathrm{JB}_G(x'_{k-2}, x'_k) \le \frac{1}{c}\left(\mathrm{JB}_G(x'_{k-2}, x'_{k-1}) + \mathrm{JB}_G(x'_{k-1}, x'_k)\right) \le \frac{4\beta}{c^2}$$

where for convenience we define $x'_{-1} := 0, x'_N := 1$. Therefore, by Equation 16, we have

$$\mathbb{E}\left[D_G(\mu_\sigma \parallel \mu_{S^{(k(\sigma))}})\right] \le \frac{8\beta}{c^2}. \tag{17}$$

It remains to bound the second summand of Equation 15, $\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right]$, which is the bulk of the proof. We proceed in two steps:

(1) (Lemma C.4) We show that $\sum_{k=1}^N \alpha_k$ is small. This means that Alice's and Bob's estimates are unlikely to lie on opposite sides of some boundary point x_k. As a consequence, Bob is highly likely to know $k(\sigma)$ with a lot of confidence
(2) (Lemma C.7) We bound the second summand as a function of $\sum_{k=1}^N \alpha_k$. The intuition is that if $\sum_k \alpha_k$ is small, then Bob is highly likely to know $k(\sigma)$ with a lot of confidence, which means that he does not learn too much from learning $k(\sigma)$.

We begin with the first step; recall our notation $\mu^- := \min(\mu_\sigma, \mu_\tau)$ and $\mu^+ := \max(\mu_\sigma, \mu_\tau)$.

Lemma C.4.

$$2\sum_{k=1}^N \alpha_k \le 4\left(\frac{\epsilon}{\beta c}\right)^{1/(1-\log_2 c)}.$$

[19] Define x'_1 so that $\mathrm{JB}_G(0, x'_1) = \frac{2\beta}{c}$ (this is possible because JB_G is continuous in its arguments). Define x'_2 so that $\mathrm{JB}_G(x'_1, x'_2) = \frac{2\beta}{c}$. Keep going until an endpoint x'_{N-3} is defined such that adding x'_{N-2} as before would leave an interval $(x'_{N-2}, 1)$ with Jensen-Bregman divergence less than $\frac{2\beta}{c}$. Now, instead of defining x'_{N-2} in this way, define it so that $\mathrm{JB}_G(x'_{N-3}, x'_{N-2}) = \mathrm{JB}_G(x'_{N-2}, 1)$. Since $\mathrm{JB}_G(x'_{N-3}, 1) \ge \frac{2\beta}{c}$, the c-approximate triangle inequality that we have by assumption tells us that $\mathrm{JB}_G(x'_{N-3}, x'_{N-2}) = \mathrm{JB}_G(x'_{N-2}, 1) \ge \beta$.

[20] If the infimum is achieved (e.g. if the space of signals to Alice and Bob is finite), then we can set $x_k := \arg\min_x \rho(x)$. Our algorithm works in more generality, at the expense of a factor of 2 in our final bound. Note that by replacing 2 with a smaller constant can arbitrarily reduce this factor.

Proof. We use the following claim, whose proof we provide afterward.

Claim C.5. *Let $I = [x^-, x^+]$ be any sub-interval of $[0,1]$ and let $\alpha = \inf_{x \in I} \rho(x)$. Then there is an increasing sequence of points $z_0 := x^-, z_1, z_2, \ldots, z_{L-1}, z_L := x^+$, such that for every $\ell \in [L]$, $\mathbb{P}[\mu_- \le z_{\ell-1}, \mu_+ \ge z_\ell] \ge \frac{\alpha}{2}$, and where*

$$L \le \frac{2}{\alpha} \sum_{\ell \in [L]} \mathbb{P}[\mu_- \le z_{\ell-1} < \mu_+ \le z_\ell].$$

We apply Claim C.5 to the intervals $I'_1, \ldots, I'_{N-1}$, with $\alpha = \alpha_k$. Let $z_{k,0}, \ldots, z_{k,L_k}$ be the points whose existence the claim proves, and let $r_k := \sum_{\ell \in [L_k]} \mathbb{P}\left[\mu_- \le z_{k,\ell-1} < \mu_+ \le z_{k,\ell}\right]$, so that $L_k \le \frac{2}{\alpha_k} r_k$. Observe that $\sum_k r_k \le 1$, because the intervals $(z_{k,\ell-1}, z_{k,\ell}]$ are disjoint for all k, ℓ. We make the following claim (we provide the proof afterward).

Claim C.6.

$$\sum_{k \in [N-1]} r_k \left(\frac{\alpha_k}{2r_k}\right)^{1-\log_2 c} \le \frac{\epsilon}{\beta c}. \tag{18}$$

We may rewrite Equation 18 as

$$\left(\sum_{k \in [N-1]} r_k \left(\frac{\alpha_k}{2r_k}\right)^{1-\log_2 c}\right)^{1/(1-\log_2 c)} \le \left(\frac{\epsilon}{\beta c}\right)^{1/(1-\log_2 c)}.$$

Recall that $\sum_k r_k \le 1$. Scaling the r_k's to add to 1 decreases the left-hand side above, so we may assume that $\sum_k r_k = 1$. Note that $x^{1-\log_2 c}$ is convex. Thus, by using a weighted Jensen inequality on the left-hand side with weights r_k, we find that

$$\frac{1}{2}\sum_k \alpha_k = \sum_k r_k \cdot \frac{\alpha_k}{2r_k} \le \left(\sum_{k \in [N-1]} r_k \left(\frac{\alpha_k}{2r_k}\right)^{1-\log_2 c}\right)^{1/(1-\log_2 c)} \le \left(\frac{\epsilon}{\beta c}\right)^{1/(1-\log_2 c)}.$$

This completes the proof of Lemma C.4. □

Proof of Claim C.5. Let $z_1 = \inf\{z : \mathbb{P}[\mu_- \le z_0 < \mu_+ \le z] \ge \frac{\alpha}{2}\}$, or x^+ if this number does not exist or is larger than x^+. Note that $\mathbb{P}[\mu_- \le z_0 < \mu_+] \ge \alpha$, as we have $\rho(z_0) = \mathbb{P}[\mu_- \le z_0 < \mu_+] + \mathbb{P}[\mu_- < z_0 = \mu_+] \ge \alpha$, so if the first term were less than α we would have some $z' > z_0$ with $\rho(z') < \alpha$. On the other hand, $\mathbb{P}[\mu_- \le z_0 < \mu_+ < z_1] \le \frac{\alpha}{2}$, since

$$\mathbb{P}[\mu_- \le z_0 < \mu_+ < z_1] = \lim_{z \to z_1 \text{ from below}} \mathbb{P}[\mu_- \le z_0 < \mu_+ \le z]$$

and if the right-hand side were more than $\frac{\alpha}{2}$ then that would contradict the definition of z_1 as an infimum. Therefore, $\mathbb{P}[\mu_- \le z_0, \mu_+ \ge z_1] \ge \frac{\alpha}{2}$.

If $z_1 = x^+$, we are done. Otherwise, let $z_2 = \inf\{z : \mathbb{P}[\mu_- \le z_1 < \mu_+ \le z] \ge \frac{\alpha}{2}\}$. Then $\mathbb{P}[\mu_- \le z_1, \mu_+ \ge z_2] \ge \frac{\alpha}{2}$. Define z_3 analogously, and so forth.

All that remains to show is the upper bound on L. This is where we use the fact that (by construction) $\mathbb{P}[\mu_- \le z_{\ell-1} < \mu_+ \le z_\ell] \ge \frac{\alpha}{2}$. Summing over all ℓ, we have

$$\sum_{\ell \in [L]} \mathbb{P}[\mu_- \le z_{\ell-1} < \mu_+ \le z_\ell] \ge \frac{\alpha}{2} L,$$

which (after rearranging) completes the proof. □

Proof of Claim C.6. First note that by construction, $\mathrm{JB}_G(I'_k) \geq \beta$ for all k. By repeated use of the c-approximate triangle inequality,[21] we find that

$$\sum_{\ell \in [L_k]} \mathrm{JB}_G(z_{k,\ell-1}, z_{k,\ell}) \geq c^{\lceil \log_2 L_k \rceil} \mathrm{JB}_G(I'_k) \geq c^{1+\log_2 \frac{2r_k}{\alpha_k}} \mathrm{JB}_G(I'_k) \geq c^{1+\log_2 \frac{2r_k}{\alpha_k}} \beta = c \left(\frac{2r_k}{\alpha_k}\right)^{\log_2 c} \beta.$$

On the other hand, we have

$$\epsilon \geq \mathbb{E}\left[\mathrm{JB}_G(\mu_\sigma, \mu_\tau)\right] = \sum_{\sigma,\tau} \mathbb{P}\left[\sigma, \tau\right] \mathrm{JB}_G(\mu_\sigma, \mu_\tau) \geq \sum_{\sigma,\tau} \mathbb{P}\left[\sigma, \tau\right] \sum_{\substack{k,\ell: \mu_- \leq z_{k,\ell-1} \\ \mu_+ \geq z_{k,\ell}}} \mathrm{JB}_G(z_{k,\ell-1}, z_{k,\ell})$$

$$= \sum_{k,\ell} \mathbb{P}\left[\mu_- \leq z_{k,\ell-1}, \mu_+ \geq z_{k,\ell}\right] \mathrm{JB}_G(z_{k,\ell-1}, z_{k,\ell}) \geq \sum_{k,\ell} \frac{\alpha_k}{2} \mathrm{JB}_G(z_{k,\ell-1}, z_{k,\ell}).$$

Here, the third step follows by the reverse triangle inequality (Fact (ii) of Proposition C.1) and the fourth step follows by rearranging the order of summation.[22] Combining the last two facts gives us that

$$\epsilon \geq \sum_k \frac{\alpha_k}{2} \cdot c \left(\frac{2r_k}{\alpha_k}\right)^{\log_2 c} \beta = \sum_k r_k \left(\frac{2r_k}{\alpha_k}\right)^{\log_2 c - 1} \beta c,$$

which rearranges to the desired identity. □

We are now ready to bound the second summand, i.e. $\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right]$, where $k(\sigma)$ is the k such that Alice's estimate μ_σ lies in I_k. For convenience we will define $k(\tau)$ for Bob by analogy as the k such that μ_τ lies in I_k. By Lemma C.4 and the preceding discussion, we know that

$$\mathbb{P}\left[k(\sigma) \neq k(\tau)\right] \leq 4 \left(\frac{\epsilon}{\beta c}\right)^{1/(1-\log_2 c)}.$$

Lemma C.7. *Let $Q = \mathbb{P}\left[k(\sigma) \neq k(\tau)\right]$. Then*

$$\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right] \leq 2\tilde{G}^*(Q).$$

The key idea is that because $k(\sigma) = k(\tau)$ with probability near 1, learning $k(\sigma)$ is unlikely to make Bob update his estimate much.

Proof. Consider any signal $\hat{\tau} \in \mathcal{T}$ and let $p(\hat{\tau}) = \mathbb{P}\left[\tau = \hat{\tau}\right]$. We have[23]

$$\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right] = \sum_{\hat{\tau} \in \mathcal{T}} p(\hat{\tau}) \mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\hat{\tau}} \parallel \mu_{\hat{\tau}}) \mid \tau = \hat{\tau}\right].$$

Note that $\mu_{\hat{\tau}} = \mathbb{E}\left[\mu_{S^{(k(\sigma))}\hat{\tau}} \mid \tau = \hat{\tau}\right]$, so by Proposition C.1 we have that

$$\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right] = \sum_{\hat{\tau} \in \mathcal{T}} p(\hat{\tau}) \left(\mathbb{E}\left[G(\mu_{S^{(k(\sigma))}\hat{\tau}}) \mid \tau = \hat{\tau}\right] - G(\mu_{\hat{\tau}})\right).$$

Let $q(\hat{\tau}) = \mathbb{P}\left[\tau = \hat{\tau}, k(\sigma) \neq k(\hat{\tau})\right]$, so $\sum_{\hat{\tau} \in \mathcal{T}} q(\hat{\tau}) = Q$. Then

$$\mathbb{E}\left[G(\mu_{S^{(k(\sigma))}\hat{\tau}}) \mid \tau = \hat{\tau}\right] - G(\mu_{\hat{\tau}}) =$$

$$\frac{p(\hat{\tau}) - q(\hat{\tau})}{p(\hat{\tau})} \left(\mathbb{E}\left[G(\mu_{S^{(k(\hat{\tau}))}\hat{\tau}}) - G(\mu_{\hat{\tau}})\right]\right) + \frac{q(\hat{\tau})}{p(\hat{\tau})} \left(\mathbb{E}\left[G(\mu_{S^{(k(\sigma))}\hat{\tau}}) \mid \tau = \hat{\tau}, k(\sigma) \neq k(\hat{\tau})\right] - G(\mu_{\hat{\tau}})\right).$$

[21] We sub-divide I'_k into $[z_{k,0}, z_{k,L_k/2}]$ and $[z_{k,L_k/2}, z_{k,L}]$, then subdivide each of these, and so on.

[22] The case that the space of signals is infinite is identical except that the summation is replaced by an integral over the probability space.

[23] This proof takes sums over $\hat{\tau} \in \mathcal{T}$ and thus implicitly assumes that $\mathcal{T}$ is finite, but the proof extends to infinite $\mathcal{T}$, with sums over τ replaced by integrals with respect to the probability measure over $\mathcal{T}$.

The second term is at most $\frac{q(\hat{\tau})}{p(\hat{\tau})}M$, since M is the range of G. To bound the first term, we note that $\mu_{S^{(k(\hat{\tau}))}\hat{\tau}}$ cannot differ from $\mu_{\hat{\tau}}$ by more than $\frac{q(\hat{\tau})}{p(\hat{\tau})-q(\hat{\tau})}$, as otherwise the average value of $\mu_{S^{(k(\sigma))}\hat{\tau}}$ could not be $\mu_{\hat{\tau}}$. Therefore, $\mathbb{E}\left[G(\mu_{S^{(k(\hat{\tau}))}\hat{\tau}}) - G(\mu_{\hat{\tau}})\right]$ is bounded by the largest possible difference in G-values of two points that differ by at most $\frac{q(\hat{\tau})}{p(\hat{\tau})-q(\hat{\tau})}$. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right] &\le \sum_{\hat{\tau}\in\mathcal{T}} p(\hat{\tau})\left(\frac{p(\hat{\tau})-q(\hat{\tau})}{p(\hat{\tau})}\tilde{G}\left(\frac{q(\hat{\tau})}{p(\hat{\tau})-q(\hat{\tau})}\right) + \frac{q(\hat{\tau})}{p(\hat{\tau})}M\right)\\ &\le QM + \sum_{\hat{\tau}\in\mathcal{T}}(p(\hat{\tau})-q(\hat{\tau}))\tilde{G}\left(\frac{q(\hat{\tau})}{p(\hat{\tau})-q(\hat{\tau})}\right),\end{aligned}$$

where $\tilde{G}$ is defined as in the statement of Lemma C.7. If G is symmetric on $[0,1]$, then $\tilde{G}(x) = G(0) - G(x)$ for $x \le \frac{1}{2}$ and M otherwise. This is a concave function, but $\tilde{G}$ is not in general concave. However, consider $\tilde{G}^*$ as defined in the lemma statement, so $\tilde{G}(x) \le \tilde{G}^*(x)$ for all x. Then

$$\begin{aligned}\mathbb{E}\left[D_G(\mu_{S^{(k(\sigma))}\tau} \parallel \mu_\tau)\right] &\le QM + \sum_{\hat{\tau}\in\mathcal{T}}(p(\hat{\tau})-q(\hat{\tau}))\tilde{G}^*\left(\frac{q(\hat{\tau})}{p(\hat{\tau})-q(\hat{\tau})}\right)\\ &\le QM + \left(\sum_{\hat{\tau}\in\mathcal{T}}(p(\hat{\tau})-q(\hat{\tau}))\right)\cdot\tilde{G}^*\left(\frac{\sum_{\hat{\tau}\in\mathcal{T}}q(\hat{\tau})}{\sum_{\hat{\tau}\in\mathcal{T}}(p(\hat{\tau})-q(\hat{\tau}))}\right)\\ &= QM + (1-Q)\tilde{G}^*\left(\frac{Q}{1-Q}\right) \le QM + \tilde{G}^*(Q) \le 2\tilde{G}^*(Q).\end{aligned}$$

Here, the second step follows by Jensen's inequality with terms $\frac{q(\hat{\tau})}{p(\hat{\tau})-q(\hat{\tau})}$ and weights $p(\hat{\tau}) - q(\hat{\tau})$, the second-to-last step follows from the fact that $\tilde{G}^*$ is convex and $\tilde{G}^*(0) = 0$, and the last step follows from the fact that $\tilde{G}^*$ is convex and $\tilde{G}^*(1) = M$. □

Since $Q \le 4\left(\frac{\epsilon}{\beta c}\right)^{1/(1-\log_2 c)}$, combining Lemma C.7 with Equation 17 gives us the following result.

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_\tau)\right] \le \frac{8\beta}{c^2} + 2\tilde{G}^*\left(4\left(\frac{\epsilon}{\beta c}\right)^{1/(1-\log_2 c)}\right).$$

Noting that $\tilde{G}^*$ is concave and $c^{-1/(1-\log_2 c)} \le 2$ (which is true for all $0 < c < 1$) completes the proof of Lemma 4.19. □

D Alternative Definitions of Agreement and Accuracy

For arbitrary Bregman divergences, there are several notions of agreement and accuracy that are worth considering. Before we discuss these, we make a note about the order of arguments in a Bregman divergence. In our context, it makes the most sense to talk of the Bregman divergence *from a more informed estimate to a less informed estimate*. By a "more informed estimate" we mean a finer-grained one, i.e. one that is informed by more knowledge. For example, in terms of estimating Y in the context of this work explores, Y is more informed than $\mu_{\sigma\tau}$, which is more informed than μ_σ and $\mu_{S\tau}$, which are each more informed than μ_{ST}, which is more informed than $\mathbb{E}[Y]$.

To see that this is the natural order of the arguments, recall that Bregman divergences are motivated by the property that they elicit the mean (see Proposition 4.2): if an agent who gives an estimate of x for the value of a random variable Y incurs a loss of $D_G(Y \parallel x)$, then the agent minimizes their expected loss by reporting $x = \mathbb{E}[Y]$. This means that the expert ought to report the expected value of Y given the information that the expert knows.

This means that given two estimates of Y, Z_1 and Z_2, of which Z_1 is more informed, the quantity $D_G(Z_1 \parallel Z_2)$ has a natural interpretation: it is the expected amount the expert gains by learning more and refining their estimate from Z_2 to Z_1. This follows by the Pythagorean theorem:

$$\mathbb{E}\left[D_G(Z_1 \parallel Z_2)\right] = \mathbb{E}\left[D_G(Y \parallel Z_2)\right] - \mathbb{E}\left[D_G(Y \parallel Z_1)\right].$$

D.1 Alternative Definitions of Agreement

One important motivation for using the Jensen-Bregman divergence to the midpoint as the definition of agreement is that this quantity serves as a lower bound on the expected amount that Charlie disagrees with Alice and Bob. Formally:

Definition D.1. *Let a, b, and c be Alice's, Bob's, and Charlie's expectations, respectively (these are random variables on Ω). Alice and Bob* ϵ-agree with Charlie *if* $\frac{1}{2}(\mathbb{E}\left[D_G(a \parallel c) + D_G(b \parallel c)\right]) \le \epsilon$.

(This is the order of arguments because Alice and Bob are more informed than Charlie.) By Proposition C.1 (i), we know that **if Alice and Bob ϵ-agree with Charlie then they ϵ-agree**.

As it happens the fact that under this (stronger) definition of agreement implies accuracy under rectangle substitutes follows immediately:

Proposition D.2. *Let $\mathcal{I} = (\Omega, \mathbb{P}, \mathcal{S}, \mathcal{T}, Y)$ be an information structure that satisfies rectangle substitutes. For any communication protocol that causes Alice and Bob to ϵ-agree with Charlie on $\mathcal{I}$, Alice and Bob are 2ϵ-accurate after the protocol terminates.*

Proof. Let S be the set of possible signals of Alice at the end of the protocol which are consistent with the protocol transcript, and define T likewise for Bob. Recall that Charlie's expectation is μ_{ST}. We have

$$\mathbb{E}\left[D_G(\mu_{\sigma\tau} \parallel \mu_{S\tau})\right] \le \mathbb{E}\left[D_G(\mu_{\sigma T} \parallel \mu_{ST})\right] \le \mathbb{E}\left[D_G(\mu_{\sigma T} \parallel \mu_{ST})\right] + \mathbb{E}\left[D_G(\mu_{S\tau} \parallel \mu_{ST})\right] \le 2\epsilon,$$

where the first inequality follows by rectangle substitutes and the last inequality follows because Alice and Bob ϵ-agree with Charlie. □

The drawback of Definition D.1 is that it is not so much a definition of Alice and Bob's agreement with each other, so much as a definition of agreement with respect to the protocol being run (since Charlie only exists within the context of the protocol). Put otherwise, it is impossible to determine whether Alice and Bob ϵ-agree with Charlie simply by knowing Alice and Bob's expectations; one must also know Charlie's expectation, which cannot be determined from Alice's and Bob's expectations. The question "how far from agreement are Alice and Bob if Alice believes 25% and Bob believes 30%?" makes sense in the context of ϵ-agreement, but not in the context of ϵ-agreement with Charlie.

A different notion of agreement, which (like ϵ-agreement) only depends on Alice's and Bob's expectations, uses the *symmetrized Bregman divergence* between these expectations: $\frac{1}{2}(D_G(a \parallel b) + D_G(b \parallel a))$.

Definition D.3. *Let a and b be Alice's and Bob's expectations, respectively (these are random variables on Ω). Alice and Bob satisfy* symmetrized ϵ-agreement *if* $\frac{1}{2}(D_G(a \parallel b) + D_G(b \parallel a))$.

By Proposition C.1 (iii), we know that **if Alice and Bob satisfy symmetrized ϵ-agreement then they ϵ-agree**.

In our context, symmetrized Bregman divergence is less natural than Jensen-Bregman divergence. This is symmetrized Bregman divergence (unlike Jensen-Bregman divergence) does not seem to closely relate to our previous discussion of the Bregman divergence from a more informed to a less informed estimate being most natural.

D.2 Alternative Notions of Accuracy

Our definition of Alice's accuracy as the expected Bregman divergence from the truth $\mu_{\sigma\tau}$ to Alice's expectation seems like the most natural one. However, one may desire a definition of accuracy that takes both Alice's and Bob's expectations into account, judging the pair's accuracy based on their consensus belief, rather than each of their individual beliefs. For instance, one could say that Alice and Bob are ϵ-*midpoint-accurate* if $\mathbb{E}\left[D_G\left(\mu_{\sigma\tau} \parallel \frac{a+b}{2}\right)\right] \leq \epsilon$. By this definition, Alice's and Bob's expectations could individually be far from the truth, but they are considered accurate because the average of their expectations is close to correct.

Proposition D.4. *If Alice and Bob are ϵ-accurate, then they are 2ϵ-midpoint-accurate.*

Proof. Observe that for all a, b, y it is the case that

$$D_G\left(y \parallel \frac{a+b}{2}\right) \leq \max(D_G(y \parallel a), D_G(y \parallel b) \leq D_G(y \parallel a) + D_G(y \parallel b).$$

The first inequality is true simply because $\frac{a+b}{2}$ lies in between a and b. Therefore,

$$\mathbb{E}\left[D_G\left(y \parallel \frac{a+b}{2}\right)\right] \leq \mathbb{E}\left[D_G(y \parallel a) + D_G(y \parallel b)\right] \leq 2\epsilon.$$

□

Another natural choice for Alice's and Bob's consensus belief is the QA pool (see [Neyman and Roughgarden, 2021b]). Proposition D.4 likewise holds for the QA pool in place of the midpoint, and indeed holds for any choice of consensus belief that is guaranteed to lie in between Alice's and Bob's expectations. Thus, any such definition will be weaker than our definition of ϵ-accuracy for Alice and Bob (up to a constant factor).

To summarize, among the above definitions of agreement, ϵ-agreement is the weakest; and among the above definitions of accuracy, Alice's and Bob's ϵ-accuracy is the strongest. This is an indication of strength for Theorem 4.18: it starts from a relatively weak premise and reaches a relatively strong conclusion.

E Implications for Communication Complexity

Our results can be framed in a communication complexity context, where they imply that "substitutable" functions can be computed with probability $1 - \delta$ (over the inputs) with a transcript length depending only on δ. This is a nonstandard and weak notion of computing the function, but sketching the reduction may inspire future work on connections between substitutes and communication complexity.

In a classic deterministic communication complexity setup (e.g. [Rao and Yehudayoff, 2020]), Alice holds $\sigma \in \mathcal{S}$, Bob holds $\tau \in \mathcal{T}$, and the goal is to compute some function $g : \mathcal{S} \times \mathcal{T} \to \{0, 1\}$ using a communication protocol (see Section 2.2). Our setting captures this model when $Y = g(\sigma, \tau)$. Observe that in this case, $Y = \mu_{\sigma\tau}$, i.e. Alice and Bob's information together determine Y completely. A communication protocol defines its output by a function $h : \Pi \to \{0, 1\}$ where Π is the space of transcripts. We can simply let $h(\pi) = \text{round}(\mu_{ST})$, i.e. rounding the *ex post* expectation $\mathbb{E}[Y \mid \pi] = \mu_{ST}$ to either zero or one. This is equivalent to the belief of "Charlie", or the common knowledge of Alice and Bob after the protocol is completed.

Definition E.1 (Rectangle substitutes, $(1-\delta)$-computes). *Given a function g and a distribution $\mathcal{D}$ over $\mathcal{S} \times \mathcal{T}$, we say $(g, \mathcal{D})$ satisfy* rectangle substitutes *if the corresponding information*

structure with $Y = g(\sigma, \tau)$ satisfies rectangle substitutes (Definition 2.6). We say a protocol $(1-\delta)$-computes g over $\mathcal{D}$ if, with probability at least $1-\delta$ over $(\sigma, \tau) \sim \mathcal{D}$, the protocol has $h(\pi) = g(\sigma, \tau)$.

By our results, under rectangle substitutes $(g, \mathcal{D})$, any agreement protocol approximately computes g over $\mathcal{D}$. More precisely, using a fast substitutes-agreement protocol similar to Proposition B.1, we obtain the following.

Corollary E.2. *Suppose $(g, \mathcal{D})$ satisfy rectangle substitutes. Then for every $\delta \in (0, 1)$, there is a deterministic communication protocol using $O(\log(1/\delta))$ bits of communication that $(1-\delta)$-computes g over $\mathcal{D}$.*

Proof. In round one, Alice sends her current expectation μ_σ rounded to a multiple of ϵ; call this message A. In round two, Bob sends his updated expectation $\mu_{S\tau}$ rounded to a multiple of ϵ; call this message B. The protocol then halts, and the output is B rounded to either zero or one. It uses $O(\log(1/\epsilon))$ bits. Let S, T be the random rectangle associated with the protocol.

By construction, $|\mu_\sigma - A| \le \epsilon$, and μ_S is the expectation of Y conditioned on A, so it follows that $|\mu_\sigma - \mu_S| \le \epsilon$. Using substitutes (just as in Proposition B.1),

$$\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2\right] \le \mathbb{E}\left[(\mu_\sigma - \mu_S)^2\right] \le \epsilon^2.$$

By construction, $|B - \mu_{S\tau}| \le \epsilon$. Therefore, by the $\frac{1}{2}$-approximate triangle inequality for squared distance (e.g. Proposition 4.13)),

$$\mathbb{E}\left[(\mu_{\sigma\tau} - B)^2\right] \;\le\; 2\mathbb{E}\left[(\mu_{\sigma\tau} - \mu_{S\tau})^2\right] + 2\mathbb{E}\left[(\mu_{S\tau} - B)^2\right] \;\le\; 2\epsilon^2.$$

Now, the protocol is incorrect if $|B - \mu_{\sigma\tau}| \ge \frac{1}{2}$. Using Markov's inequality,

$$\begin{aligned}\Pr[|B - \mu_{\sigma\tau}| \ge \tfrac{1}{2}] &= \Pr[(B - \mu_{\sigma\tau})^2 \ge \tfrac{1}{4}] \\ &\le 4\mathbb{E}\left[(B - \mu_{\sigma\tau})^2\right] \\ &\le 8\epsilon^2.\end{aligned}$$

Therefore, given $\delta \in (0, 1)$, we run the protocol with $\epsilon = \sqrt{\delta/8}$. The probability of an incorrect output is at most δ, and we use $O(\log(1/\epsilon) = O(\log(1/\delta))$ bits of communication. □