

Quantifying information and uncertainty

Alexander Frankel and Emir Kamenica
University of Chicago

May 2019

Abstract

We examine ways to measure the amount of information generated by a piece of news and the amount of uncertainty implicit in a given belief. Say a measure of information is valid if it corresponds to the value of news in some decision problem. Say a measure of uncertainty is valid if it corresponds to expected utility loss from not knowing the state in some decision problem. We axiomatically characterize all valid measures of information and uncertainty. We show that if measures of information and uncertainty arise from the same decision problem, then they are coupled in that the expected reduction in uncertainty always equals the expected amount of information generated. We provide explicit formulas for the measure of information that is coupled with any given measure of uncertainty and vice versa. Finally, we show that valid measures of information are the only payment schemes that never provide incentives to delay information revelation.

JEL classification: D80, D83

Keywords: value of information, Bregman divergences, entropy

1 Introduction

Suppose we observe some pieces of news. How might we quantify the amount of information contained in each piece? One desideratum might be that the measure should correspond to the instrumental value of information for *some* decision problem. We call such measures of information valid. Another approach would be to specify that the measure should satisfy the following properties: (i) news cannot contain a negative amount of information, (ii) news that does not affect beliefs generates no information, and (iii) the order in which the news is read does not, on average, change the total amount of information generated. The first result of this paper is that these two approaches are equivalent. A measure of information is valid if and only if it satisfies the three aforementioned properties.

A related question is: how might we quantify the uncertainty of a belief? Again, one approach would be to measure uncertainty by its instrumental cost, i.e., by the extent to which it reduces a decision-maker's utility relative to an omniscient benchmark. We call such measures of uncertainty valid; every concave function that is zero at degenerate beliefs is valid. Another approach would couple the measure of uncertainty to some measure of information and insist that, on average, observing news reduces uncertainty to the extent that it generates information. We show that these two approaches are equivalent: a measure of uncertainty reflects the instrumental cost for some decision-maker if and only if the expected reduction in uncertainty always equals the expected amount of information generated (by a measure based on that decision-maker's instrumental value of information).

These results have various implications. First, they tell us that some seemingly sensible ways of measuring information, such as the Euclidean distance between the prior and the posterior, are not valid (under our definition of this term). In fact, no metric is valid: there does not exist a decision problem whose instrumental value of information is a metric on beliefs. Second, our results provide decision-theoretic foundations for standard measures of information and uncertainty, such as Kullback-Leibler divergence and entropy or quadratic distance and variance. Third, our results introduce a notion of “coupling” between measures of information and measures of uncertainty, which reflects the fact that Kullback-Leibler divergence complements entropy while quadratic distance complements variance. In fact, every valid measure of information is coupled with a unique valid

measure of uncertainty (and vice versa), but we cannot mix and match. We also derive a functional form that pins down the measure of uncertainty coupled with a given measure of information and vice versa. These functional forms in turn provide an easy way of verifying whether a given measure of information is valid. Finally, we establish that the measures of information we term valid coincide with the set of incentive-compatible payment schemes in a class of dynamic information-elicitation environments.

In contrast to much of the existing literature, we focus on ex post rather than ex ante measures of information.¹ Taking a decision problem as given, it is straightforward to quantify the ex ante instrumental value of a Blackwell experiment, i.e., of a news-generating process.² The ex post perspective, however, is the more relevant one in a number of economic settings. First, measuring the amount (or the value) of news after it has been observed plays a key role in decisions ranging from a media outlet selecting its headline story to a journal editor determining whether to publish a paper (Frankel and Kasy 2018). Second, as emphasized in the recent survey by Bergemann and Bonatti (2018), many markets for information are structured so that the payment for information is based on the signal realization rather than the ex ante signal (see Babaioff et al. 2012 and Bergemann and Bonatti 2015). Understanding consumers' willingness to pay in such markets necessarily entails the ex post perspective. Finally, if one is interested in the uncertain ex ante value of some information source, one approach to estimating it would be to average the ex post value the source generates. Our results provide a full characterization of all unbiased estimators of ex ante value of information. Indeed, if an estimator satisfies certain desiderata (namely that it is unbiased and that its range lies in the constrained parameter space), it must be the sample average of a valid (ex post) measure of information.

It is important to note that our approach is focused on the benefits of information and the benefits of the reduction in uncertainty. A complementary literature analyzes functional forms over information and uncertainty that arise from costs of acquisition or processing of information.³

¹In footnote 14, we extend our notion of validity to ex ante measures of information and provide a simple characterization of such measures.

²Blackwell (1951) provides an ordinal comparison of experiments without a reference to a specific prior or decision problem. Lehmann (1988), Persico (2000), and Cabrales et al. (2013) consider ordinal comparisons on a restricted space of priors, problems, and/or experiments. One could also consider ordinal rankings of pieces of news from an ex post perspective, but we do not take that route in this paper. De Lara and Gossner (2017) express the cardinal value of an experiment based on its influence on decisions.

³Arrow (1971) discusses the distinction between microfounding entropy via the benefit vs. the cost of information

The literature on rational inattention (e.g., Sims 2003) associates such costs with the expected reduction in entropy.⁴ Hebert and Woodford (2017), Morris and Strack (2017), Mensch (2018b), and Pomatto et al. (2019) provide microfoundations for various measures of uncertainty that reflect costs of information acquisition. Of course, there is no reason to expect that these measures should satisfy our notion of validity, since they reflect costs rather than benefits of information acquisition.

2 Set-up

2.1 The informational environment

There is a finite state space $\Omega = \{1, \dots, n\}$ with a typical state denoted ω . A *belief* q is a distribution on Ω that puts weight q^ω on the state ω . We denote a belief that is degenerate on ω by δ_ω .

Information is generated by signals. We follow the formalization of Green and Stokey (1978) and Gentzkow and Kamenica (2017) and define a signal π as a finite partition of $\Omega \times [0, 1]$ s.t. $\pi \subset S$, where S is the set of non-empty Lebesgue-measurable subsets of $\Omega \times [0, 1]$. We refer to an element $s \in S$ as a *signal realization*. The interpretation is that a random variable x drawn uniformly from $[0, 1]$ determines the signal realization conditional on the state; the probability of observing $s \in \pi$ in ω is the Lebesgue measure of $\{\omega, x \mid (\omega, x) \in s\}$. As a notational convention, we let α denote the S -valued random variable induced by signal π_α .

Given a prior q , we denote the posterior induced by signal realization s by $q(s)$.⁵ Observing realizations from both π_α and π_β induces the posterior $q(\alpha \cap \beta)$ since it reveals that $(\omega, x) \in \alpha \cap \beta$.

For every signal π_α , we have $\mathbb{E}[q(\alpha)] = q$.⁶

acquisition.

⁴These cost functions have been generalized along a number of dimensions (Caplin and Dean 2013, Gentzkow and Kamenica 2014, Yoder 2016, Steiner et al. 2017, Mensch 2018a). These generalizations typically still maintain that the cost of an experiment is proportional to the expected reduction of some measure of uncertainty. Caplin et al. (2017) refer to cost functions in this class as posterior separable.

⁵Note that, as long as the probability of s is strictly positive given q , Bayes' rule implies a unique $q(s)$ which does not depend on the signal π from which s was realized. If probability of s is zero, we can set $q(s)$ to an arbitrary belief.

⁶In the expression $\mathbb{E}[q(\alpha)]$, the expectation is taken over the realization of α , whose distribution depends on q .

2.2 Decision problems

A decision problem $\mathcal{D} = (A, u)$ specifies an action set A and a utility function $u : A \times \Omega \rightarrow \mathbb{R} \cup \{-\infty\}$.⁷ Given the extended range, we assume that there exists some action a such that $u(a, \omega)$ is finite for every ω ; moreover, we use the convention that $-\infty \times 0 = 0$ so that taking an action that yields $-\infty$ in some state is not costly if that state has zero probability. Finally, we assume the decision problem is well-defined in the sense that $\arg \max_{a \in A} \mathbb{E}_q [u(a, \omega)]$ is non-empty for all beliefs q .

Given a decision problem $\mathcal{D} = (A, u)$, a *value of information* for \mathcal{D} , denoted $v_{\mathcal{D}} : \Delta(\Omega) \times \Delta(\Omega) \rightarrow \mathbb{R}$, is given by⁸

$$v_{\mathcal{D}}(p, q) = \mathbb{E}_p [u(a^*(p), \omega)] - \mathbb{E}_p [u(a^*(q), \omega)],$$

where for belief q , $a^*(q) \in \arg \max_{a \in A} \mathbb{E}_q [u(a, \omega)]$.⁹ A given decision problem does not necessarily imply a unique $v_{\mathcal{D}}(p, q)$ when the decision-maker is indifferent across multiple actions at belief q , but any two functions that are a value of information for the same decision problem will coincide for almost every q . From the perspective of an agent with belief p , the payoff to $a^*(q)$ is $\mathbb{E}_p [u(a^*(q), \omega)]$, whereas the payoff from taking the “correct” action given this belief is $\max_{a \in A} \mathbb{E}_p [u(a, \omega)]$. Thus, $v_{\mathcal{D}}(p, q)$ reflects the ex post value of a piece of information that updates beliefs from q to p for a decision-maker who faces the decision problem \mathcal{D} . It is important to note that $v_{\mathcal{D}}$ captures the value of a piece of information in isolation. If other information is available, valuing one piece of information could entail considerations of the extent to which it substitutes or complements the other pieces (Börgers et al. 2013).

If u is denominated in money, we can think of $v_{\mathcal{D}}(p, q)$ as the greatest price at which the decision-maker could have purchased a signal that moved her belief from q to p such that she does not regret the purchase. Another interpretation of $v_{\mathcal{D}}(p, q)$ is the instrumental loss from believing q when available data indicates p ; Gossner and Steiner (2018) introduce this function to study costs

⁷We explain below the benefit of including $-\infty$ in the range of the utility function.

⁸We restrict the domain of v to be pairs (p, q) such that the support of p is a subset of the support of q , i.e., $q^\omega = 0 \Rightarrow p^\omega = 0$. We need not concern ourselves with movements of beliefs outside of this domain since they are not compatible with Bayesian updating.

⁹We also allow for $a^*(q)$ to be a distribution over optimal actions in which case $u(a^*(q), \omega)$ is interpreted to mean the expectation of $u(a, \omega)$ given that distribution.

of misperception. The same function also arises in models of belief-based utility that emphasize potential optimality of inaccurate beliefs (e.g., Caplin and Leahy 2001, Brunnermeier and Parker 2005).

The *cost of uncertainty* for $\mathcal{D} = (A, u)$ is

$$C_{\mathcal{D}}(q) = \mathbb{E}_q \left[\max_a u(a, \omega) \right] - \max_a \mathbb{E}_q [u(a, \omega)].$$

The term $\mathbb{E}_q [\max_a u(a, \omega)]$ is the expected payoff to the decision-maker if she were to learn the true state of the world before taking the action. The term $\max_a [\mathbb{E}_q [u(a, \omega)]]$ is the expected payoff from the action that is optimal given belief q . Thus, the cost of uncertainty simply reflects how much lower the decision-maker's payoff is because she does not know the state of the world. This function is sometimes also called the expected value of perfect information.

Example 1. (*Simple decision problem*) Consider a decision problem with two actions $a \in \{0, 1\}$, each of which is optimal in the corresponding state of the world $\omega \in \{0, 1\}$. We call such a problem *simple*. Normalizing the payoff of $a = 0$ to zero in both states, $a = 1$ is an optimal action if $qu(1, 1) + (1 - q)u(1, 0) \geq 0$, i.e., $q \geq r \equiv \frac{-u(1, 0)}{u(1, 1) - u(1, 0)}$, where q denotes the probability of $\omega = 1$. We can further normalize the denominator $u(1, 1) - u(1, 0)$ to 1, which yields utility function u^r given by $u^r(1, 1) = 1 - r$ and $u^r(1, 0) = -r$ (with $u^r(0, 0) = u^r(0, 1) = 0$). Thus, every simple decision problem is characterized by some $r \in [0, 1]$ and is denoted \mathcal{D}^r .

Any value of information for \mathcal{D}^r equals $|p - r|$ if $r \in (\min\{p, q\}, \max\{p, q\})$ – i.e., r is strictly between p and q – and zero if $r > \max\{p, q\}$ or $r < \min\{p, q\}$. To see this, note that if r is not between p and q , the optimal action does not change so the value of information is zero. If $q < r < p$, the optimal action switches from 0 to 1, and the value of information is $pu^r(1, 1) + (1 - p)u^r(1, 0) = p(1 - r) + (1 - p) - r = p - r$. Likewise, if $p < r < q$, the optimal action switches from 1 to 0, and the value of information is $(r - p)$. Thus, when r is between p and q , value of information is proportional to $|p - r|$. When $q = r$, we have flexibility in specifying the value of information since it depends on the way the decision-maker breaks her indifference. For concreteness, we suppose the decision-maker takes the two actions with equal probability, which yields the value of $\frac{|p - r|}{2}$. Hence,

the value of information for \mathcal{D}^r is

$$v^r(p, q) = \begin{cases} |p - r| & \text{if } r \in (\min\{p, q\}, \max\{p, q\}) \\ \frac{|p-r|}{2} & \text{if } r = q \\ 0 & \text{otherwise} \end{cases}.$$

Note that, under v^r , belief movements that are “similar” do not necessarily generate similar value: if the cutoff belief is $r = 0.5$, moving from 0.49 to 0.9 generates much more value than moving from 0.51 to 0.9 since the former changes the action to the ex post optimal one, whereas the latter leaves the action unchanged.

The cost of uncertainty for problem \mathcal{D}^r is the triangular function

$$C^r(q) = \begin{cases} q(1-r) & \text{if } q \leq r \\ (1-q)r & \text{if } q > r \end{cases}.$$

Example 2. (*Quadratic loss estimation*) Suppose $\Omega \subset \mathbb{R}$, $A = co(\Omega)$, and $u^Q(a, \omega) = -(a - \omega)^2$, where co denotes the convex hull. The optimal action given belief q is $\mathbb{E}_q[\omega]$, the value of information is¹⁰

$$v^Q(p, q) = (\mathbb{E}_p[\omega] - \mathbb{E}_q[\omega])^2,$$

10

$$\begin{aligned} v^Q(p, q) &= \mathbb{E}_p[-(\mathbb{E}_p[\omega] - \omega)^2 + (\mathbb{E}_q[\omega] - \omega)^2] \\ &= \mathbb{E}_p[-((\mathbb{E}_p[\omega])^2 - 2\mathbb{E}_p[\omega]\omega + \omega^2) + ((\mathbb{E}_q[\omega])^2 - 2\mathbb{E}_q[\omega]\omega + \omega^2)] \\ &= \mathbb{E}_p[-(\mathbb{E}_p[\omega])^2 + 2\mathbb{E}_p[\omega]\omega + (\mathbb{E}_q[\omega])^2 - 2\mathbb{E}_q[\omega]\omega] \\ &= (-(\mathbb{E}_p[\omega])^2 + 2(\mathbb{E}_p[\omega])^2 + (\mathbb{E}_q[\omega])^2 - 2\mathbb{E}_q[\omega]\mathbb{E}_p[\omega]) \\ &= ((\mathbb{E}_p[\omega])^2 + (\mathbb{E}_q[\omega])^2 - 2\mathbb{E}_q[\omega]\mathbb{E}_p[\omega]) \\ &= (\mathbb{E}_p[\omega] - \mathbb{E}_q[\omega])^2. \end{aligned}$$

and the cost of uncertainty is¹¹

$$C^Q(q) = \text{Var}_q[\omega].$$

Example 3. (*Brier elicitation*) Consider some arbitrary (finite) Ω , let $A = \Delta(\Omega)$, and suppose $u^B(a, \omega) = -\|a - \delta_\omega\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. In other words, the action is a report of a probability distribution, and the utility is the Brier score of the accuracy of the report (Brier 1950). The optimal action given belief q is to set $a = q$, the value of information is

$$v^B(p, q) = \|p - q\|^2,$$

and the cost of uncertainty is

$$C^B(q) = \sum_{\omega} q^\omega (1 - q^\omega).$$

Ely et al. (2015) refer to C^B as residual variance.

Note that quadratic loss estimation coincides with Brier elicitation (modulo a scaling factor) when the state space is binary.

2.3 Measures of information and uncertainty

A *measure of information* d is a function that maps a pair of beliefs to a real number.¹² We interpret $d(p, q)$ as the amount of information in a piece of news that moves a Bayesian's belief from prior q to posterior p .¹³ A *measure of uncertainty* H is a function that maps a belief to a real number. We interpret $H(q)$ as the amount of uncertainty faced by a decision-maker with belief q .

¹¹

$$\begin{aligned} C^Q(q) &= \mathbb{E}_q \left[\max_a [-(a - \omega)^2] \right] - \max_a \mathbb{E}_q [-(a - \omega)^2] \\ &= 0 - \max_a \mathbb{E}_q [-(a - \omega)^2] \\ &= -\mathbb{E}_q [-(\mathbb{E}_q[\omega] - \omega)^2] \\ &= \text{Var}_q[\omega]. \end{aligned}$$

¹²As with value of information, we restrict the domain of d to be pairs (p, q) such that the support of p is a subset of the support of q .

¹³An alternative approach would be to take news as a primitive, define a piece of news n based on its likelihood in each state $n(\omega)$, and introduce a measure of newswiness $\eta(n, q)$ given a prior q . As long as two pieces of news n and n' with the same likelihood ratios yield $\eta(n, q) = \eta(n', q)$, a measure of newswiness delivers a measure of information by setting $d(p, q) = \eta(n, q)$ for $p^\omega = \frac{n(\omega)q^\omega}{\sum_{\omega'} n(\omega')q^{\omega'}}$.

One natural measure of information is the Euclidean distance between the beliefs, i.e., $d(p, q) = \|p - q\|$. Another candidate measure would be distance-squared, $\|p - q\|^2$. Measures of information can also depend only on some “summary statistic”, e.g., if $\Omega \subset \mathbb{R}$, we might set $d(p, q) = (\mathbb{E}_p[\omega] - \mathbb{E}_q[\omega])^2$. Finally, note that a measure of information need not be symmetric, e.g., it might be Kullback-Leibler divergence, $\sum_{\omega} p^{\omega} \log \frac{p^{\omega}}{q^{\omega}}$.

Commonly encountered measures of uncertainty include variance $H(q) = \text{Var}_q[\omega]$ (when $\Omega \subset \mathbb{R}$) and Shannon entropy $H(q) = -\sum_{\omega} q^{\omega} \log q^{\omega}$.

We are interested in whether the aforementioned (and other) measures of information and uncertainty quantify attitudes toward information in some decision problem. Say *a measure of information d is valid* if there is a decision problem \mathcal{D} such that d is a value of information for \mathcal{D} .¹⁴ Say *a measure of uncertainty H is valid* if there is a decision problem \mathcal{D} such that H is the cost of uncertainty for \mathcal{D} .

We are also interested in whether a given measure of information and a measure of uncertainty are “consistent” with each other. We introduce two notions of consistency. Say that d and H are *jointly valid* if they arise from the same decision problem, i.e., there is a decision problem \mathcal{D} such that d is a value of information for \mathcal{D} and H is the cost of uncertainty for \mathcal{D} . We say that d and H are *coupled* if for every prior q and every signal π_s , we have $\mathbb{E}[d(q(s), q)] = \mathbb{E}[H(q) - H(q(s))]$; in other words, the expected amount of information generated equals the expected reduction in uncertainty.¹⁵

3 Main results

In this section, we answer the two questions posed at the end of the previous section: (i) which measures of information and uncertainty are valid, and (ii) which measures of information and uncertainty are consistent with each other.

¹⁴We could also define an ex ante measure of information $D(\pi, q)$ that captures the amount of information generated by a signal π given prior q , and an ex ante value of information for a given decision problem $V_{\mathcal{D}}(\pi, q)$ that captures the expected utility benefit of observing signal π given prior q in decision problem \mathcal{D} . Then, we might analogously say that D is valid if there is a \mathcal{D} such that $D = V_{\mathcal{D}}$. It is easy to show that $D(\pi, q)$ is valid if and only if $D(\pi_{\alpha}, q) = \mathbb{E}[H(q) - H(q(\alpha))]$ for some concave H .

¹⁵We will sometimes say “ d is coupled with H ” or “ H is coupled with d ” in place of “ d and H are coupled.”

3.1 Valid measures of information

Consider the following potential properties of a measure of information.

(*Null-information*) $d(q, q) = 0 \forall q$.

(*Positivity*) $d(p, q) \geq 0 \forall (p, q)$.

(*Order-invariance*) Given any prior and pair of signals, the expected sum of information generated by the two signals is independent of the order in which they are observed. Formally, for every q , π_α , and π_β , we have

$$\mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[d(q(\beta), q) + d(q(\alpha \cap \beta), q(\beta))].$$

Null-information and Positivity impose ex post properties on the realized amount of information, while Order-invariance imposes an ex ante property regarding the expected amount of information that will be generated.¹⁶

In our approach, a measure of information is a function of just the prior and the posterior. It explicitly excludes news that does not affect the posterior. This is analogous to our focus on valuing a piece of information without consideration of the extent to which it substitutes or complements other information available. Accordingly, Null-information states that news that does not change beliefs does not generate information.

Positivity requires that the amount of information received is always weakly positive, even from the ex post perspective. This puts it in contrast with the approach common in information theory – and often used in the economics literature on rational inattention – that measures information as the expected reduction in entropy. Because a particular piece of news can increase entropy, ex post information measured by reduction in entropy can be negative.

Order-invariance concerns sequences of signals. It echoes a crucial feature of Bayesian updating, namely that the order in which information is observed cannot affect its overall content. That is, observing π_α followed by π_β leads to the same final distribution of posteriors as observing π_β followed by π_α . Order-invariance requires that these two sequences must generate the same expected sum of

¹⁶Note that requiring that d satisfy the ex post version of Order-invariance would be overly restrictive. Only $d(p, q)$ identically equal to zero everywhere would satisfy Null-information, Positivity, and the ex post version of Order-invariance.

ex post information. While Order-invariance is stated in terms of pairs of signals, in combination with Null-information it guarantees that sequences of any number of signals that lead to the same final distribution of posteriors necessarily generate the same expected sum of information.

Theorem 1. *A measure of information is valid if and only if it satisfies Null-information, Positivity, and Order-invariance.*

Proofs are in the Appendix. It is immediate that a measure of information is valid only if it satisfies Null-information and Positivity. The less obvious parts of Theorem 1 are that (i) every valid measure satisfies Order-invariance, and (ii) any measure that satisfies these three properties is valid. To establish (i), we first show that every valid measure of information d is necessarily coupled with some measure of uncertainty H . (In fact, if d is a value of information for some decision problem \mathcal{D} , then it is coupled with the cost of uncertainty for \mathcal{D} ; see Proposition 6.) Moreover, it is straightforward to see that to be coupled with any measure of uncertainty, a measure of information must satisfy Order-invariance (see Lemma 4). To establish (ii), we show how to explicitly construct a decision problem that underlies any given measure of information that satisfies the three axioms (see proof of Proposition 7). Specifically, we set $A = \Delta(\Omega)$ and $u(a, \omega) = -d(\delta_\omega, a)$.

A property closely related to Order-invariance would be that the expected sum of information generated by observing one signal and then the other is the same as the expected amount of information generated by observing the two signals at once. Formally, say that d satisfies *Combination-invariance* if for every q , π_α , and π_β , we have

$$\mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[d(q(\alpha \cap \beta), q)].$$

It turns out that Combination-invariance is equivalent to Null-information and Order-invariance (see Appendix A.4).¹⁷ Therefore, Theorem 1 can alternatively be stated in terms of Positivity and Combination-invariance.

Theorem 1 not only identifies the precise set of properties that a measure of information compatible with a decision-theoretic foundation must satisfy, it also provides an easy way to test whether a given measure is valid or not. It is of course straightforward to check Null-information and

¹⁷This echoes the equivalence of divisibility and order independence axioms in Cripps (2019).

Positivity, and in Section 3.4 we establish two simple tests for Order-invariance.

3.2 Valid measures of uncertainty

It is easy to see that for every decision problem, the cost of uncertainty is concave and is equal to zero when the decision-maker knows the state. These two properties are not only necessary but sufficient for validity. Formally, consider the following two properties.

(Null-uncertainty) $H(\delta_\omega) = 0$ for all ω .

(Concavity) H is concave.

Theorem 2. *A measure of uncertainty is valid if and only if it satisfies Null-uncertainty and Concavity.*

Note that Null-uncertainty and Concavity jointly imply that H is positive everywhere. Moreover, the concavity of H implies that observing any signal necessarily reduces expected uncertainty, i.e., for every q and π_s , $\mathbb{E}[H(q(s))] \leq H(q)$.

We present Theorem 2 as a formal result, but we do not see this result as a substantive contribution of this paper. Theorem 2 is basically a restatement of the well-known result that for every decision problem, the maximized expected utility is convex in the decision-maker's belief and moreover, every convex function of beliefs is the maximized expected utility for some decision problem.¹⁸

3.3 Consistency of measures of information and uncertainty

In Section 2.3, we introduced two distinct notions of consistency, joint validity and coupling. Here we show that (given valid measures of information and uncertainty) these two notions of consistency coincide. Moreover, there are specific functional forms that relate the two measures.

Recall that a *supergradient* of a concave function H at belief q is a vector $\nabla H(q)$ such that for every belief p we have $H(q) + \nabla H(q) \cdot (p - q) \geq H(p)$. For every concave function, $\nabla H(q)$ exists for all q .¹⁹ When H is smooth at q , $\nabla H(q)$ is unique and equal to $H'(q)$. A *Bregman divergence*

¹⁸See Mensch (2018a) for a recent statement and application of this result.

¹⁹When q is interior, it is immediate that $\nabla H(q)$ always exists. If q is at the boundary of $\Delta(\Omega)$, we need to be somewhat careful. In particular, if $q^\omega = 0$, we set $\nabla H(q)^\omega = \infty$ with $\nabla H(q)^\omega (p - q)^\omega = 0$ if $p^\omega = q^\omega$; similarly, if $q^\omega = 1$, we set $\nabla H(q)^\omega = -\infty$ with $\nabla H(q)^\omega (p - q)^\omega = 0$ if $p^\omega = q^\omega$. These conditions guarantee that $\nabla H(q)$ also exists at the boundary.

of a concave function H is a function from (p, q) to the real numbers equal to

$$H(q) - H(p) + \nabla H(q) \cdot (p - q)$$

for some supergradient $\nabla H(q)$ of H at q (Bregman 1967).

Theorem 3. *Given a valid measure of information d and a valid measure of uncertainty H , the following are equivalent:*

- (1) d and H are jointly valid,
- (2) d and H are coupled,
- (3) $H(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$,
- (4) d is a Bregman divergence of H .

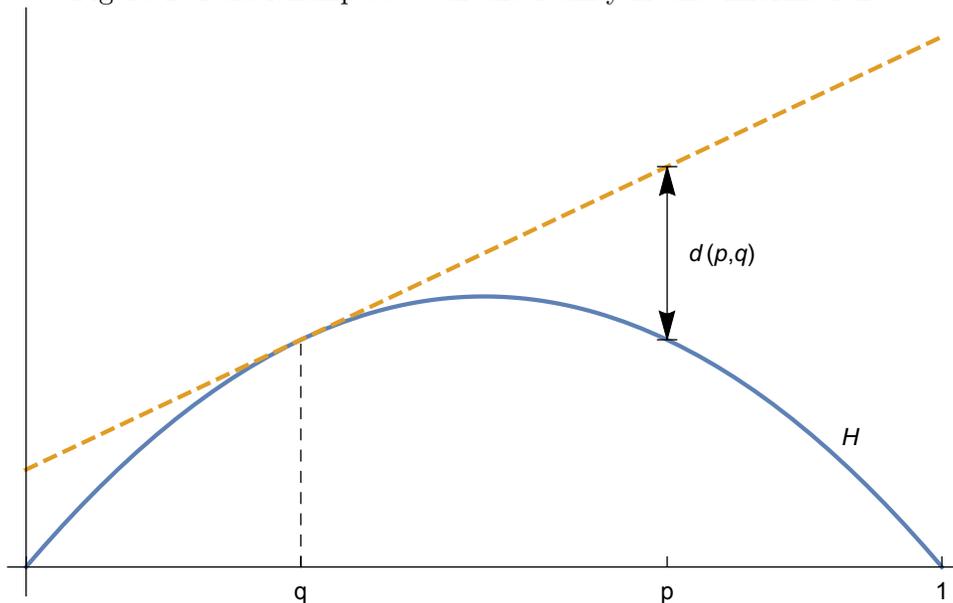
Moreover, as Propositions 4 and 5 below clarify, given any valid H , d is jointly valid with H if and only if it is the Bregman divergence of H , and given any valid d , H is jointly valid with d if and only if $H(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$.

We see the equivalence of (1) and (2) as the main contribution of Theorem 3. It tells us that we can associate the (expected) reduction in uncertainty with the (expected) amount of information that is generated if and only if we can microfound the two measures using the same decision problem. The relationship between d and H expressed by part (3) has a simple interpretation: uncertainty is the expected amount of information generated by observing a fully revealing signal. Finally, to gain intuition for part (4) of the Theorem, first note that if a measure of information d is coupled with some H , it must be the case that $d(p, q) = H(q) - H(p) + f(q)(p - q)$ for some function f . (No matter what f is, for every signal π_s , $\mathbb{E}[f(q)(q(s) - q)]$ must equal 0; hence, $\mathbb{E}[d(q(s), q)] = \mathbb{E}[H(q) - H(q(s)) + f(q)(q(s) - q)] = \mathbb{E}[H(q) - H(q(s))]$.) Then, to ensure that d only takes on positive values, we must have f be a supergradient of H , i.e., d must be a Bregman divergence of H .²⁰

Part (4) of Theorem 3 highlights a geometric relationship between uncertainty and information through the Bregman divergence characterization. Figure 1 illustrates the connection between an

²⁰The literature on axiomatic foundations of various divergences (including Bregman) is summarized in Csiszar (2008). Banerjee et al. (2005) use Bregman divergences as a foundation for measures of information, with applications to clustering problems in machine learning. Zanardo (2017) provides an axiomatic derivation of divergences suitable for measuring disagreement between decision-makers.

Figure 1: Relationship between uncertainty H and information d



arbitrary valid measure of uncertainty and its coupled measure of information. Gossner and Steiner (2018) utilize a closely related figure to illustrate that what we term value of information can – in the case of binary states – be derived from the second derivative of $\mathbb{E}_p [u(a^*(p), \omega)]$ with respect to p .

3.4 Checking validity

Theorem 2 provides an easy way to confirm whether a given measure of uncertainty is valid. Applying Theorem 1 to check whether a given measure of information is valid may not seem as straightforward because it requires checking Order-invariance, which is not a self-evident property. Fortunately, however, there are two ways around this issue.

First, if a measure of information is smooth, it is possible to confirm whether it satisfies Order-invariance simply by inspecting its derivatives:

Proposition 1. *Suppose a measure of information d is twice-differentiable in p for all q and satisfies Null-information. Then, d satisfies Order-invariance if and only if $\frac{\partial^2 d(p,q)}{\partial p^2}$ is independent of q .*

Second, the functional forms from Theorem 3 provide an easy way to check whether a measure of information (smooth or not) satisfies Order-invariance.

Proposition 2. *Suppose a measure of information d satisfies Null-information and Positivity. Then, d satisfies Order-invariance if and only if d is a Bregman divergence of $\sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$.*

Proposition 2 is easily applied since confirming whether

$$d(p, q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q) - \sum_{\omega} p^{\omega} d(\delta_{\omega}, p) + \nabla \left(\sum_{\omega} q^{\omega} d(\delta_{\omega}, q) \right) \cdot (p - q)$$

is a straightforward computation for any given d . Moreover, since a Bregman divergence cannot be a metric, Proposition 2 also implies that any metric on the space of beliefs violates Order-invariance.²¹

Corollary 1. *If a measure of information d is a metric on $\Delta(\Omega)$, it does not satisfy Order-invariance and thus cannot be valid.*

Corollary 1 further implies that no metric can be coupled with any measure of uncertainty, since (as shown below in Lemma 4 in Appendix A.2) a coupled measure of information necessarily satisfies Order-invariance.²²

3.5 Discussion

Consider again the examples of decision problems from Section 2. Theorem 1 thus tells us that the value of information functions v^r , v^Q , and v^B all satisfy Null-information, Positivity, and Order-invariance,²³ while Theorem 2 confirms that the cost of uncertainty functions C^r , C^Q , and C^B satisfy Null-uncertainty and Concavity. Moreover, Theorem 3 tells us that each of these corresponding pairs of measures is coupled. If we want our measure of uncertainty to be “consistent” with our measure of information (in the sense that every signal generates information to the same extent that it reduces uncertainty), then we can measure uncertainty with C^r and information with v^r , or we can measure uncertainty with C^Q and information with v^Q , etc., but we cannot “mix and match.”

²¹The only way that a Bregman divergence d can satisfy the triangle inequality on a triplet p, q, r when p is a convex combination of q and r is if $d(p, q) = 0$. Therefore, a Bregman divergence cannot satisfy both the triangle inequality and $d(p, q) = 0 \implies p = q$, and hence cannot be a quasimetric. (Recall that a quasimetric is a function that satisfies the axioms for a metric with the possible exception of symmetry.)

²²Augenblick and Rabin (2018) also note that there is no measure of uncertainty that can be coupled with Euclidean distance.

²³While the functional forms make Null-information and Positivity easy to see, simply inspecting these functions does not make it obvious that they satisfy Order-invariance. There is a simple direct proof of Order-invariance of v^Q and v^B , but directly establishing Order-invariance of v^r is not as straightforward.

In fact, our results imply that every concave function (equal to zero at degenerate beliefs) is a valid measure of uncertainty and is consistent with some measure of information – namely its Bregman divergence. Take entropy. Our results tell us that entropy $H(q) = -\sum_{\omega} q^{\omega} \log q^{\omega}$ is coupled with its Bregman divergence, which, as shown by Bregman (1967), is Kullback-Leibler divergence, $d(p, q) = \sum_{\omega} p^{\omega} \log \frac{p^{\omega}}{q^{\omega}}$. For every signal, the expected reduction in entropy equals the expected Kullback-Leibler divergence. Such coupling (with a suitable measure of information) is not a special feature of entropy, but rather holds for every concave function normalized to zero at degenerate beliefs. Indeed, the fact that for every signal, the expected reduction in residual variance equals the expected quadratic distance between the prior and the posterior plays a central role in the derivation of suspense-optimal entertainment policies in Ely et al. (2015). Augenblick and Rabin (2018) emphasize this result (cf: their Proposition 1) and use it as a cornerstone for an empirical test of rationality of belief updating. They also introduce a class of measures of uncertainty and coupled measures of information, which they term “measures of movement,” that encompass variance (coupled with quadratic distance) and entropy (coupled with Kullback-Leibler divergence).²⁴

Moreover, our results tell us that we can “microfound” any (normalized, concave) measure of uncertainty – and its coupled measure of information – as arising from some decision problem. In fact, as we discussed above, the constructive proof of Proposition 7 in Appendix A.2 provides an explicit algorithm for finding the underlying decision problem. Given the H and the d , if we set $A = \Delta(\Omega)$ and $u(a, \omega) = -d(\delta_{\omega}, a)$, we obtain H as the cost of uncertainty and d as the value of information.²⁵ These decision problems can be interpreted as follows: the action is a report of a probability distribution of an unknown variable, and utility is maximized by reporting one’s true belief. In other words, the decision problem corresponds to a “proper scoring rule” for eliciting an honest report of beliefs. The relationship between proper scoring rules, decision problems, and

²⁴Assuming a binary state space $\{0, 1\}$, Augenblick and Rabin (2018) define a *surprisal function* as any decreasing $\gamma : [0, 1] \rightarrow \mathbb{R}_+$ with $\gamma(1) = 0$ and let $H(q) = q\gamma(q) + (1-q)\gamma(1-q)$. They show each such H is coupled with $d(p, q) = p(\gamma(q) - \gamma(p)) + (1-p)(\gamma(1-q) - \gamma(1-p))$. Measures of uncertainty that are generated this way are all symmetric around $q = \frac{1}{2}$, but they need not be concave nor equal to zero at degenerate beliefs. Likewise, the corresponding measures of information are not necessarily positive (even when H is concave). Thus, these measures of information and uncertainty are more restrictive than ours (due to symmetry), but are not always compatible with our decision-theoretic microfoundations (since they need not be valid). We discuss additional implications of our results for the questions explored by Augenblick and Rabin (2018) below.

²⁵As we explain below, we in fact need to set $u(a, \omega) = \begin{cases} -d(\delta_{\omega}, a) & \text{if } a^{\omega} > 0 \\ -\infty & \text{otherwise} \end{cases}$.

Bregman divergences has previously been explored by Gneiting and Raftery (2007).

Consider entropy. If we set $A = \Delta(\Omega)$ and $u^{KL}(a, \omega) = -\log a^\omega$ (letting $u(a, \omega) = -\infty$ if $a^\omega = 0$), we get entropy as the cost of uncertainty ($C^{KL}(q) = -\sum_{\omega} q^\omega \log q^\omega$) and Kullback-Leibler divergence as the value of information ($v^{KL}(p, q) = \sum_{\omega} p^\omega \log \frac{p^\omega}{q^\omega}$).²⁶ This microfoundation for entropy was first provided by Kelly (1956). The utility function used corresponds to the familiar logarithmic scoring rule (Good 1952). This example also illustrates why we needed to extend the potential range of the utility function to negative infinity in our definition of decision problems (see footnote 7). Some measures of uncertainty, such as entropy, have the property that the marginal reduction of uncertainty goes to infinity as the belief approaches certainty; the extended range allows us to include such measures. There is no decision problem with a finite-valued utility function that has entropy as its cost of uncertainty.

That said, a decision problem that has some particular d and H as its value of information and cost of uncertainty is not unique. In information theory, for example, entropy and Kullback-Leibler divergence are often presented as arising from a decision problem that is quite different from $A = \Delta(\Omega)$ and $u(a, \omega) = -\log a^\omega$. If a decision-maker needs to choose a *code*, i.e., a map from ω to a string in some alphabet, aiming to minimize the expected length of the string, entropy arises as the cost of uncertainty and Kullback-Leibler divergence arises as the value of information (Cover and Thomas 2012).²⁷

In Appendix A.1, we provide yet another foundation for entropy and Kullback-Leibler divergence in terms of collections of simple decision problems. In fact, we show that when the state space is binary, *every* valid measure of information or uncertainty can be expressed as arising from some measure over simple decision problems. In other words, the simple decision problems form a basis of sorts for all decision problems. The uniform measure (a constant density of cutoffs r) yields the cost of uncertainty equal to residual variance, while the measure with density $\frac{1}{r(1-r)}$ yields entropy. Thus, residual variance and entropy can be seen as two elements of a parametric family. Consider measures over simple decision problems with density equal to $r^\gamma (1-r)^\gamma$ for $\gamma > -2$. If $\gamma = 0$, we get residual variance; if $\gamma = -1$, we get entropy. For any $\gamma > -1$, the density integrates

²⁶Note that $v^{KL}(\delta_\omega, a) = \sum_{\omega'} \delta_{\omega'} \log \frac{\delta_{\omega'}}{a^{\omega'}} = -\log a^\omega$.

²⁷There are also various axiomatic approaches to deriving entropy and Kullback-Leibler divergence (see survey by Csiszar 2008).

to a finite amount and thus can be scaled to a probability distribution (a symmetric Beta). For $\gamma \leq -1$, the measure is not finite. Thus, entropy can be seen as the “border case” between costs of uncertainty that are derived from probability distributions and those that can only arise from infinite measures.²⁸

These results provide additional insight into the tests of rationality proposed by Augenblick and Rabin (2018). They emphasize that if a sequence of beliefs was formed through Bayes’ rule, the expected sum of belief movements as measured by quadratic variation must equal residual variance of the initial belief, though as we mentioned in footnote 24, they discuss a broader (but non-comprehensive) class of tests. Our results (see Lemma 5 in Appendix A.2) imply that the set of all tests of Bayesian rationality that confirm whether expected reduction in uncertainty equals expected belief movement is spanned by picking *any* pair of functions H and f and then measuring uncertainty by H and belief movement by $H(q) - H(p) + f(q)(p - q)$. Moreover, if one wishes to guarantee that belief movements are non-negative, H must be concave and f must be a supergradient of H .

That said, at least for the case of binary states, there is one sense in which Augenblick and Rabin’s (2018) choice to focus on quadratic variation is the most natural. Every test of the Augenblick-Rabin variety that defines belief movement as non-negative – and thus uses a valid measure of information as its measure of belief movement – implicitly puts a certain “weight” on movements that cross specific beliefs. For example, suppose we use v^r with $r = 0.5$ as the measure of belief movement. If a sequence of beliefs (q_0, \dots, q_T) was formed through Bayes’ rule and q_T is degenerate, it must be the case that $\sum_{t=1}^T v^{0.5}(q_t, q_{t-1})$ in expectation equals $C^{0.5}(q_0)$. This test only “counts” movements that cross the belief $r = 0.5$, ignoring all others. If we were to use Kullback-Leibler divergence as the measure of belief movement, we would be putting weight $-\frac{1}{r(1-r)}$ on movements across r . Quadratic variation, the test used by Augenblick and Rabin (2018), corresponds to uniform weights.

Our results are also useful insofar as they reveal that certain seemingly sensible measures of information are not valid, i.e., cannot be a microfoundation in our decision-theoretic terms. For example, Euclidean distance between the prior and the posterior is not a valid measure of infor-

²⁸When $\gamma \leq -2$, the measure is not equivalent to any decision problem since it would imply an infinite cost of uncertainty for interior beliefs.

mation. While this measure satisfies Null-information and Positivity, it is easy to see that it does not satisfy Order-invariance. Under this measure, any partially informative signal followed by a fully informative signal yields a higher expected sum of information than a fully informative signal does on its own. In fact, Corollary 1 tells us that every metric violates Order-invariance, and hence there does not exist a decision problem whose value of information is a metric on beliefs.

Finally, our results have implications for estimation of the ex ante value of sources of information. Suppose a (frequentist) econometrician wishes to estimate the value W of some information source – an unknown signal π_α – for a decision-maker with a prior q who faces some decision problem $\mathcal{D} = (A, u)$. The econometrician has data on how other decision-makers (with prior q)²⁹ updated their posteriors after observing information from the source. Let $\mathbf{p} = (p_1, \dots, p_N)$ denote this dataset of posteriors. Since the value of π_α can be written as the expected reduction in the cost of uncertainty, $W = \mathbb{E}[C_{\mathcal{D}}(q) - C_{\mathcal{D}}(p)]$, one natural estimator of W would be $\hat{W}(\mathbf{p}) = \frac{\sum_{i=1}^N C_{\mathcal{D}}(q) - C_{\mathcal{D}}(p_i)}{N}$. That said, since W can also be expressed as the average of the ex post value of information, $W = \mathbb{E}[v_{\mathcal{D}}(q(\alpha), q)]$, another reasonable estimator would be $\hat{W}(\mathbf{p}) = \frac{\sum_{i=1}^N v_{\mathcal{D}}(p_i, q)}{N}$. Both of these estimators are unbiased. In fact, if we consider estimators of the form $\hat{W}(\mathbf{p}) = \frac{\sum_{i=1}^N w(p_i)}{N}$, Lemma 5 implies that \hat{W} is an unbiased estimator of W if and only if $w(p) = C_{\mathcal{D}}(q) - C_{\mathcal{D}}(p) + k(p - q)$ for some constant k . The two estimators above correspond to setting $k = 0$ and $k = \nabla C_{\mathcal{D}}(q)$, respectively. Note, however, that the first estimator sometimes takes on negative values even though we know that $W \geq 0$. If we impose the desideratum that \hat{W} lies in the constrained parameter space (i.e., that $\hat{W}(\mathbf{p}) \geq 0$ for all \mathbf{p}), we must set $k = \nabla C_{\mathcal{D}}(q)$. Hence, if some $\hat{W}(\mathbf{p}) = \frac{\sum_{i=1}^N w(p_i)}{N}$ is an unbiased estimator of the value of an information source for a decision problem and it respects the parameter constraints, then it must be a sample average of a valid ex post measure of information.

4 Buying information

In this section, we establish a relationship between valid measures of information and intertemporal incentive-compatibility constraints faced by a seller of information. Consider the following model of a buyer who compensates a seller for the information that the seller reveals.

²⁹We focus on homogeneous priors for ease of exposition. Our results readily extend to heterogeneous priors.

The prior is q . There are two periods $t \in \{1, 2\}$ and two available signals π_{α^*} (which arrives in period 1) and π_{β^*} (which arrives in period 2). The seller will eventually reveal all of the information from these signals, but may delay doing so. There are two types of delay to consider.

First, the seller can delay the arrival of information from the ex ante perspective. He can choose to observe π_{α} (instead of π_{α^*}) in period 1 and π_{β} (instead of π_{β^*}) in period 2, but is restricted to π_{α} and π_{β} such that the distribution of $q(\alpha^*)$ is a mean-preserving spread of the distribution of $q(\alpha)$ and the distribution of $q(\alpha^* \cap \beta^*)$ is the same as that of $q(\alpha \cap \beta)$. In other words, he can choose to get a (Blackwell) less informative signal in the first period and “transfer” the foregone information to the second period signal.

Second, he can delay the revelation of information in the interim stage. Following the realization α of π_{α} , he can either reveal α or reveal no information. If he chooses to reveal α in period 1, in period 2 he will just reveal β . If he chooses to reveal nothing in period 1, in period 2 he must reveal both α and β . In other words, he can “hide” what he learned in period 1 and only reveal it in period 2 along with the new information that arrived in period 2.

The seller is paid for the information he reveals. Before any information is revealed, the prior is q . The payment to the seller in period 1 is $t(p_1, q)$, and in period 2, it is $t(p_2, p_1)$ for some payment function t , with p_1 and p_2 determined as follows. At period 1, if the seller revealed α , the posterior is $p_1 = q(\alpha)$; if the seller revealed no information, we set $p_1 = q$.³⁰ In either case, in period 2, the posterior is $p_2 = q(\alpha \cap \beta)$. The seller’s objective is to maximize the expected sum of transfers.

We make two assumptions about t , namely that $t(q, q) = 0$ and $t(p, q) \geq 0$ for every p and q . The latter assumption can be seen as a form of limited liability. We are interested in how the structure of t interacts with the seller’s incentives to delay information revelation. We say that t is *ex ante incentive compatible* if for every prior q and every pair of signals π_{α^*} and π_{β^*} , the seller weakly prefers to set $\pi_{\alpha} = \pi_{\alpha^*}$ (and thus $\pi_{\beta} = \pi_{\beta^*}$). We say that t is *interim incentive compatible* if for every prior q and every pair of signals π_{α} and π_{β} , the seller weakly prefers to reveal every signal realization α in period 1. If t is both ex ante and interim incentive compatible, we simply say it is *incentive compatible*.³¹

³⁰Note that if the decision not to reveal α conveys information about ω , the equilibrium posterior p_1 would not equal q . However, we are interested in settings where the seller is paid for the explicit, verifiable information he provides and thus we rule out t being a function of updating based on implicit information.

³¹As is typical, we only require that the seller weakly prefer not to delay. It is worthwhile to note that, not only

At first glance, it might seem that interim incentive compatibility is more difficult to satisfy than ex ante incentive compatibility since the seller can condition the decision of whether to delay on the first period’s signal realization. This turns out not to be the case; in fact, ex ante incentive compatibility is the stronger condition. One example of a payment function that is interim but not ex ante incentive compatible is $t(p, q) = \|p - q\|$.³² But, there is no payment function that is ex ante but not interim incentive compatible:

Lemma 1. *Every payment function that is ex ante incentive compatible is interim incentive compatible.*

We now turn to the question of which payment functions are (ex ante) incentive compatible. We have seen that $t(p, q) = \|p - q\|$ does not work. It turns out that this fact is closely tied to the observation that Euclidean distance is not a valid measure of information.

Theorem 4. *A payment function is incentive compatible if and only if it is a valid measure of information.*

The idea behind the proof is to show that if a payment function satisfies Order-invariance, then delaying any signal (at the ex ante or interim stage) has zero impact on the expected payment. Thus, the seller is always indifferent between revealing or delaying information. One may wish to find a payment function that induces a strict preference against delay, but since Theorem 4 is an if-and-only-if result, it tells us that any payment scheme that leads to a strict preference not to delay some information necessarily induces a strict preference to delay other information. Thus, making the seller indifferent about the delay is the only way to ensure incentive compatibility.

is it not possible to construct a payment function that always gives a strict incentive not to delay, Theorem 4 below implies that any payment function that ever gives a strict incentive not to delay cannot be incentive compatible.

³²Given any q and signal realizations α and β , if $q(\alpha)$ is a convex combination of q and $q(\alpha \cap \beta)$, the payment under $t(p, q) = \|p - q\|$ is the same whether α had been revealed or not; otherwise, the payment is higher if α had been revealed. Hence, the payment function is interim incentive compatible. To see it is not ex ante incentive compatible, consider π_{α^*} that is informative and π_{β^*} that provides no information. Then, it is a profitable deviation to “split” the informational content of π_{α^*} across the two periods.

A Appendix

A.1 Collections of simple decision problems

As we noted in the paper, two decision problems can be equivalent in the sense that they have the same cost of uncertainty. Similarly, some collection of decision problems can induce the same attitude toward information and uncertainty as some particular decision problem \mathcal{D} . In this section, we show that, when the state space is binary, every decision problem corresponds to some collection of simple decision problems. Formally, a *simple decision environment* μ is a measure on $[0, 1]$ with the interpretation that the decision-maker faces a collection of simple decision problems with the measure indicating their prevalence. The *cost of uncertainty of q given μ* – denoted $K_\mu(q)$ – is the reduction in the decision-maker’s payoff, aggregated across the decision problems in μ , due to her ignorance of the state of the world, i.e.,

$$K_\mu(q) = \int \left[\mathbb{E}_q \left[\max_a [u^r(a, \omega)] \right] - \max_a [\mathbb{E}_q [u^r(a, \omega)]] \right] d\mu(r).$$

Given a binary state space, we say that a decision problem \mathcal{D} is *equivalent* to a simple decision environment μ if $C_{\mathcal{D}}(q) = K_\mu(q)$ for all q .³³

Our main result in this section is that every decision problem is equivalent to some simple decision environment. This basically means that (when the state space is binary) we can think of the simple decision problems as a “basis” for all decision problems, at least as far as value of information is concerned. This result is closely related to the characterization of proper scoring rules in Schervish (1989). The argument we present, which emphasizes the role of the second derivative of the cost of uncertainty, follows along similar lines as the analysis in Gossner and Steiner (2018).

Proposition 3. *Suppose the state space is binary. Every decision problem is equivalent to some simple decision environment.*

The formal proof is below, but to get some intuition for this result, first consider a simple decision environment that puts measure 1 on some simple decision problem r . Its cost of uncertainty, as

³³We can also define value of information for a simple decision environment and note that if some function is a value of information both for some decision problem \mathcal{D} and for some environment μ , then \mathcal{D} and μ are equivalent. That said, we define equivalence in terms of the cost of uncertainty, since that (unlike value of information) is unique for a given decision problem.

noted above, is

$$C^r(q) = \begin{cases} q(1-r) & \text{if } q \leq r \\ (1-q)r & \text{if } q > r \end{cases},$$

i.e., a piecewise linear function with slope $1-r$ for $q < r$ and slope $-r$ for $q > r$; in other words, the decrease in slope at r is 1. Next, consider an environment that puts measure η on some simple decision problem r . Its cost of uncertainty is just $\eta C^r(q)$ with the decrease in slope at r of η . Now, consider an environment $\mu = ((\eta_1, r_1), \dots, (\eta_k, r_k))$ that puts measure η_i on simple decision problem r_i for $i \in \{1, \dots, k\}$. Its cost of uncertainty $K_\mu(q)$ is $\sum_i \eta_i C^{r_i}(q)$, which is a piecewise linear function whose slope decreases by η_i at each r_i .

Hence, given any decision problem \mathcal{D} with a finite action space A – and therefore a piecewise linear cost of uncertainty $C_{\mathcal{D}}(q)$ – we can find an equivalent simple decision environment by putting a measure $\eta_i = \lim_{q \rightarrow r_i^-} C'_{\mathcal{D}}(q) - \lim_{q \rightarrow r_i^+} C'_{\mathcal{D}}(q)$ (i.e., the decrease in slope) at each kink r_i of $C_{\mathcal{D}}(q)$.

For example, consider $A = \{x, m, y\}$ and $\Omega = \{X, Y\}$, where $u(a, \omega)$ is indicated by the matrix

$$\begin{array}{cc} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} x \\ m \\ y \end{array} & \begin{array}{cc} 1 & 0 \\ \frac{1}{3} & \frac{3}{4} \\ 0 & 1 \end{array} \end{array}.$$

Under these preferences, x is optimal when the probability of Y is $q \leq \frac{8}{17}$, m is optimal for $q \in [\frac{8}{17}, \frac{4}{7}]$, and y is optimal when $q \geq \frac{4}{7}$. Then, $C_{\mathcal{D}}(q)$ is piecewise linear with kinks at $r_1 = \frac{8}{17}$ and $r_2 = \frac{4}{7}$ and slope decreases of $\frac{17}{12}$ at r_1 and $\frac{7}{12}$ at r_2 . Therefore, this problem is equivalent to a simple decision environment that puts measure $\eta_1 = \frac{17}{12}$ on $r_1 = \frac{8}{17}$ and $\eta_2 = \frac{7}{12}$ on $r_2 = \frac{4}{7}$.

The logic above can be extended to decision problems with a continuous action space and smooth cost of uncertainty by setting the density of the simple decision environment equal to the infinitesimal decrease in slope of $C_{\mathcal{D}}(q)$, i.e., $-C''_{\mathcal{D}}(q)$.³⁴ For example, consider the decision problem

³⁴The formal proof handles mixtures of kinks and smooth decreases in $C'_{\mathcal{D}}(q)$.

$A = [0, 1]$ and $\Omega = \{X, Y\}$ with

$$u(a, \omega) = \begin{cases} -\frac{a^2}{4} & \text{if } \omega = X \\ -\frac{(1-a)^2}{4} & \text{if } \omega = Y \end{cases},$$

whose cost of uncertainty $C_{\mathcal{D}}(q) = \frac{q(1-q)}{2}$ is proportional to residual variance. Since $-C_{\mathcal{D}}''(q) = 1$ for all q , this decision problem is equivalent to a uniform measure over simple decision problems. In other words, a decision-maker who is equally likely to face any simple decision problem has value of information and cost of uncertainty proportional to quadratic variation and residual variance, respectively.

Similarly, suppose the decision-maker faces the decision problem whose cost of uncertainty is entropy, i.e., $A = [0, 1]$ and $\Omega = \{X, Y\}$ with

$$u(a, \omega) = \begin{cases} -\log(1-q) & \text{if } \omega = X \\ -\log(q) & \text{if } \omega = Y \end{cases}.$$

Since entropy is given by $-q \log q - (1-q) \log(1-q)$, its second derivative is $-\frac{1}{q(1-q)}$. Thus, a decision-maker has entropy-like attitude toward information and uncertainty when there is a high prevalence of simple decision problems with very low and very high cutoffs relative to those with cutoffs near $\frac{1}{2}$. In fact, while the entropy-equivalent measure on any interval $(\underline{z}, \bar{z}) \subset (0, 1)$ is finite, it diverges to infinity as \underline{z} approaches 0 or \bar{z} approaches 1. This is the primary reason why we needed to define simple decision environments as general measures (rather than probability distributions). The formulation in terms of measures allows us to accommodate costs of uncertainty – such as entropy – that have infinite slopes at the boundary.³⁵

Proof of Proposition 3. Suppose the state space is binary. Consider some decision problem \mathcal{D} . Let $C(q)$ denote its cost of uncertainty. For $q \in (0, 1)$, let $F(q) = -C'_+(q)$ where C'_+ denotes the right-derivative. Since C is concave, F is well-defined, increasing, and right-continuous. Hence, there exists a measure μ on $[0, 1]$ such that $\mu((\underline{z}, \bar{z}]) = F(\bar{z}) - F(\underline{z})$ for all $0 < \underline{z} < \bar{z} < 1$ and

³⁵If μ is equivalent to some \mathcal{D} with cost of uncertainty $C_{\mathcal{D}}(q)$, μ must have measure of $C'_{\mathcal{D}}(0) - C'_{\mathcal{D}}(1)$ on the unit interval.

$\mu(\{0\}) = \mu(\{1\}) = 0$. We wish to show that μ is a simple decision environment equivalent to \mathcal{D} .

Let $g_q(r) = \mathbb{E}_q[\max_a [u^r(a, \omega)]] - \max_a [\mathbb{E}_q [u^r(a, \omega)]]$, i.e.,

$$g_q(r) = H^r(q) = \begin{cases} q(1-r) & \text{if } q \leq r \\ (1-q)r & \text{if } q > r \end{cases}.$$

The cost of uncertainty of μ can be expressed as a Riemann–Stieltjes integral $K_\mu(q) = \int_0^1 g_q(r) dF(r)$. Applying integration by parts, we get $K_\mu(q) = \lim_{r \rightarrow 1} g_q(r) F(r) - \lim_{r \rightarrow 0} g_q(r) F(r) - \int_0^1 F(r) dg_q(r)$. We first want to show that $\lim_{r \rightarrow 1} g_q(r) F(r) = \lim_{r \rightarrow 0} g_q(r) F(r) = 0$. (It is immediate that $g_q(1) = g_q(0) = 0$, but $F(r)$ might approach infinity as r approaches 0 or 1.) For every r sufficiently small, we have that $g_q(r) = r(1-q)$ and $C'_+(r) \geq 0$. Hence, since C is concave with $C(0) = 0$, we have $0 \leq rC'_+(r) \leq C(r)$. Hence, $\lim_{r \rightarrow 0} rC'_+(r) = 0$ and thus $\lim_{r \rightarrow 0} g_q(r) F(r) = -(1-q) \lim_{r \rightarrow 0} rC'_+(r) = 0$. A similar argument establishes $\lim_{r \rightarrow 1} g_q(r) F(r) = 0$. Thus, $K_\mu(q) = -\int_0^1 F(r) dg_q(r)$. Now, since g_q is absolutely continuous, the Riemann–Stieltjes integral $\int_0^1 F(r) dg_q(r)$ equals the Riemann integral $\int_0^1 F(r) g'_q(r) dr$, with

$$g'_q(r) = \begin{cases} 1-q & \text{if } r < q \\ -q & \text{if } r \geq q \end{cases}.$$

Thus,

$$\begin{aligned} \int_0^1 F(r) g'_q(r) dr &= \int_0^q F(r) (1-q) dr + \int_q^1 F(r) (-q) dr \\ &= (1-q)(C(0) - C(q)) - q(C(q) - C(1)) \\ &= -C(q) + qC(q) - qC(q) \\ &= -C(q). \end{aligned}$$

Hence, $K_\mu(q) = C(q)$. □

A.2 Proofs of Theorems 1-3

We begin with a Lemma that will be referenced in a number of proofs.

Lemma 2. *For every prior q , signals π_α and π_β , and d and H that are coupled, $\mathbb{E}[d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[H(q(\alpha)) - H(q(\alpha \cap \beta))]$.*

Proof of Lemma 2. By the law of iterated expectations:

$$\begin{aligned} \mathbb{E}[d(q(\alpha \cap \beta), q(\alpha))] &= \mathbb{E}[\mathbb{E}[d(q(\alpha \cap \beta), q(\alpha)) | \alpha]] \\ &= \mathbb{E}[\mathbb{E}[H(q(\alpha)) - H(q(\alpha \cap \beta)) | \alpha]] \\ &= \mathbb{E}[H(q(\alpha)) - H(q(\alpha \cap \beta))]. \end{aligned}$$

□

Next we present two Propositions that relate coupling of d and H with the properties of d and H .

Proposition 4. *Consider a measure of information d that satisfies Null-information, Positivity, and Order-invariance. The unique measure of uncertainty that satisfies Null-uncertainty and Concavity, and is coupled with d , is $H(q) = \sum_\omega q^\omega d(\delta_\omega, q)$.*

To establish Proposition 4, we first establish that Null-information and Order-invariance alone suffice to establish that $H(q) = \sum_\omega q^\omega d(\delta_\omega, q)$ is coupled with d and satisfies Null-uncertainty. We then show that the addition of Positivity of d implies Concavity of H . Finally, we establish that this is the only function that is coupled with d and satisfies Null-uncertainty.

Lemma 3. *Given a measure of information d that satisfies Null-information and Order-invariance, let $H(q) = \sum_\omega q^\omega d(\delta_\omega, q)$. Then, H is coupled with d and satisfies Null-uncertainty.*

Proof of Lemma 3. Given a measure of information d that satisfies Null-information, and Order-invariance, let $H(q) = \sum_\omega q^\omega d(\delta_\omega, q)$. Consider some q and some signal π_α . To show that H is coupled with d , we need to show $\mathbb{E}[d(q(\alpha), q)] = \mathbb{E}[H(q) - H(q(\alpha))]$. Let π_β be a fully informative signal. First, consider observing π_β followed by π_α . Since π_β is fully informative, $\mathbb{E}[d(q(\beta), q)] =$

$\sum_{\omega} q^{\omega} d(\delta_{\omega}, q) = H(q)$. Furthermore, π_{α} cannot generate any additional information so $\alpha \cap \beta = \beta$, and hence by Null-information of d , we have that $\mathbb{E}[d(q(\alpha \cap \beta), q(\beta))] = 0$. Thus, the expected sum of information generated by observing π_{β} followed by π_{α} (i.e., $\mathbb{E}[d(q(\beta), q) + d(q(\alpha \cap \beta), q(\beta))]$) equals $H(q)$. Now consider observing π_{α} followed by π_{β} . This generates expected sum of information equal to $\mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))]$. Moreover, $\mathbb{E}[d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[\sum_{\omega} q^{\omega}(\alpha) d(\delta_{\omega}, q(\alpha))] = \mathbb{E}[H(q(\alpha))]$. Hence, by Order-invariance, we have $H(q) = \mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[d(q(\alpha), q) + H(q(\alpha))]$, i.e., $\mathbb{E}[d(q(\alpha), q)] = \mathbb{E}[H(q) - H(q(\alpha))]$. Hence, H is coupled with d . Finally, Null-information and $H(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$ jointly imply Null-uncertainty. \square

We now turn to the proof of Proposition 4.

Proof of Proposition 4. Consider a measure of information d that satisfies Null-information, Positivity, and Order-invariance. Let $H(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. To show H is concave, we need to establish that for every q and every π_s , we have $\mathbb{E}[H(q) - H(q(s))] \geq 0$. By Lemma 3, we know that $\mathbb{E}[H(q) - H(q(s))] = \mathbb{E}[d(q(s), q)]$. By Positivity, $d(q(s), q) \geq 0$ for every s . Hence, $\mathbb{E}[H(q) - H(q(s))] \geq 0$. It remains to show that $H(q)$ is the unique function that is coupled with d and satisfies Null-uncertainty. Consider a fully informative signal π_s and some \hat{H} that is coupled with d and satisfies Null-uncertainty. We have that $\hat{H}(q) - \mathbb{E}[\hat{H}(q(s))] = \mathbb{E}[d(q(s), q)] = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. Null-uncertainty implies that $\mathbb{E}[\hat{H}(q(s))] = 0$. Hence, $\hat{H}(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. \square

Proposition 5. *Given a measure of uncertainty H that satisfies Null-uncertainty and Concavity, d is a measure of information that satisfies Null-information, Positivity, and Order-invariance, and is coupled with H if and only if d is a Bregman divergence of H .*

We begin the proof with two Lemmas of independent interest:

Lemma 4. *If a measure of information d is coupled with some measure of uncertainty H , d satisfies Order-invariance.³⁶*

Proof of Lemma 4. Consider any d and H that are coupled. Given any q and pair of signals π_{α}

³⁶This Lemma is reminiscent of results about the relationship between path-independence and the existence of a potential function, with Order-invariance playing the role of path-independence and the coupled measure of uncertainty playing the role of the potential.

and π_β , applying Lemma 2,

$$\begin{aligned} \mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] &= \\ \mathbb{E}[(H(q) - H(q(\alpha))) + (H(q(\alpha)) - H(q(\alpha \cap \beta)))] &= \\ \mathbb{E}[H(q) - H(q(\alpha \cap \beta))], \end{aligned}$$

and by the same argument $\mathbb{E}[d(q(\beta), q) + d(q(\alpha \cap \beta), q(\beta))] = \mathbb{E}[H(q) - H(q(\alpha \cap \beta))]$. Hence, d satisfies Order-invariance. \square

Lemma 5. *Given any measure of uncertainty H , a measure of information d is coupled with H if and only if $d(p, q) = H(q) - H(p) + f(q)(p - q)$ for some function f .*

Proof of Lemma 5. Suppose some d and H are coupled. Fix any q . We know that for every signal π_s , $\mathbb{E}[d(q(s), q) - H(q) + H(q(s))] = 0$. Since this expression is constant across all signals, $\mathbb{E}_{p \sim \tau}[d(p, q) - H(q) + H(p)]$ is constant across all distributions of posteriors τ s.t. $\mathbb{E}_{p \sim \tau}[p] = q$ (Kamenica and Gentzkow 2011). This in turn implies that $d(p, q) - H(q) + H(p)$ is some affine function of p , $f(q)p + g(q)$. Now, since $\mathbb{E}_{p \sim \tau}[f(q) \cdot p + g(q)] = 0$ for all τ s.t. $\mathbb{E}_{p \sim \tau}[p] = q$, it must be that $g(q) = -f(q)q$. Hence, $d(p, q) - H(q) + H(p) = f(q) \cdot (p - q)$. \square

We are now ready to prove Proposition 5.

Proof of Proposition 5. We first establish the “if” direction. Consider some H that satisfies Null-uncertainty and Concavity. Let $d(p, q) = H(q) - H(p) + \nabla H(q) \cdot (p - q)$ for some supergradient $\nabla H(q)$. Note that for every q and every π_s , we have $\mathbb{E}[\nabla H(q) \cdot (q(s) - q)] = 0$ and thus $\mathbb{E}[d(q(s), q)] = \mathbb{E}[H(q) - H(q(s))]$. Since this holds for all signals, d is coupled with H . Next, by Lemma 4, d satisfies Order-invariance. It is obvious that d satisfies Null-information and since it is a Bregman divergence of a concave function, it satisfies Positivity. To establish the “only if” direction, consider any d that is coupled with H and satisfies Positivity. Lemma 5 shows that $d(p, q) = H(q) - H(p) + f(q) \cdot (p - q)$ for some function f . Positivity implies that $H(q) - H(p) + f(q) \cdot (p - q) \geq 0$ for all pairs (p, q) , which means that $f(q)$ is a supergradient of $H(q)$. \square

We now turn to two Propositions that relate validity of d and H with the properties of d and H .

Proposition 6. *Given a decision problem \mathcal{D} , let $v_{\mathcal{D}}$ be a value of information for \mathcal{D} and let $C_{\mathcal{D}}$ be the cost of uncertainty for \mathcal{D} . Then:*

- (1) $v_{\mathcal{D}}$ satisfies Null-information, Positivity, and Order-invariance,
- (2) $C_{\mathcal{D}}$ satisfies Null-uncertainty and Concavity,
- (3) $v_{\mathcal{D}}$ and $C_{\mathcal{D}}$ are coupled.

Proof of Proposition 6. To establish (2), we note that $\mathbb{E}_q [\max_a [u(a, \omega)]]$ is linear in q while $\max_a [\mathbb{E}_q [u(a, \omega)]]$ is convex in q ; thus $C_{\mathcal{D}}$ is concave. It is immediate that $C_{\mathcal{D}}$ satisfies Null-uncertainty. To establish (3), consider some q and some signal π_s . Then,

$$\begin{aligned} \mathbb{E}[C_{\mathcal{D}}(q) - C_{\mathcal{D}}(q(s))] &= \mathbb{E} \left[\mathbb{E}_q \left[\max_a [u(a, \omega)] \right] - \max_a [\mathbb{E}_q [u(a, \omega)]] - \mathbb{E}_{q(s)} \left[\max_a [u(a, \omega)] \right] + \max_a [\mathbb{E}_{q(s)} [u(a, \omega)]] \right] \\ &= \mathbb{E} \left[\max_a [\mathbb{E}_{q(s)} [u(a, \omega)]] - \max_a [\mathbb{E}_q [u(a, \omega)]] \right] \end{aligned}$$

since $\mathbb{E} [\mathbb{E}_q [\max_a [u(a, \omega)]] - \mathbb{E}_{q(s)} [\max_a [u(a, \omega)]]] = 0$ by the law of iterated expectations. Moreover,

$$\begin{aligned} \mathbb{E}[v_{\mathcal{D}}(q(s), q)] &= \mathbb{E} \left[\max_a [\mathbb{E}_{q(s)} [u(a, \omega)]] - \mathbb{E}_{q(s)} [u(a^*(q), \omega)] \right] \\ &= \mathbb{E} \left[\max_a [\mathbb{E}_{q(s)} [u(a, \omega)]] - \max_a [\mathbb{E}_q [u(a, \omega)]] \right] \end{aligned}$$

for any optimal action $a^*(q)$ since for every such action $\mathbb{E} [\mathbb{E}_{q(s)} [u(a^*(q), \omega)]] = \mathbb{E}_q [u(a^*(q), \omega)] = \max_a [\mathbb{E}_q [u(a, \omega)]]$. Thus, $\mathbb{E}[C_{\mathcal{D}}(q) - C_{\mathcal{D}}(q(s))] = \mathbb{E}[v_{\mathcal{D}}(q(s), q)]$. Finally, to establish (1), note that Null-information and Positivity are immediate while Order-invariance follows from (3) by Lemma 4. \square

Proposition 7. *Suppose a measure of information d satisfies Null-information, Positivity, and Order-invariance; a measure of uncertainty H satisfies Null-uncertainty and Concavity; and d and H are coupled. There exists a decision problem \mathcal{D} such that (i) d is a value of information for \mathcal{D} and (ii) H is the cost of uncertainty for \mathcal{D} .*

Proof of Proposition 7. Suppose a measure of information d satisfies Null-information, Positivity, and Order-invariance; a measure of uncertainty H satisfies Null-uncertainty and Concavity; and d and H are coupled. Let $\mathcal{D} = (A, u)$ with $A = \Delta(\Omega)$ and

$$u(a, \omega) = \begin{cases} -d(\delta_\omega, a) & \text{if } a^\omega > 0 \\ -\infty & \text{otherwise} \end{cases}.$$

(While our proof employs payoffs of $-\infty$, these are not needed if H is continuous and has finite derivatives.) First, we note that for every p and q such that $q^\omega > 0 \Rightarrow p^\omega > 0$ we have:

$$\begin{aligned} \mathbb{E}_q[u(q, \omega) - u(p, \omega)] &= \mathbb{E}_q[-d(\delta_\omega, q) + d(\delta_\omega, p)] \\ &= \mathbb{E}_q[-H(q) + H(\delta_\omega) - \nabla H(q)(\delta_\omega - q) + H(p) - H(\delta_\omega) + \nabla H(p)(\delta_\omega - p)] \\ &= H(p) - H(q) + \nabla H(p)(q - p) = d(q, p), \end{aligned}$$

where the third equality holds because $\mathbb{E}_q[\delta_\omega] = q$. Any optimal action clearly satisfies $q^\omega > 0 \Rightarrow (a^*(q))^\omega > 0$, so for every action p that might be optimal, we have

$$\mathbb{E}_q[u(q, \omega) - u(p, \omega)] = d(q, p) \geq 0. \tag{1}$$

Hence, at every belief q , action q yields as high a payoff as any alternative action p . The value of information for \mathcal{D} , moving from prior q to posterior p is $\mathbb{E}_p[u(a^*(p), \omega)] - \mathbb{E}_p[u(a^*(q), \omega)] = \mathbb{E}_p[u(p, \omega) - u(q, \omega)]$. By the equality in Equation (1), $\mathbb{E}_p[u(p, \omega) - u(q, \omega)] = d(p, q)$. Hence, d is a value of information for \mathcal{D} . By Proposition 6, d thus must be coupled with the cost of uncertainty for \mathcal{D} which satisfies Null-uncertainty and Concavity. By Proposition 4, H is the unique measure of uncertainty that satisfies Null-uncertainty and Concavity and is coupled with d . Hence, H must be the cost of uncertainty for \mathcal{D} . \square

We are now ready to prove the main Theorems.

Proof of Theorem 1. Suppose some measure of information d is valid. By Proposition 6, it satisfies

Null-information, Positivity, and Order-invariance. Suppose d satisfies Null-information, Positivity, and Order-invariance. By Proposition 4, it is coupled with some measure of uncertainty that satisfies Null-uncertainty and Concavity. Hence, by Proposition 7, it is valid. \square

Proof of Theorem 2. Suppose some measure of uncertainty H is valid. By Proposition 6, it satisfies Null-uncertainty and Concavity. Suppose H satisfies Null-uncertainty and Concavity. By Proposition 5, it is coupled with some measure of information that satisfies Null-information, Positivity, and Order-invariance. Hence, by Proposition 7, it is valid. \square

Proof of Theorem 3. (1) implies (2) by Proposition 6. (2) implies (1) by Theorems 1 and 2 and Proposition 7. (2) is equivalent to (3) by Theorems 1 and 2 and Proposition 4. (2) is equivalent to (4) by Theorems 1 and 2 and Proposition 5. \square

A.3 Proofs of results about checking validity

Proof of Proposition 1. Consider a measure of uncertainty d that is twice-differentiable in p for all q and satisfies Null-information. Suppose d satisfied Order-invariance. Lemma 3 implies that d is coupled with some H . Therefore, by Lemma 5, it has the form $d(p, q) = H(q) - H(p) + f(q) \cdot (p - q)$. Thus, $\frac{\partial^2 d(p, q)}{\partial p^2}$ is independent of q . Now, suppose $\frac{\partial^2 d(p, q)}{\partial p^2}$ is independent of q . That means it is of the form $d(p, q) = g(q) - h(p) + f(q)p$, or equivalently $d(p, q) = \hat{g}(q) - h(p) + f(q)(p - q)$. Since d satisfies Null-information, we have $\hat{g}(q) = h(q)$ for all q , so d is of the form $d(p, q) = h(q) - h(p) + f(q)(p - q)$. Hence, by Lemma 5, d is coupled with h , and hence by Lemma 4, it satisfies Order-invariance. \square

Proof of Proposition 2. Suppose a measure of information d satisfies Null-information and Positivity. If it satisfies Order-invariance, then by Proposition 4 it is coupled with some \hat{H} . Consider a fully informative signal π_s . Since \hat{H} is coupled with d , we have $\hat{H}(q) - \mathbb{E}[\hat{H}(q(s))] = \mathbb{E}[d(q(s), q)] = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. Since $\mathbb{E}[\hat{H}(q(s))] = \sum_{\omega} q^{\omega} \hat{H}(\delta_{\omega})$, we have $\hat{H}(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q) + \sum_{\omega} q^{\omega} \hat{H}(\delta_{\omega})$.

By Lemma 5, $d(p, q) = \hat{H}(q) - \hat{H}(p) + f(q)(p - q)$ for some function f . Positivity of d implies that f is a supergradient of \hat{H} and thus that d is a Bregman divergence of \hat{H} . Since d is a Bregman divergence of $\hat{H}(q)$, it is also a Bregman divergence of $\hat{H}(q) + g(q)$ for every affine function g . Setting $g(q) = -\sum_{\omega} q^{\omega} \hat{H}(\delta_{\omega})$ yields that d is a Bregman divergence of $\sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. Now suppose that d is the Bregman divergence of $\sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. By Lemma 5, we know that d is coupled with $\sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$, and thus Lemma 4 implies it satisfies Order-invariance. \square

Proof of Corollary 1. Suppose some measure of information d satisfies Order-invariance. By Theorem 3, it is a Bregman divergence of a weakly concave function H . First, suppose H is strictly concave. Consider p, q , and r with p a convex combination of q and r . We have $d(r, q) = d(p, q) + d(r, p) + (\nabla H(q) - \nabla H(p))(r - p)$. We have that $(\nabla H(q) - \nabla H(p))(r - p) = k(\nabla H(q) - \nabla H(p))(p - q)$ for some positive k , and since H is strictly concave, we have $(\nabla H(q) - \nabla H(p))(p - q) > 0$. Hence, d does not satisfy the triangle inequality and is thus not a metric. Next, if H is weakly concave, it must be affine on some interval (p, q) , and hence for every p', q' on this interval we have $d(p', q') = 0$ even if $p' \neq q'$ and hence d is not a metric. \square

A.4 Combination-invariance and Proof of Theorem 4

Proposition 8. *Combination-invariance is equivalent to Null-information and Order-invariance.*

Proof of Proposition 8. Consider some d that satisfies Null-information and Order-invariance. By Lemma 3, d is coupled with some H . Consider some q, π_{α} , and π_{β} . Then, applying Lemma 2,

$$\begin{aligned} \mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] &= \mathbb{E}[(H(q) - H(q(\alpha))) + (H(q(\alpha)) - H(q(\alpha \cap \beta)))] \\ &= \mathbb{E}[H(q) - H(q(\alpha \cap \beta))] \\ &= \mathbb{E}[d(q(\alpha \cap \beta), q)]. \end{aligned}$$

Hence, d satisfies Combination-invariance.

Now, consider some d that satisfies Combination-invariance. We first show that d is coupled with $H(q) = \sum_{\omega} q^{\omega} d(\delta_{\omega}, q)$. The proof follows the same logic as the proof of Lemma 3. Fix

some q and some signal π_α . To show that H is coupled with d , we need to show $\mathbb{E}[d(q(\alpha), q)] = \mathbb{E}[H(q) - H(q(\alpha))]$. Let π_β be a fully informative signal. First, consider observing both π_α and π_β . Since π_β is fully informative, $q(\alpha \cap \beta)$ equals δ_ω when the state is ω ; hence, $\mathbb{E}[d(q(\alpha \cap \beta), q)] = \sum_\omega q^\omega d(\delta_\omega, q) = H(q)$. Next, consider observing π_α followed by π_β . This generates expected sum of information equal to $\mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))]$. Moreover, $\mathbb{E}[d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[\sum_\omega q^\omega(\alpha) d(\delta_\omega, q(\alpha))] = \mathbb{E}[H(q(\alpha))]$. So, by Combination-invariance, we have that $H(q) = \mathbb{E}[d(q(\alpha \cap \beta), q)] = \mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] = \mathbb{E}[d(q(\alpha), q) + H(q(\alpha))]$, i.e., $\mathbb{E}[d(q(\alpha), q)] = \mathbb{E}[H(q) - H(q(\alpha))]$. Hence, H is coupled with d and thus by Lemma 4, d satisfies Order-invariance. Finally, to show that d satisfies Null-information, consider any q and π_α and π_β , both of which are completely uninformative. Then, $\mathbb{E}[d(q(\alpha), q) + d(q(\alpha \cap \beta), q(\alpha))] = 2d(q, q)$ and $\mathbb{E}[d(q(\alpha \cap \beta), q)] = d(q, q)$. Hence, Combination-invariance implies $2d(q, q) = d(q, q)$ or $d(q, q) = 0$. \square

We now turn to the proof of Theorem 4. First consider a (seemingly) stronger notion of combination invariance. Say that a measure of information d satisfies *Interim combination-invariance* if for every q , $\alpha \in S$, and π_β ,

$$d(q(\alpha), q) + \mathbb{E}[d(q(\alpha \cap \beta), q(\alpha)) | \alpha] = \mathbb{E}[d(q(\alpha \cap \beta), q) | \alpha].$$

Lemma 6. *A measure of information satisfies Combination-invariance if and only if it satisfies Interim combination-invariance.*

Proof of Lemma 6. It is immediate that Interim combination-invariance implies Combination-invariance. Now suppose d satisfies Combination-invariance. To establish Interim combination-invariance, we first note that Proposition 8 jointly with Lemmas 3 and 5 implies that $d(p, q) = H(q) - H(p) + f(q)(p - q)$ for some measure of uncertainty H and some function $f(q)$. Hence, given some $\alpha \in S$

and some signal π_β , we have

$$\begin{aligned}
& d(q(\alpha), q) + \mathbb{E}[d(q(\alpha \cap \beta), q(\alpha)) | \alpha] \\
&= H(q) - H(q(\alpha)) + f(q)(q(\alpha) - q) + \\
&\quad \mathbb{E}[H(q(\alpha)) - H(q(\alpha \cap \beta)) + f(q(\alpha))(q(\alpha \cap \beta) - q(\alpha)) | \alpha] \\
&= H(q) + f(q)(q(\alpha) - q) + \mathbb{E}[-H(q(\alpha \cap \beta)) | \alpha] \\
&= \mathbb{E}[H(q) - H(q(\alpha \cap \beta)) + f(q)(q(\alpha) - q) | \alpha] \\
&= \mathbb{E}[H(q) - H(q(\alpha \cap \beta)) + f(q)(q(\alpha) - q) + f(q)(q(\alpha \cap \beta) - q(\alpha)) | \alpha] \\
&= \mathbb{E}[H(q) - H(q(\alpha \cap \beta)) + f(q)(q(\alpha \cap \beta) - q) | \alpha] \\
&= \mathbb{E}[d(q(\alpha \cap \beta), q) | \alpha].
\end{aligned}$$

□

Lemma 7. Consider a payment function $t(p, q)$ and a measure of information $d(p, q)$ such that $t(p, q) = d(p, q)$. If t is ex ante incentive compatible, then d satisfies Combination-invariance.

Proof of Lemma 7. Suppose t is an ex ante incentive compatible payment function and $d(p, q) = t(p, q)$. Consider some prior q and a pair of signals, $\pi_{\alpha'}$ and $\pi_{\beta'}$. We need to show

$$\mathbb{E}[t(q(\alpha'), q) + t(q(\alpha' \cap \beta'), q(\alpha'))] = \mathbb{E}[t(q(\alpha' \cap \beta'), q)]. \quad (2)$$

We denote the signal that is equivalent to observing both $\pi_{\alpha'}$ and $\pi_{\beta'}$ by $\pi_{\alpha'} \vee \pi_{\beta'}$.³⁷ Suppose the seller faces exogenous signals π_{α^*} and π_{β^*} with $\pi_{\alpha^*} = \pi_{\alpha'} \vee \pi_{\beta'}$ and π_{β^*} is completely uninformative. If the seller does not delay revelation at the ex ante stage and sets $\pi_\alpha = \pi_{\alpha^*}$, his payoff is $\mathbb{E}[t(q(\alpha^* \cap \beta^*), q)] = \mathbb{E}[t(q(\alpha' \cap \beta'), q)]$. The seller has a possible deviation of setting $\pi_\alpha = \pi_{\alpha'}$ and $\pi_\beta = \pi_{\alpha'} \vee \pi_{\beta'}$, which would give him the payoff $\mathbb{E}[t(q(\alpha'), q) + t(q(\alpha' \cap \beta'), q(\alpha'))]$. Since t is ex ante incentive compatible, we have

$$\mathbb{E}[t(q(\alpha'), q) + t(q(\alpha' \cap \beta'), q(\alpha'))] \leq \mathbb{E}[t(q(\alpha^* \cap \beta^*), q)] = \mathbb{E}[t(q(\alpha' \cap \beta'), q)]. \quad (3)$$

³⁷Since the set of all signals with the refinement order is a lattice, \vee indicates the join operator.

Now, suppose the seller faces exogenous signals $\pi_{\alpha^*} = \pi_{\alpha'}$ and $\pi_{\beta^*} = \pi_{\beta'}$. If the seller does not delay revelation at the ex ante stage, his payoff is $\mathbb{E}[t(q(\alpha'), q) + t(q(\alpha' \cap \beta'), q(\alpha'))]$. The seller has a possible deviation of setting π_{α} as uninformative and $\pi_{\beta} = \pi_{\alpha'} \vee \pi_{\beta'}$, which would give him the payoff $\mathbb{E}[t(q(\alpha' \cap \beta'), q)]$. Since t is ex ante incentive compatible, we have

$$\mathbb{E}[t(q(\alpha'), q) + t(q(\alpha' \cap \beta'), q(\alpha'))] \geq \mathbb{E}[t(q(\alpha' \cap \beta'), q)]. \quad (4)$$

Combining (3) and (4) yields (2). □

We are now ready to prove Lemma 1.

Proof of Lemma 1. Suppose that t is ex ante incentive compatible. By Lemma 7, we know that it satisfies Combination-invariance, so Lemma 6 in turn implies it satisfies Interim combination-invariance. Now, consider any π_{α} and π_{β} and some realization α from π_{α} . If the seller reveals α , his payoff is $t(q(\alpha), q) + \mathbb{E}[t(q(\alpha \cap \beta), q(\alpha)) | \alpha]$ while if he withholds α , his payoff is $t(q, q) + \mathbb{E}[t(q(\alpha \cap \beta), q) | \alpha]$. Hence, Interim combination-invariance and $t(q, q) = 0$ imply that the payoffs from revealing α and withholding it are the same. This shows that t is interim incentive compatible. □

Finally, we turn to the proof of Theorem 4.

Proof of Theorem 4. Suppose t is a valid measure of information. By Theorem 1 and Proposition 8, we know t satisfies Combination-invariance. Hence, it is ex ante incentive compatible and thus, by Lemma 1, it is incentive compatible. Now suppose that we have some t that is incentive compatible. Since we assume $t(q, q) = 0$ and $t(p, q) \geq 0$, by Theorem 1 and Proposition 8, it will suffice to establish t satisfies Combination-invariance. This follows directly from Lemma 7. □

References

- Arrow, Kenneth. 1971. *Essays in the Theory of Risk-Bearing*. Chicago: Markham Publishing Co.
- Augenblick, Ned and Matthew Rabin. 2018. Belief movement, uncertainty reduction, & rational updating. Working paper.
- Babaioff, Moshe, Robert Kleinberg, and Renato Paes Leme. 2012. Optimal mechanisms for selling information. *Proceedings of the 13th ACM Conference on Electronic Commerce*. 92–109.
- Banerjee, Arindam, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. 2005. Clustering with Bregman divergences. *Journal of Machine Learning Research*. 6(10): 1705-1749.
- Bergemann, Dirk and Alessandro Bonatti. 2015. Selling cookies. *American Economic Journal: Microeconomics*. 7(3): 259-294.
- Bergemann, Dirk and Alessandro Bonatti. 2018. Markets for information. *Annual Review of Economics*. forthcoming.
- Blackwell, David. 1951. Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93-102, University of California Press, Berkeley, Calif.
- Börgers, Tilman, Angel Hernando-Veciana, and Daniel Krähmer. 2013. When are signals complements or substitutes? *Journal of Economic Theory*. 148(1): 165-195.
- Bregman, Lev. 1967. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*. 7(3): 200–217.
- Brier, Glenn W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 78: 1-3.
- Brunnermeier, Markus and Jonathan Parker. 2005. Optimal expectations. *American Economic Review*. 95(4): 1092-1118.
- Cabrales, Antonio, Olivier Gossner, and Roberto Serrano. 2013. Entropy and the value of information for investors. *American Economic Review*. 103(1): 360-377.
- Caplin, Andrew and Mark Dean. 2013. The behavioral implications of rational inattention with Shannon entropy. Working paper.

- Caplin, Andrew, Mark Dean, and John Leahy. 2017. Rationally inattentive behavior: characterizing and generalizing Shannon entropy. Working paper.
- Caplin, Andrew and John Leahy. 2001. Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*. 116(1): 55-79.
- Cover, Thomas M. and Joy A. Thomas. 2012. *Elements of information theory*. John Wiley & Sons.
- Cripps, Martin W. 2019. Divisible updating. Working paper.
- Csiszar, Imre. 2008. Axiomatic characterizations of information measures. *Entropy*. 10(3): 261-273.
- De Lara, Michael and Olivier Gossner. 2017. An instrumental approach to the value of information. Working paper.
- Ely, Jeff, Alex Frankel, and Emir Kamenica. 2015. Suspense and surprise. *Journal of Political Economy*. 103(1): 215-260.
- Frankel, Alexander and Maximilian Kasy. 2018. Which findings should be published? Working paper.
- Gentzkow, Matthew and Emir Kamenica. 2014. Costly persuasion. *American Economic Review Papers and Proceedings*. 104(5): 457-462.
- Gentzkow, Matthew and Emir Kamenica. 2017. Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*. 104: 411-429.
- Gneiting, Tilmann and Adrian Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*. 102(447): 359-378.
- Good, I.J. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*. 14: 107-114.
- Gossner, Olivier and Jakub Steiner. 2018. On the cost of misperception: General results and behavioural applications. *Journal of Economic Theory*. 177: 816-847.
- Green, Jerry R. and Nancy L. Stokey. 1978. Two representations of information structures and their comparisons. Working paper.
- Hebert, Benjamin and Michael Woodford. 2017. Rational inattention with sequential information sampling. Working paper.

- Kamenica, Emir and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review*. 101(6): 2590-2615.
- Kelly, J.L. 1956. A new interpretation of information rate. *Bell System Technical Journal*. 35: 917-926.
- Lehmann, Erich L. 1988. Comparing location experiments. *The Annals of Statistics*. 16(2): 521-533
- Mensch, Jeffrey. 2018a. Cardinal representations of information. Working paper.
- Mensch, Jeffrey. 2018b. Bayesian learning by polling. Working paper.
- Morris, Stephen and Philip Strack. 2017. The Wald problem and the equivalence of sequential sampling and static information costs. Working paper.
- Persico, Nicola. 2000. Information acquisition in auctions. *Econometrica*. 68(1): 135-148.
- Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2019. The cost of information. Working paper.
- Schervish, Mark J. 1989. A general method for comparing probability assessors. *Annals of Statistics*. 17(4): 1856-1879.
- Sims, Christopher. 2003. Implications of rational inattention. *Journal of Monetary Economics*. 50(3): 665-690.
- Steiner, Jakub, Colin Stewart, and Filip Matejka. 2017. Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*. 85(2): 521-553.
- Yoder, Nathan. 2016. Designing incentives for academic research. Working paper.
- Zanardo, Enrico. 2017. How to measure disagreement. Working paper.