CONSUMER THEORY
FOR CHEAP INFORMATION

Gary Baker

This version: 19 March 2021

[PRELIMINARY AND INCOMPLETE – PLEASE DO NOT CIRCULATE]

LINK TO THE MOST RECENT VERSION

ABSTRACT

In many settings, a decision-maker facing uncertainty must decide not only *how much* information to purchase, but also *from which* sources. It is already known that when uncertainty is about a binary variable a decision-maker with a sufficiently large budget will spend their entire budget on signals from a single source: the most precise one. When there are more than two possible states of the world, however, the story is different: different information sources may differ in the pair of states they are worst at distinguishing, and the decision-maker may therefore purchase information from different sources which cover for each others' weaknesses. I characterize tradeoffs between samples from different sources in a setting with low costs (or large budgets), where the probability of a mistake is small and well-described by large deviations theory. I show that, for large samples, the marginal rate of substitution between samples from distinct information sources is approximately given by the ratio of precision-like indices for each source, and that the asymptotically optimal bundle satisfies a maxi-min rule: it maximizes the precision per dollar for the worst case pair of states. Since the precision of each signal does not depend on the characteristics of the decision-maker (their prior beliefs or the payoffs of the decision problem they face), all decision-makers agree on the fraction of signals from each source in an asymptotically optimal bundle. Different information sources with the same worst-case pair of states—the two states the signal is least precisely able to differentiate—are redundant and are never demanded together. The asymptotically optimal bundle occurs either at a corner, or at one of a finite number of interior "kink points" where the worst-case pair of states switches. To illustrate these results, I consider a number of basic consumer theory exercises and discuss implications for information demand.

1 INTRODUCTION

Often a decision-maker may find herself in a position to acquire information prior to making a decision under an uncertain state of the world, and many cases, she must not only decide how much information to purchase, but also from where to acquire it. For example, a reader of the news must every day decide both how long and from which sites to read. Or a researcher studying the effects of a new drug on a disease might have multiple available tests of differing cost for the disease.

In these scenarios, how might one determine the optimal bundle of information? Are corner solutions always optimal? And if not, when?

Answering these questions ought to be little more than a principles exercise of setting the marginal rate of substitution (MRS) equal to the price ratio along a budget constraint. Unfortunately, such an exercise presupposes understanding of a utility function—no mean feat when the value for information is notoriously ill-behaved.[1]

I provide an answer to these questions for environments with finitely many possible underlying states, valid when information is sufficiently cheap (or budgets sufficiently large), by precisely characterizing asymptotic tradeoffs between conditionally independent samples from different information sources (Blackwell experiments) using methods from the theory of large deviations.

In particular, I develop a precision-like measure (the Chernoff precision) that quantifies the degree to which an information source can distinguish between a given pair of underlying states. I then show that, for large samples, indifference curves[2] are loci of sample bundles with equal precision of their worst-case dichotomy (the state pair for which the bundle is least precise). Furthermore, for any given state pair, the precision of a sample bundle is the sum of component precisions from each constituent information source, and the component precision of a given source is always weakly lower when an information source is a part of a heterogeneous bundle than it is by itself. Put more simply: for a single pair of states, composite information sources are weakly worse that the sum of their parts.

In a two-state environment, these results imply indifference curves eventually bow outward, and that corner solutions are thus optimal at low costs. In richer environments, however, because only the least-precision dichotomy matters for large samples, a source may complement another by covering for the other's weaknesses. In this case, indifference curves are outer contours of multiple outward bowing curves (one for each dichotomy), and thus inward-pointing kinks occur whenever the worst-case pair of states changes. The optimal bundle must then eventually lie near either a corner, or one of finitely many interior kink points. Fig. 1 stylistically illustrates this result graphically for an example environment with three states (and thus three dichotomies).

1 Most famously, value of information is often non-concave, and thus marginal value of information is often rising. See, for example (Radner & Stiglitz 1984) or Chade & Schlee (2002).

2 Samples are a fundamentally discrete quantity, so non-singleton indifference sets typically don't exist. I will be more formal later.

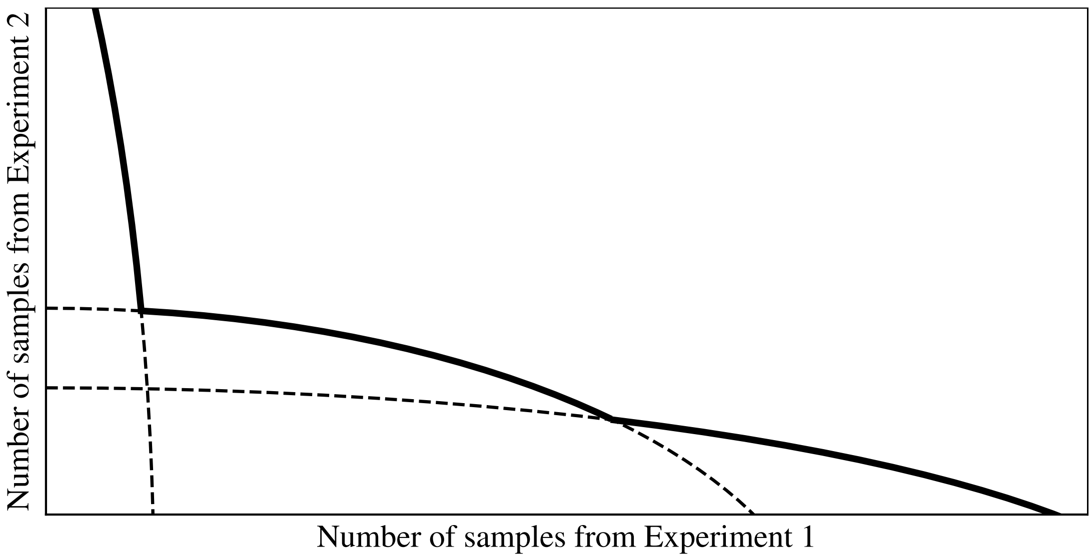

Figure 1: The geometry of iso-precision lines in a three-state environment with two available information sources. Iso-precision lines for a fixed dichotomy (dashed lines) bow out, but the iso-least-precision line (solid line) will have inward pointing kinks.

I then apply my formulae to derive asymptotically precise expressions for information demand given a budget constraint and cost vector. These results have certain testable consumer theoretic implications. For example, because interior solutions always lie near an indifference curve kink, for small price changes, information sources behave as perfect Marshallian complements with fixed bundle proportions.

Except near finitely many discontinuities when the optimum jumps between kinks/corners, demand curves are inelastic, with finitely many small cost intervals of high elasticity when jumping between kinks/corners.

Furthermore, component precisions depend only on the relative proportions of each source in a bundle, so information value is approximately homothetic for large samples. Thus all information sources are normal goods—and more specifically, they have unit income elasticities—for low enough prices or high enough budgets.

Importantly, because the Chernoff precision only depends on characteristics of the information sources, and not on decision-maker specifics such as prior or payoff structure, (generically) all decision makers agree on tradeoffs between information sources at low enough price, and for a given price vector, all agree of the relative proportions of each information source in an asymptotically optimal bundle.

My results fit primarily into two different strains of the literature: one economic and the other statistical.

First, this paper contributes to the statistical literature on asymptotic relative efficiency (ARE) which traditionally asks how many samples, n_1, from one test are required to perform as well as n_2 from another for large samples. In particular, my results build on those of Chernoff (1952) who showed that, in a two-state/two-action environment (e.g., null/alternative hypothesis testing), the ratio, n_2/n_1, of two equally performing sample sizes is given by the ratio of each test's respective Chernoff precision β_1/β_2. My results expand upon this in two ways: first, I generalize the problem to tradeoffs between arbitrary finite-action/finite-state decision scenarios, and second, I consider local tradeoffs between interior bundles, rather than just the extreme tradeoffs between corner bundles.

Second, this paper sits in the broader information economics literature, especially that on information demand. Of particular relevance—and crucial for my results—are the results of Moscarini & Smith (2002), henceforth MS02, who characterized the asymptotic marginal value of information using methods from the theory of large deviations to show that, in a two-state world, the probability of a mistake is falling exponentially, and that in a many-state environment, the only most likely mistake (that is, the state pair with the least precision) matters for large samples. They then use these results to calculate a particularly accurate approximation for information demand at low prices. Their demand approximation implies demand is always inelastic and thus linear pricing cannot be optimal for an information monopolist. I generalize these results to a world with multiple available information sources and show there is a rich consumer theory arising from the exercise.

Finally, the Chernoff precision fits in a smaller literature (somewhat adjacent to the information design literature) on quantifying the information content of signals. Pomatto et al. (2018) characterize a unique class of "linear" information measures based on weighted sums of the information source's various relative entropies of its conditional distributions. The linear nature of their measure lends itself to capturing costs of information *production*. By contrast, I argue that the Chernoff precision is more useful on the *consumption* side of the equation, characterizing approximate tradeoffs between different information sources.

These results lend themselves to multiple applications. First and foremost, they provide a general approach for analysis of information markets. In particular, they imply that information can be treated with a fairly standard consumer theory toolkit and suggest particular functional forms for information values and demand that may prove useful for modeling purposes in certain applied settings.

Additionally, this paper provides a prescriptive recommendation for optimal experiment design that accounts not only for statistical properties of the various design options but also their costs. Although the finite-state assumption limits the immediate applicability of my results in most experimental contexts, the approach provides a basis for developing an economically-motivated design criterion.[3]

The remainder of the paper is structured as follows: Section 2 explains the model structure and assumptions. Section 3 introduces the relevant large deviations tools and develops the main approximation results, and Section 5 uses these results to build a consumer theory for information, and considers some implications for information demand. Section 6 explores the performance of the approximations both theoretically and numerically. Finally, Section 7 concludes and discusses areas for future research.

3 Experiment design is a large and varied literature within statistics going back to Fisher (and arguably Gauss). For a textbook treatment, see Montgomery (2012). I consider the experiment design implications more carefully in a separate working paper that generalizes these results to certain regression settings.

A decision-maker (DM) must choose an action a from a finite set, A, under an uncertain decision-relevant state of the world θ drawn from a finite set, Θ. The DM has a full-support prior, $p \in \Delta\Theta$, and state-dependent (Bernoulli) utility function, $u(a, \theta)$. To avoid trivialities, assume no action is dominated, and for expositional simplicity assume that each state has a unique optimal action (so necessarily $|A| \geq |\Theta|$).

Denote a particular information source—formally, a *Blackwell experiment*—as $\mathscr{E} = \langle F(\cdot \,|\, \theta), \theta \in \Theta \rangle$. That is, each signal is a collection of state-dependent distributions over some arbitrary space of outcomes (or *realizations*), R. Purely for the sake of simplifying exposition, I will assume $R = \mathbb{R}$ and that each state-dependent distribution has a density, $f(\cdot \,|\, \theta)$, but the results and proofs allow arbitrary realization space.

To avoid degenerate outcomes, assume all experiments are composed of distinct distributions—so signals are informative—and that those distributions are mutually absolutely continuous—so no signal realization completely rules in or rules out any strict subset of the states. Under the density assumption, this implies that for any state pair, $\theta\theta'$, and for all realizations (except perhaps a measure zero set) the likelihood ratio, $f(r \,|\, \theta')/f(r \,|\, \theta)$, is non-zero and finite.

The DM has access to a finite menu of such information sources, $\mathscr{E}_1, \ldots, \mathscr{E}_J$, from each of which she may purchase an arbitrary number of conditionally independent samples, $n = [n_1, \ldots, n_J]$ at costs $c = [c_1, \ldots, c_J]$ per sample. The DM then observes a realization from the joint signal $\bigotimes_{j=1}^{J} \mathscr{E}_j^{n_j}$, where $\mathscr{E}^n$ denotes the compound signal generated by n i.i.d. samples from $\mathscr{E}$, and $\mathscr{E}_1 \otimes \mathscr{E}_2$ is the experiment generated by one sample from each $\mathscr{E}_1$ and $\mathscr{E}_2$. The DM has a large but finite budget, Y, to spend on information.

After observing the realization, r, of the chosen signal bundle, the DM updates her beliefs according to Bayes's rule and the chooses an action to maximize her expected payoff. Putting this all together, we can define the value with information: the expected payoffs from acting after observing the signal realization.[4] That is,

$$v(n) = \int_{r \in R} \max_a \left\{ \sum_{\theta \in \Theta} u(a, \theta) \mathbb{P}(\theta \mid r) f_n(r \,|\, \theta) \right\} dr$$

where f_n is the conditional density of signal realizations from the chosen sample bundle.

The DM chooses n to maximize this value, subject to the usual budget constraint, $n \cdot c \leq Y$.

Bayes's rule for probabilities is somewhat tricky to work with so we'll typically work with log-likelihood ratios (LLRs) where Bayes's rule becomes a sum:

$$\log\left(\frac{\mathbb{P}(\theta' \,|\, r)}{\mathbb{P}(\theta \,|\, r)}\right) = \log\left(\frac{p_{\theta'}}{p_\theta}\right) + \log\left(\frac{f(r \,|\, \theta')}{f(r \,|\, \theta)}\right)$$

4 Not to be confused with the value *of* information, which measures the expected *gain* in payoff from acting after observing the realization. Both values are ordinally equivalent (they differ only by a constant), so I use the value with information to simplify notation.

We'll thus need some assumptions on the state-dependent distributions of LLRs for each signal:

First, for each signal, $\mathscr{E}_j$, for any pair of states, θ and θ', denote the moment generating function (MGF) of the LLR distribution in the denominator state—that is, the distribution of $\log(f_j(r\,|\,\theta')/f_j(r\,|\,\theta))$ in state θ—as

$$\begin{aligned} \mathrm{M}_i(t\,;\theta\theta') &\equiv \int_{r\in\mathrm{R}} e^{t\log\left(\frac{f_j(r\,|\,\theta')}{f_j(r\,|\,\theta)}\right)} f_j(r\,|\,\theta)\mathrm{d}r \\ &= \int_{r\in\mathrm{R}} f_j(r\,|\,\theta)^t f_j(r\,|\,\theta')^{1-t}\mathrm{d}r \end{aligned} \tag{1}$$

Assume that for any pair of states, this MGF is defined within an open interval containing zero, and thus that the LLR distribution has all its moments.[5] Finally, to simplify the main proof, assume that for any pair of states and for any experiment, that the log-likelihood ratios of possible signal realizations do not form a lattice on the reals—that is, for $b, k \in \mathbb{R}$

$$\left\{\log\left(\frac{\mathrm{dF}_j(r\,|\,\theta')}{\mathrm{dF}_j(r\,|\,\theta)}\right) \,:\, r\in\mathrm{R}\right\} \not\subseteq \{b+kc \,:\, k\in\mathbb{Z}\}$$

In other words, the support of the log-likelihood ratio distribution does not exclusively consist of equally spaced points.

3 BACKGROUND

3.1 Large deviations

For the sake of illustrating the relevant tools, first consider the two-state/two-action world. I pose the problem as a classic statistical dichotomy: there are two states, $\Theta = \{\mathrm{H}_0, \mathrm{H}_1\}$, representing a null hypothesis and alternative hypothesis, respectively. The DM must choose whether to *reject*, $a = \mathscr{R}$, or *accept*, $a = \mathscr{A}$, the null hypothesis. Let $u(\mathscr{A}, \mathrm{H}_0) > u(\mathscr{R}, \mathrm{H}_0)$ and $u(\mathscr{R}, \mathrm{H}_1) > u(\mathscr{A}, \mathrm{H}_1)$, so rejecting the null is preferred when the alternative is true. Denote the prior probability that the alternative is true as p. Finally denote the cut-off belief—i.e. the one where the DM is exactly indifferent between the two actions—as $\bar{p}$.

For now, in order to introduce the necessary technical concepts, I review the single-source case examined by MS02. In this case, we can write the value after receiving n samples as

$$\begin{aligned} v(n) = (1-p)\big(&\alpha_{\mathrm{I}}(n)\,u(\mathscr{R},\mathrm{H}_0) + (1-\alpha_{\mathrm{I}}(n))\,u(\mathscr{A},\mathrm{H}_0)\big) \\ -p\,\big(&\alpha_{\mathrm{II}}(n)\,u(\mathscr{A},\mathrm{H}_1) + (1-\alpha_{\mathrm{II}}(n))\,u(\mathscr{R},\mathrm{H}_1)\big) \end{aligned} \tag{2}$$

5 applying dominated convergence to Eq. 1 shows that the MGF is finite for $t \in [0,1]$ for mutually absolutely continuous distributions.

where $\alpha_{\mathrm{I}}(n)$ and $\alpha_{\mathrm{II}}(n)$ denote the respective Type-I and Type-II error probabilities with the given number of samples from each experiment.

For the sake of algebraic simplicity, it will usually be convenient to work with the *full-information gap* (FIG): the difference between the value of a perfect information source, $\bar{v} = pu(\mathscr{R}, \mathrm{H}_1) + (1-p)u(\mathscr{A}, \mathrm{H}_0)$, and the value of the given information purchases.

$$\begin{aligned} \mathrm{FIG}(n) = (1-p)\,\alpha_{\mathrm{I}}(n))\,(u(\mathscr{A}, \mathrm{H}_0) - u(\mathscr{R}, \mathrm{H}_0)) \\ + p\,\alpha_{\mathrm{II}}(n))\,(u(\mathscr{R}, \mathrm{H}_1) - u(\mathscr{A}, \mathrm{H}_1)) \end{aligned} \tag{3}$$

Notice that the FIG goes to zero with the error probabilities, and that constrained *minimization* of the FIG is equivalent to constrained *maximization* of the value. Equally, the FIG can be seen as the expected loss for the loss function $\mathrm{L}(a, \theta) = u(a^*(\theta), \theta) - u(a, \theta)$.

We can now characterize the asymptotic behavior of α_{I} and α_{II} using a large-deviations approach:

First, as previously discussed, Bayes's rule is a sum for log-likelihood ratios, so we can write

$$\alpha_{\mathrm{I}}(n) = \mathbb{P}\left(l + \sum_{k=1}^{n} s_k > \bar{l} \,\middle|\, \theta = \mathrm{H}_0 \right)$$

where $l \equiv \log(p/(1-p))$ is the prior log-likelihood ratio, $\bar{l} \equiv \log(\bar{p}/(1-\bar{p}))$ is the indifference belief log-likelihood ratio, and $s_k \equiv \log(\mathrm{dF}(r_k \,|\, \mathrm{H}_0)/\mathrm{dF}(r_k \,|\, \mathrm{H}_1))$ is the log-likelihood ratio of the realization from the k-th sample.

Because $\mathbb{E}(s_k \,|\, \theta = \mathrm{H}_0) < 0$, the posterior log-likelihood will stochastically drift towards negative infinity (certainty that the truth is H_0). Asymptotic approximation of the mistake probability thus falls into the realm of *large deviations theory* which considers the distribution of sample means far from the true mean.[6]

To attack such a problem, Cramér (1938) canonically proved that for a random variable X with negative expectation, the probability that the sum of n i.i.d. draws from X is positive is falling roughly exponentially with rate given by the minimized value of the MGF of X. Chernoff (1952) further developed this result to show that the ratio of two equally performing sample sizes from two statistical ratio tests on a binary state is given by the ratio of log minimized MGFs.[7]

Motivated by this approach, define $\rho \equiv \min_{t \in [0,1]} \mathrm{M}(t\,; \mathrm{H}_1, \mathrm{H}_0)$ as the *Chernoff number*[8] of the information source and τ as the minimizer. Because MGFs are always convex, this

6 In contrast to *small deviations* central limit theorems which consider the distribution of a sample mean near the true mean.

7 I grossly simplify Cramér's and Chernoff's by only considering implications for log-likelihood ratio tests. Both results are considerably more general than this. For a more thorough coverage of large deviations methods, see Dembo & Zeitouni (1998).

8 I follow Torgersen (1991) here. Moscarini & Smith (2002) call this the *efficiency index*.

minimum must be unique, and, because the MGF is 1 when $t \in \{0,1\}$ (see Eq. 1), we must have $\tau \in (0,1)$.

Note that $\mathrm{M}(t\,;\mathrm{H}_1,\mathrm{H}_0) = \mathrm{M}(1-t\,;\mathrm{H}_0,\mathrm{H}_1)$ so the minimized value is the same for both log-likelihood ratios and thus each test has a single Chernoff number in the two-state world. In the two-state world, I thus suppress the state-pair argument: $\mathrm{M}(t) \equiv \mathrm{M}(t;\mathrm{H}_1,\mathrm{H}_0)$.

The Chernoff number arises from a fairly abstract derivation so it's worth examining a transformation of it for the sake of intuition. We can consider $\beta \equiv -\log(\rho)$ a measure of a test's *precision.* For example, consider a (homoskedastic) Gaussian signal that is distributed $\mathcal{N}(0,1/\gamma)$ when $\theta = \mathrm{H}_0$ and $\mathcal{N}(1,1/\gamma)$ otherwise. In this case, γ is the signal's precision in the classical sense, and direct computation reveals that for this experiment we have $\beta = (1/8)\gamma$. In reference to this example, I will call β_j the *Chernoff precision* of the test. The following result establishes a number of other properties one might expect of a measure with such a name:

Lemma 1 (Properties of Chernoff precisions). *In an environment with two states, for any information source with mutually absolutely continuous state-dependent distributions, the Chernoff precision has the following properties:*

1. $\rho \in (0,1)$, *and thus* $\beta > 0$;
2. *A test composed of* k *i.i.d. samples from* $\mathscr{E}$ *will have Chernoff precision* $k\beta$; *and,*
3. *If* $\mathscr{E}_1 \succsim \mathscr{E}_2$ *in the Blackwell sense, then* $\beta_1 \geq \beta_2$.

Proof. From Eq. 1, it is immediate that the MGF takes a value of 1 at the corners of the unit interval, and is strictly between 0 and 1 on the interior. The first property then follows. The second follows because the log-likelihood ratio MGF of k samples is M^k because the MGF of an i.i.d. sum is the product of MGFs. The last property follows from (Thm. 12, Blackwell 1951) which states that composites of Blackwell-dominant experiments dominate composites of dominated ones. Thus, Chernoff's result implies that dominant experiments must have lower Chernoff numbers and thus higher Chernoff precisions. □

With these tools we can now state a version of MS02's main result:

Lemma 2 (MS02, Thm. 1). *The probability of a type-I error is proportional to*[9]

$$\alpha_{\mathrm{I}}(n) \propto \frac{\rho^n}{\sqrt{n}}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \tag{4}$$

9 O is the usual big-O Landau notation: a function $g(n)$ is $O(f(n))$, if there exists some finite constant, such that $|g(n)| \leq |\mathrm{C}f(n)|$ for all n large enough. I will additionally use the little-o notation: $g(n)$ is $o(f(n))$ if the previous inequality holds *for any* $\mathrm{C} > 0$ for large enough n, or, equivalently, that $g(n)/f(n)$ goes to 0 as n grows large.

The type-II error probability is the same, up to the proportionality constant. For the appropriate constant,[10] B, *the FIG is thus given by*

$$\mathrm{FIG}(n) = \mathrm{B}\frac{\rho^n}{\sqrt{n}}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \tag{5}$$

To prove this, MS02 use a classic approach similar to Cramér (1938) and Chernoff (1952), namely, an exponential change of measure and Edgeworth series approximation. In this context, that method yields a conservative $O(n^{-1/2})$ bound on the error. Later, I will state a version of this result with sharper bounds on the error using a saddlepoint approach.

From this result we can immediately derive the Chernoff's asymptotic relative efficiency for this setting:

Corollary 1 (Chernoff ARE). *If $\mathscr{E}_1$ and $\mathscr{E}_2$ are two information sources, then, at large samples, the minimum number of independent samples n_2 from $\mathscr{E}_2$ required to perform at least as well as n_1 from $\mathscr{E}_1$ satisfies*

$$\frac{n_2}{n_1} \simeq \frac{\beta_1}{\beta_2}$$

where β_j are the sources' respective Chernoff precisions.

Proof sketch. Take the log of the expression for FIG given by Eq. 5 for both information sources and choose n_1, n_2 to set them equal.

Note that these results do not depend on the binary actions I've considered here. Lem. 2 gives the asymptotic probability that the LLR sum exceeds some finite cutoff, so by taking differences, one can approximate the probability that the LLR falls in any interval whose lower bound is not $-\infty$. Thus any decision problem with finitely many actions will have a FIG that falls exponentially with the Chernoff number of the information source.

Notice that prior and payoff structure don't enter into this approximation (except through the proportionality constant). Given two information sources, for large enough samples any two DMs[11] eventually agree on which is superior at large samples!

3.2 The many state case

Thus far I've relied heavily on LLRs to turn Bayes's rule into a sum and apply large deviations. However, with more than two states, we can no longer express beliefs in terms of a single LLR; we need an LLR for every state relative to some base state. Furthermore,

10 MS02 give a precise form for the proportionality constant. The interested reader can find the precise form for the constant in Appendix A.??, but it doesn't affect my results, so I omit it from the discussion.

11 There is a subtlety here. For large enough samples, any two DM's agree, but there may not be any sample size such that *all* DM's agree. Azrieli (2014) calls the latter, stronger property *eventual Blackwell sufficiency* and provides a sufficient condition for it.

the mistake probabilities become the probability that this vector of LLRs (whose entries are non-independent) falls in some non-interval subset.

Although at first glance, it seems we would need a non-trivial extension to Lem. 2, we can get some simplification out of the exponentially falling nature of the mistake probability.

First, define $a^*(\theta)$ as the optimal action in state θ, and $\alpha(n\,;a,\theta)$ as the probability of mistakenly choosing suboptimal action a in state θ. Then write the FIG as

$$\mathrm{FIG}(n) = \sum_{\theta} p_\theta \sum_{a \neq a(\theta)} \alpha(n\,;a,\theta)(u(a^*(\theta),\theta) - u(a,\theta))$$

So, the FIG generalizes to a similar form: the ex-ante expectation over all possible mistakes. Thus as before, approximating the FIG is a matter of approximating each error probability.

Now, for each dichotomy, $\theta\theta'$, we can define a MGF for the LLR of the signal realizations:

$$\mathrm{M}(t;\theta\theta') = \int f(r\,|\,\theta)^t f(r\,|\,\theta')^{1-t}\mathrm{d}r$$

and thus for each pair of states, we can define a Chernoff number, $\rho(\theta\theta')$, and precision, $\beta(\theta\theta')$.

Based on the two-state results, we might guess that each dichotomy FIG is roughly proportional to a term falling exponentially in that dichotomy's Chernoff number, and thus asymptotically all dichotomies except the one with the highest Chernoff number (lowest precision) are negligible for large samples. MS02's Thm. 4 makes this intuition rigorous:

Lemma 3 (MS02, Thm. 4). *Suppose there is a unique least-precision dichotomy,* $\theta\theta'$*, (generically true). Then for some* $\bar{\rho} < 1$*, we can write*

$$\mathrm{FIG}(n) = (p_\theta + p_{\theta'})\mathrm{FIG}^*_{\theta\theta'}(n)(1 + O(\bar{\rho}^n))$$

where $\mathrm{FIG}^*_{\theta\theta'}(n)$ *is the FIG from observing n indpendent samples from the information source when the state is known to be either* θ *or* θ'*.*

Proof sketch. First, clearly $\mathrm{FIG}(n) \geq (p_\theta + p_{\theta'})\mathrm{FIG}^*_{\theta\theta'}(n)$ because the right-hand side is the FIG from observing n samples *plus* a signal that perfectly the reveals the state provided the state is neither θ nor θ' (more information implies lower FIG). The remainder of the proof consists of showing that each state-conditional mistake probability $\alpha(n;a,\theta)$ is $O(\max_{\theta'\neq\theta} \rho(\theta\theta')^n)$, heuristically because each mistake probability requires a large deviation in at least one of the LLRs, and the most likely large deviation is the one for the dichotomy with the least precision. See MS02 for the full proof. □

Using this result plus Lem. 2, with appropriate choice of constant, B, we can rewrite

$$\mathrm{FIG}(n) = \mathrm{B}\frac{\rho(\theta\theta')^n}{\sqrt{n}}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

where $\theta\theta'$ is the least-precision dichotomy.

Simply put, for large samples, an information sources asymptotic performance is entirely given by the pair of states which is most struggles to tell apart.

Thus, all our conclusions from the two-state case still hold: because the least-precision dichotomy is only a function of signal properties, all DMs eventually agree on which of two signals is optimal at large samples, and the Chernoff ARE is still given by the ratio of precisions—though now the ratio of each source's respective worst-case dichotomy precision.

Now if all we cared about was the ability to compare corner strategies, we could stop here—Chernoff's ARE tells us everything we need; however, Chernoff tells us nothing about tradeoffs between tests at interior bundles. In order to get a complete picture, we need to extend the theory to cover composite information sources.

4 RESULTS

4.1 Composite experiments

First, consider the two-state/two-action case again. Suppose our DM has access to a menu of various conditionally independent information sources, $\mathscr{E}_1, \dots, \mathscr{E}_J$. If she purchases a bundle $\mathrm{n} = [n_1, \dots, n_J]$ of samples, then the MGF of the LLR distribution of this composite experiment will simply be

$$\mathrm{M}_{\mathrm{n}}(t) = \prod_{j=1}^{J} \mathrm{M}_j(t)^{n_j}$$

because the MGF of a sum of independent random variables is the product of the constituent MGFs. Define the *composite factor*, $\omega \equiv [n_1/\mathrm{N}, \dots, n_J/\mathrm{N}]$ where $\mathrm{N} \equiv \sum n_j$ is the *total* sample size. We can then treat the composite signal as N samples from a single information source with LLR MGF given by $\mathrm{M}_\omega(t) \equiv \prod \mathrm{M}_j(t)^{\omega_j}$ and write the composite MGF as[12]

$$\mathrm{M}_{\mathrm{n}}(t) = \mathrm{M}_\omega(t)^{\mathrm{N}}$$

Not surprisingly given the previous section, we will be minimizing this object: define $\rho_\omega \equiv \min_t \mathrm{M}_\omega(t)$ and $\beta_\omega \equiv -\log \rho_\omega$ as the *composite* Chernoff number and precision respectively, and $\tau_\omega \equiv \arg\min_t \mathrm{M}_\omega(t)$.

The composite Chernoff number and precision can then be broken up by contribution from each information source by writing $\rho_{\omega i} \equiv \mathrm{M}_j(\tau_\omega)$ and $\beta_{\omega j} \equiv -\log \rho_{\omega j}$. Call these the ω-component Chernoff numbers and precisions respectively. So we can then write the total

12 Note that M_ω will not generally be a valid MGF, but for our purposes it will suffice to treat it as such.

Chernoff number and precision from an n bundle as

$$\rho_\omega^{\mathrm{N}} = \prod_{j=1}^{\mathrm{J}} \rho_{\omega j}^{n_j}$$

$$\mathrm{N}\beta_\omega = \sum_{j=1}^{\mathrm{J}} n_j \beta_{\omega j}$$

We can now state an augmented version of Lem. 2:

Proposition 1. *For a two-state/two-action decision problem, the probability of a type-I error given a bundle of* $\mathrm{n} = [n_1, \ldots, n_{\mathrm{J}}]$ *conditionally independent samples from each information source is proportional to*

$$\alpha_{\mathrm{I}}(\mathrm{n}) \propto \frac{\prod_{j=1}^{\mathrm{J}} \rho_{\omega j}^{n_j}}{\sqrt{\mathrm{N}}} \left(1 + \mathcal{O}\left(\frac{1}{\mathrm{N}}\right)\right) \tag{6}$$

where N *is the sum of samples from each source. The type-II error is the same up to the proportionality constant. Merging terms into* B_ω*, the FIG can thus be written*

$$\mathrm{FIG}(\mathrm{n}) = \mathrm{B}_\omega \frac{\prod_{j=1}^{\mathrm{J}} \rho_{\omega j}^{n_j}}{\sqrt{\mathrm{N}}} \left(1 + \mathcal{O}\left(\frac{1}{\mathrm{N}}\right)\right) \tag{7}$$

Proof sketch. The result follows from a direct application of a *saddlepoint approximation* due to Lugannani & Rice (1980), building on a method due to Daniels (1954). Roughly, the idea is that, because the MGF is analytic, it's minimum along the real axis corresponds to a maximum along the perpindicular imaginary axis, and thus, the integrand of the MGF (characteristic function) inversion formula becomes almost entirely determined by that maximum at high sample sizes. The full proof is a bit technical and thus relegated to the appendix.[13]

Besides the generalization to a multiple-source setting, Prop. 1 improves on Lem. 2 in two ways: first, it tightens the error bound to $O(\mathrm{N}^{-1})$, and, second, the proof approach implies an asymptotic expansion that accounts for each source's contribution to the approximation error, and may be numerically useful. Because the expansion doesn't directly affect my results, I leave further discussion of it to the Appendix A.??.

We're primarily interested in *ordinal* properties of the information value, so we can simplify things further by taking a monotone transformation:

$$-\log \mathrm{FIG}(\mathrm{n}) \;=\; \sum_{j=1}^{\mathrm{J}} n_j \beta_{\omega j} \left(1 + O\left(\frac{\log(\mathrm{N})}{\mathrm{N}}\right)\right) \tag{8}$$

13 Butler (2017) gives a relatively non-technical treatment of saddlepoint methods.

Then we have that constrained maximization of the above is equivalent to minimization of the FIG, and thus equivalent to maximization of the value, and thus that large sample bundles of equal precision, must have roughly equal information value for any decision maker.

From this it seems like tests must eventually be perfectly substitutable. But it's worse than that because we minimized the product of all the MGFs, rather than each of them individually, so we have

$$\begin{aligned}
\min_t\{M_\omega(t)\} &\geq \prod_{j=1}^{J} \min_t\{M_j(t)^{\omega_j}\}\} \\
\Leftrightarrow \qquad \beta_\omega &\leq \sum_{j=1}^{J} \omega_j \beta_j
\end{aligned} \tag{9}$$

with equality holding only for bundles of information sources all sharing the same minimizer. So, at least when it comes to distinguishing a single pair of states, all composite experiments are weakly worse than the sum of their parts in a large sample setting.

Given this, the optimal bundle for a constrained decision-maker follows immediately:

Corollary 2 (Corner bundles are (near) optimal for dichotomies). *Suppose a budget-constrained decision-maker has a collection of available information sources, $\mathscr{E}_1, \dots, \mathscr{E}_J$, from each of which she may purchase an arbitrary number of samples at per-sample costs $c_1, \dots, c_J$ respectively. Suppose there is a unique source with the highest precision per dollar (true for generic costs), and without loss assume it is $\mathscr{E}_1$. Then for any two-state decision problem, the composite factor of the optimal bundle approaches $\omega_1 = 1$ as the budget goes to infinity.*

Proof. From Eq. 8, it suffices to show that the asymptotic payoff-per-dollar of the non-composite experiment is higher than that of any composite experiment. Ignoring lower-order terms, we want

$$\frac{\beta_1}{c_1} \geq \frac{\omega\beta_1(\omega) + (1-\omega)\beta_2(\omega)}{\omega c_1 + (1-\omega)c_2}$$

but this follows immediately from inequality 9. □

Note that this is not quite the same as corner bundles being precisely optimal. For example, the discrete nature of the problem often makes it optimal to consume some of a cheaper signal if only to make better use of the full budget[14] Nonetheless, because β_ω approaches the Chernoff precision of the non-composite experiment as ω approaches 0 or 1, if an interior bundle is optimal for large samples, it must have a composite factor arbitrarily close to the optimal corner.

14 e.g. if $\mathscr{E}_1$ has a price of \$2 per sample and $\mathscr{E}_2$ a cost of \$1 per sample, it will often be optimal to purchase 1 sample from $\mathscr{E}_2$ simply to avoid leftover budget.

Now, if we only cared about two-state decision problems, all of this effort would have been of little use: corners (or at least near corners) are always optimal for large samples, so Chernoff's original result would have been all we needed. But things get more interesting when we move to the many state world.

Recall from Section 3 that now an experiment is characterized by its collection of LLR MGFs, one for each state pair. We thus then have a Chernoff index/precision for each state pair: denote, $\rho_\omega(\theta\theta') \equiv \min M_\omega(t;\theta\theta')$ and $\beta_\omega(\theta\theta') \equiv -\log \rho_\omega(\theta\theta')$, and the ω-component parts, $\rho_{\omega j}(\theta\theta')$ and $\beta_{\omega j}(\theta\theta')$, similarly.

Then, using Lem. 3 (MS02's Thm. 4) we can generalize Prop. 1 to the many state case:

Proposition 2. *Assume that for all but a measure-zero set of composite factors, the least-precision dichotomy,*

$$D_\omega \equiv \underset{\theta,\theta'\ s.t.\ \theta\neq\theta'}{\arg\min}\ \max_t \left\{ -\sum_{j=1}^{J} \omega_j \log M_j(t;\theta\theta') \right\} \equiv \underset{\theta,\theta'\ s.t.\ \theta\neq\theta'}{\arg\min} \sum_{j=1}^{J} \omega_j \beta_{\omega j}(\theta\theta')$$

is unique (generically true). Then for a finite-state/finite-action decision problem, the FIG from $n = [n_1, \ldots, n_J]$ *conditionally independent samples from each information source is given by*

$$\mathrm{FIG}(n) = B_\omega \frac{\prod_{j=1}^{J} \rho_{\omega j}(D_\omega)^{n_j}}{\sqrt{N}} \left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right)$$

where B_ω *is a uniformly bounded constant that depends only on decision-maker characteristics (prior and payoff structure) and the minimizer of the composite MGF for the least-precise dichotomy.*

Proof. Follows immediately from plugging the expression for the two-state FIG given by Prop. 1 into the many-state FIG equation given by Lem. 3. □

As before, we can monotonically transform this for simplicity letting D_ω continue to be the least-precision dichotomy:

$$-\log \mathrm{FIG}(n) \;=\; \sum_{j=1}^{J} n_j \beta_{\omega j}(D_\omega)\left(1 + O\left(\frac{\log(N)}{N}\right)\right) \tag{10}$$

But unlike before, we can see the potential for complementarities between sources because

$$\min_{\theta,\theta' s.t. \theta\neq\theta'} \max_t \left\{ -\sum_{j=1}^{J} \omega_j \log M_j(t;\theta\theta') \right\}$$

is not necessarily smaller than

$$\sum_{j=1}^{J} \min_{\theta,\theta' s.t. \theta\neq\theta'} \max_{t}\{-\log M_j(t;\theta\theta')\}$$

Roughly, because the value of information for large samples only depends on the pair of states hardest to distinguish, information sources may complement each other by covering for each other's weaknesses. So even though the composite test might be less precise for every individual state pair, the *lowest* precision may actually be higher. Putting everything together, we have a natural criterion for finding a low-price optimal bundle:

Corollary 3 (Maxi-Min precision bundles are (near) optimal). *Let* $\omega^*(c)$ *be the composite factor that uniquely maximizes the mininum precision per dollar for a given set of information sources and cost vector,* c

$$\omega^*(c) = \arg\max_{\omega} \min_{\theta,\theta' s.t. \theta\neq\theta'} \frac{\beta_\omega(\theta\theta')}{\omega\cdot \mathrm{c}}$$

Then the composite factor of the optimal bundle, $\omega(\mathrm{c},\mathrm{Y})$ *satisfies* $\omega^*(c)$

$$|\omega(\mathrm{c},\mathrm{Y}) - \omega^*(\mathrm{c})| = O(\log(\mathrm{Y})\mathrm{Y}^{-1})$$

as the budget goes to infinity.

Proof. The proof is the same as for Cor. 2, except for a minor technical complication, handled in Appendix A.2, arising from applying the approximation to bundles where the least-precision dichotomy is non-unique. □

In order to get a handle on when interior solutions exists and what form they typically take, it will be useful to think about the problem more geometrically. To that end, I'll now extend the notion of asymptotic relative efficiency to cover small tradeoffs between interior bundles by defining a notion of "marginal rate of substitution" appropriate for a discrete setting.

5 CONSUMER THEORY FOR CHEAP INFORMATION

5.1 *"Marginal" Rate of Substitution*

The previous section implies a natural maxi-min criterion for finding the low-price optimal bundle: for a given cost vector, find the relative proportion of samples from each bundle that maximizes the lowest precision per dollar. We can say a bit more about the nature of such solutions by applying some marginal analysis.

First, note that the definition and computation of Chernoff precisions for composite sources, doesn't depend at all on the samples being discrete. Thus, we can happily pretend for now that samples are a continuous quantity, and, thus, at any point where the least-precision dichotomy is unique, we can differentiate Eq. 10 to find the slope of an iso-least-precision line at a bundle with composite factor ω:

$$-\frac{\mathrm{d}n_2}{\mathrm{d}n_1} = \frac{\beta_{\omega 1}(\mathrm{D}_\omega)}{\beta_{\omega 2}(\mathrm{D}_\omega)}$$

where D_ω is the least precision dichotomy for that composite factor as before. Because, for fixed dichotomies we have that composite sources are less precise than the sum of their parts, we might expect the iso-precision lines to be outward bowing, and a quick computation confirms this: if the iso-least-precision line has a defined derivative, it must be *increasing* in magnitude with n_1. However, because we only care about the least-precision, there are finitely many inward pointing kink points where the least-precision dichotomy changes (recall Fig. 1 from Section 1).

With two available information sources, we clearly then must generically have that the maxi-min precision per dollar composite factor, $\omega^*(c)$, must be at either a corner or one of finitely many interior kink points where two dichotomies have the same precision. With more than two sources, we can apply this same geometric reasoning in higher dimensions to characterize the general maxi-min precision per dollar bundle:

Proposition 3 (Optimal bundle has as many sources as dichotomies of equal precision). *Suppose a decision-maker has a collection of available information sources, $\mathscr{E}_1, \ldots, \mathscr{E}_\mathrm{J}$, from each of which she may purchase an arbitrary number of samples at per-sample costs $c_1, \ldots, c_\mathrm{J}$ respectively. Then generically the relative proportions of each source in a bundle that maximizes the minimum precision per dollar generically must lie either at a corner (all but a single ω_j is zero), or must lie at a kink point where k dichotomies have the same precision, where k is the number of distinct information sources in the bundle, $\left|\{\mathscr{E}_j \mid \omega_j > 0\}\right|$.*

Proof sketch. Follows from a geometric reasoning: If there are k sources in the maxi-min precision per dollar bundle, that bundle lies at the tangency of a k-dimension hyperplane defined by the cost vector for those k sources and the iso-least-precision contour. Because the iso-least-precision contour is composed of multiple inward-bowing faces, the tangency must generically lie at an intersection of k faces. (Intuitively, if you drop a spiky ball, it generically lands on a point.)

Prop. 3 has an intuitive meaning: in a typical scenario, each test covers in the optimal bundle for a particular weakness of the other tests.

Alas, we cannot forever avoid the fact that samples are fundamentally discrete.[15] At the end of the day, we'd like to be able to draw meaningful conclusion about tradeoffs between

15 We could get around the issue by convexifying the choice space in some way. For example, we could consider lotteries over different sample bundles (which themselves define a new information source). One particularly elegant way to do this is to consider the *Poissonization* of an information source, where instead of getting n samples for sure, you get a random, Poisson-distributed number of samples. The Poisson parameter, a continuous object, then naturally takes the place of sample size (the sum of two Poisson draws is itself Poisson distributed). Although mathematically elegant, such an approach adds technical complication for little benefit.

discrete numbers of samples from the continuous curvature of the iso-precision lines, but in order to do so, we first need a formal discrete analog to the marginal rate of substitution:

$$k_2 = \min\{k : v(n_1 - k_1, n_2 + k, n_2, \ldots, n_J) \geq v(n_1, n_2, n_2, \ldots, n_J)\}$$

That is, k_2 is the *minimum* number of additional samples of $\mathscr{E}_2$ required to at least compensate for a loss of k_1 from $\mathscr{E}_1$, all else held fixed.[16] We can then define a k_1-discrete rate of substitution as

$$\mathrm{DRS}_{12}(\mathrm{n}; k_1) \equiv \frac{k_2}{k_1}$$

Now, recall that the component precisions only depend on ω. Furthermore, the change in ω from the above substitution is eventually small—roughly $O(k_1/\mathrm{N})$—so for large samples sources are *locally* perfectly substitutable, provided the substitution is small relative to total sample size and the least-precision dichotomy is unique at the bundle's composite factor. This suggests a natural definition for an asymptotic marginal rate of substitution:

Define the *asymptotic marginal rate of substitution* (AMRS) as

$$\mathrm{AMRS}_{12}(\omega) \equiv \lim_{\mathrm{N}\to\infty} \mathrm{DRS}_{12}(\omega_1\mathrm{N}, \ldots, \omega_J\mathrm{N}; k(\mathrm{N}))$$

Where the substitution size, $k(\mathrm{N})$, goes to infinity, but at a much smaller rate than N—i.e., $k(\mathrm{N}) = o(\mathrm{N})$.[17]

Putting aside notation, the AMRS tells us that for large enough total sample sizes, giving up k samples from $\mathscr{E}_1$ in exchange for more than $k \times \mathrm{AMRS}$ yields higher payoff. Symmetrically, receiving less than $k \times \mathrm{AMRS}$ in exchange yields a lower payoff. Graphically, it defines the slope of the boundary between upper and lower contour sets of a bundle with the given composite factor for large enough total sample size.

Using this definition, the AMRS between two information sources takes the form one might expect:

Proposition 4 (Substitutability of information sources). *Suppose the least-precision dichotomy is unique for composite factor* ω*, then for large enough* N*, the number of samples of* $\mathscr{E}_2$ *required to just compensate for a loss of* k *samples from* $\mathscr{E}_1$ *is*

$$k\mathrm{DRS}_{12}(\mathrm{n}; k) = \left\lceil k \frac{\beta_{\omega 1}(\mathrm{D}_\omega)}{\beta_{\omega 2}(\mathrm{D}_\omega)} \right\rceil$$

where D_ω *is the least-precision dichotomy for composite factor* ω*, and* $\lceil \cdot \rceil$ *denotes the ceiling*

16 k_1 can be negative.

17 Of course, we must start the limit sequence from high enough N so that the substitution is possible.

function. The AMRS is thus given by

$$\mathrm{AMRS}_{12}(\omega) = \frac{\beta_{\omega 1}(\mathrm{D}_\omega)}{\beta_{\omega 2}(\mathrm{D}_\omega)} \tag{11}$$

Proof sketch. Heuristically, we can solve for an indifference condition using Eq. 10, given a loss of k samples from $\mathscr{E}_1$, ignoring changes in the composite factor since they are eventually very small for a given substitution:

$$n_1\beta_{\omega 1}(\mathrm{D}_\omega) + n_2\beta_{\omega 2}(\mathrm{D}_\omega) = (n_1 - k)\beta_{\omega 1}(\mathrm{D}_\omega) + (n_2 + k\mathrm{AMRS})\beta_{\omega 2}(\mathrm{D}_\omega)$$

Rearranging terms to solve for AMRS gives the claimed formula. The formal proof is mostly a computational exercise and is thus relegated to the appendix.

Fig. 2 illustrates the AMRS numerically against a numerically computed lower boundary of an upper contour set. The top example illustrates a two-state environment where the AMRS always traces a bowed-out curve, and the bottom plot represents a three-state environment where one of the three dichotomies is never least-precision (similar to the example in Fig. 1). Section 6 will later discuss some typical properties of this approximation for typical information sources as sample sizes increase.

With this, we now have a formal basis for using the slope of iso-least-precision lines as an approximation for the rate at which samples from different sources may be exchanged for large samples. We thus have everything we need to characterize demand for information using standard consumer theory.

5.2 Demand under constrained optimization

With the results from the previous section, we can now easily derive some properties of information demand in this setting:

First, because the AMRS only depends on the relative proportions of each source (geometrically, the ray through the origin) for a bundle, preferences over information sources are homothetic by definition, and thus the optimal bundle composition is eventually income-independent. That is, when budgets are large enough, no source is ever either inferior or a luxury.

Proposition 5 (All sources are eventually normal goods). *For a generic cost vector—i.e. those that have a unique $\omega^*(c)$—the income elasticity of demand for all sources approaches unity as the budget grows large.*

Second, because optimal bundles must lie at one of finitely many kink points where the least-precision dichotomy is non-unique, we must have that demand for information behaves somewhat like the demand for perfect complements under Leonteif preferences at least locally. That is, for generic costs, the composition of the optimal bundle doesn't change with small changes in cost, but demand changes discontinuously around finitely many costs where the optimal kink changes.

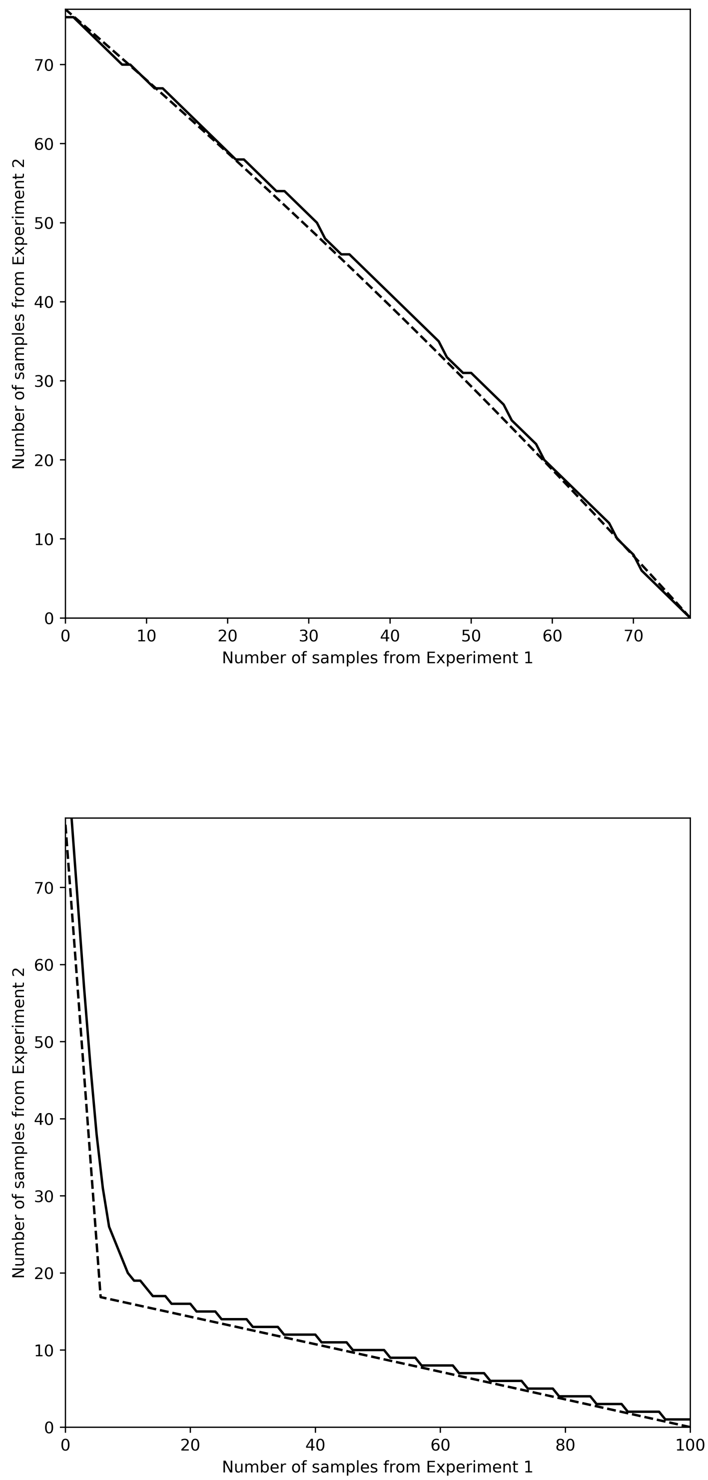

Figure 2: Top: Locus of bundles that perform *just* better than the lower right corner bundle (solid line), against the boundary predicted by Prop. 4 (dashed line) in a two-state decision problem. Bottom: The same plot for a three-state decision problem similar to the one illustrated by Fig. 1. The slope of the dashed line in both cases is by definition AMRS_{12}

We can thus write the demand elasticities as

Proposition 6 (Price elasticities). *For generic costs, the price elasticity of demand for all sources given a change in cost of source $\mathscr{E}_1$ is approaching*

$$\frac{-\omega_1^*(c)c_1}{\omega^*(c)\cdot c} + O(\%\Delta c_1)$$

as the budget grows large. Holding all other costs fixed, there are finitely many costs around which elasticities explode.

Put another way, for small prices changes, information sources are all perfect Marshallian complements, and changes in demand are entirely attributable to income effects.

These results further reinforce the peculiar nature of information goods, especially when considering the implications for competition between information sellers. Of course in many cases information is non-rival (and often difficult to make excludable), so there was already little reason to think competition between firms with linear pricing was sustainable, but these results show that, even ignoring such issues, a market for information goods, at least in simple environments with relatively few possible states,[18] requires some special treatment.

For example, Prop. 6 implies that for any cost ratio with an interior optimal bundle, demand is inelastic, and thus, all else fixed, a monopolistically competitive firm can always locally improve its profits by raising its prices at least up to the point where the optimal bundle jumps to the next kink. Thus, in a fixed budget scenario, a monopolistic competition is likely to yield to either monopoly or some form of non-linear pricing such as a subscription scheme.

6 HOW GOOD IS THE APPROXIMATION?

It's worth considering whether any of this even matters. After all, in principle, the approximation only holds when many samples are purchased and error probabilities are vanishing. Nonetheless, I claim that these approximation are in fact useful. To that end, I briefly consider the performance of the considered approximations, both theoretically and numerically.

6.1 Theoretical performance

First, recall that the full-information gap approximation on which these results were based has a relative error roughly proportional to $1/N$, which by asymptotic approximation standards is respectable. By comparison, the standard central limit theorem approximation, a staple method for approximating standard errors in applied work, typically has an error proportional to $1/\sqrt{N}$. Although the approximations are not substitutes for one another, the central limit theorem is a useful reference point due to it's ubiquity in statistical approximations.

18 The approximations work best when there are relatively few states. See the next section for further discussion.

Second, from a practical standpoint, even though mistakes are astronomically unlikely, the relative nature approximation implies that the predicted optimal bundle eventually performs *much* better than any alternative:

Proposition 7 (The optimal policy is eventually much better than any other). *Let* n_Y^* *be the sample vector with composition closest to* $\omega^*(c)$ *feasible under budget* Y. *Further let* n_Y^ω *be a policy with composition factor* $\omega \neq \omega^*(c)$. *Then we have*

$$\lim_{Y\to\infty} \frac{\mathrm{FIG}(n_Y^*)}{\mathrm{FIG}(n_Y^\omega)} = 0$$

Proof. It suffices to show that $\log(\mathrm{FIG}(n_Y^\omega)) - \log(\mathrm{FIG}(n_Y^*))$ ratio goes to infinity for large budgets. But this follows immediately because Eq. 8 and Eq. 10 imply that $-\log(\mathrm{FIG})$ is the budget times the precision-per-dollar of the of the bundle. □

More succinctly, as budgets get large, the FIG from following the optimal policy is eventually arbitrarily smaller than the FIG from any other policy.

Put another way, suppose we were to flip the problem asking instead: what is the minimum budget required to achieve a certain performance? In this case, Prop. 7 implies that as the target FIG gets small, the budget required for the optimal strategy is arbitrarily smaller than the budget required for some other strategy, and thus that the budget necessary to achieve a fixed level of performance is extremely sensitive to the choice of bundle for large samples.

6.2 Numerical observations

Lastly, let's consider how the approximation works in actual practice.

- include tables and plots showing approximation error for
  - DRS
  - Demand in the constrained case (and elasticities)
  - Demand in the quasilinear case (and elasticities)

  as sample sizes get large.

Finally, I note some qualitative properties of the approximations in practice, many of which are apparent in Fig. 2:

First, although indifference curves are definitively bowed outward for generic pairs of sources, in practice the outward bow is very subtle. The top plot of Fig. 2 was specifically selected to maximize the bow, and even then, the bow is barely perceptible. Given this, except in extraordinarily extreme cases, if two information sources differ in their least precision dichotomy, then interior solutions will exist for a non-empty set of price ratios.[19]

Second, the approximation performs relatively poorly very close a kink. Heuristically, there are two reasons for this: first, the approximation throws out all mistakes but the most likely one, but close to a kink, the second most likely mistake is very close to the most likely, so discarding it leads to a sizable overestimate of value, and second, because the kinks are inward pointing, the total sample size tends to be less than at the corners. For example, in Fig. 2, total sample size is roughly 30 at the kink, but around 100 at the corners.

Finally, because the true contour set boundary tends to be relatively smooth in the vicinity of a kink (although a very tight curve at large samples), the substitution effects will tend to be larger if there are more kinks. Since environments with more possible underlying states tend to have more kinks, it's reasonable to expect that contour sets would look more like the smooth curves of a textbook indifference curve. I thus caution against applying these approximations outside of simple environments with relatively few alternatives.

7 CONCLUSION AND FUTURE RESEARCH

This paper contributes to the literature on information values by applying previous results from the theory of large deviations to the evaluation of composite information sources. In particular, I've applied and extended the results of the results of Moscarini & Smith (2002) to develop an ordinal theory of tradeoffs for information sources.

I've shown that, for low costs/large budgets, maximizing information value is equivalent to maximizing the total Chernoff precision of the dichotomy that the information bundle most struggles to tell apart, and thus interior solutions can occur whenever two sources differ in their least-precision dichotomies.

Furthermore, I've explored implications for the standard consumer theory questions such as price, cross-price, and income elasticities for information demand. Specifically, in the "simple" (finite state) decision setting considered here, information goods will only be consumed in one of finitely many possible relative proportions, and thus for small changes in price, will behave like perfect complements.

These results naturally contribute to the broad literature in economics on information demand and suggest possible functional forms for information demand in applied settings such as media. On a more practical note, these approximations suggest a criterion (maxi-min precision per dollar) for the optimal design of experiments when the space of alternatives is small. Of course, the finite state assumption is somewhat restrictive, but the large deviation approach considered here should serve as a starting point for a general criterion for experiment design founded in Bayesian decision theory.[20]

19 The only way this can fail is if the iso-precision line for one dichotomy is sufficiently bowed so as to intersect another twice. This would lead to two kink points not on the convex hull of an upper contour set, and thus are never optimal.

20 In a separate project I show that Fisher information is the continuous-state analog to the Chernoff precision.

Appendix A OMITTED PROOFS

A.1 *Proof of Prop. 1*

Similar to Moscarini & Smith (2002), I start with the simple hypothesis testing problem. That is, we have two states, θ_0 and θ_1, with prior that θ_1 is true given by p, and two actions, accept and reject, where reject is optimal when θ_1 is true. We can then write the expected loss as

$$\mathrm{L}(n) = (1-p)\alpha_{\mathrm{I}}(n)\mathrm{L}_{\mathrm{I}} + p\alpha_{\mathrm{II}}(n)\mathrm{L}_{\mathrm{II}}$$

where L_{I} and L_{II} are the ex-post losses from Type-I (rejecting when θ_0 is true) and Type-II errors respectively, and α_{I} and α_{II} are the probabilities of those errors under a Bayesian decision rule.

In this case, we can write the Type-I error probability as the probability the probability that the posterior log-likelihood ratio is above the rejection threshold when the true state is θ_0:

$$\mathbb{P}\left(\sum_{k=0}^{n} \log\left(\frac{\mathrm{dF}(x_k|\theta_1)}{\mathrm{dF}(x_k|\theta_0)}\right) > \bar{l} \,\middle|\, \theta_0\right)$$

That is, because Bayes's rule is a sum when working with log-likelihood ratios, we can apply one of any number of statistical results for approximating distributions of sums and sample means. In particular, the above probability is one well-approximated a *large deviations* approach because at large sample sizes $\bar{l}$ is very far from the log-likelihood ratio sum's expectation (which drifts to $-\infty$ as n gets large). Moscarini & Smith (2002) apply a classic change-of-measure approach similar to Cramér (1938) to prove their main result. In contrast, I use a *saddlepoint* approach which gives a tighter bound on the approximation error:

The saddlepoint approach roughly works by applying the method of steepest descents (see Ch. 17 Jeffreys & Jeffreys 1956) to a inversion of the characteristic function. Daniels (1954) first applied this approach using the classic inversion formula for a density, but our log-likelihood ratio doesn't necessarily have a density. So instead, I rely on an approximation due to Lugannani & Rice (1980) who used a variation on the Gil-Pelaez (1951) characteristic inversion formula:

Lemma (Characteristic function inversion). *If* Y *is a random variable with characteristic function* ϕ, *then the survivor function of* Y *is*

$$\mathbb{P}(\mathrm{Y} \geq y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-iuy}\phi(u)}{iu}\mathrm{d}u$$

where the path of integration is perturbed to avoid the singularity at the origin.

We can then approximate the survivor function for the sample average by applying the method of steepest descents to the characteristic function of the sample sum:

Lemma (Lugannani & Rice, 1980). *Let* Y *be a real-valued random variable with* $\mathbb{E}(\mathrm{Y}) < 0$, *and* $\bar{\mathrm{Y}}_n$ *be the sample average of* n *i.i.d. draws of* Y. *If* Y*'s characteristic function* $\phi(u)$, *analytic through a strip,* $\{u : -\mathfrak{I}(u) \in (\tau-\epsilon, \tau+\epsilon)\}$, *where* $\tau \equiv \arg\min_t \mathbb{E}(e^{t\mathrm{Y}})$, *then for* ξ *close enough to zero*

$$\mathbb{P}(\bar{\mathrm{Y}}_n \geq \xi) = \frac{e^{n(\mathrm{K}(\tau(\xi))-\tau(\xi)\xi)}}{\tau(\xi)\sqrt{2\pi n \mathrm{K}''(\tau(\xi))}}\left(1 + O\left(\frac{1}{n}\right)\right) \tag{12}$$

where $\mathrm{K}(t) \equiv \log(\mathrm{M}(t))$ *is the cumulant generating function of* Y *and* $\tau(\xi)$ *is the minimizer of* $\mathrm{K}(t) - tx$.

It's worth noting for later that this saddlepoint approximation is a valid approximation of the characteristic function inversion of ϕ^n, even when ϕ isn't a valid characteristic function, so long as the inversion formula is still integrable and ϕ is analytic on an appropriate strip containing a unique maximum (of course, in such a setting, the resulting function won't be a valid survivor function for any distribution).

We can quickly verify that the distribution of log-likelihood ratios satisfies the above assumptions by writing the moment-generating function of the log-likelihood ratio as

$$\mathrm{M}(t;\theta_1\theta_0) = \int \exp\left(t\log\left(\frac{\mathrm{dF}(x|\theta_1)}{\mathrm{dF}(x|\theta_0)}\right)\right)\mathrm{dF}(x|\theta_0) = \int \mathrm{dF}(x|\theta_1)^t\mathrm{dF}(x|\theta_0)^{1-t} = \mathrm{M}(t;\theta_0\theta_1)$$

Because of this symmetry, I will simplify notation by writing $\mathrm{M}(t) \equiv \mathrm{M}(t;\theta_1\theta_0)$. Now, recall we assumed that the log-likelihood ratios had all finite moments, so $\mathrm{M}(t;\theta_1\theta_0)$ must be infinitely differentiable at $t = 0$ and $t = 1$, and thus all points between (by dominated convergence and convexity of e^{tx}). Because the characteristic function is $\phi(u) = \mathrm{M}(-iu)$, we must have ϕ analytic for any u such that $-iu$ is in the unit interval. Lastly, the minimizer of M must lie in $(0,1)$ because moment-generating functions are (log) convex and $\mathrm{M}(0) = \mathrm{M}(1) = 1$.

To finish the proof, we need now only let $\xi_n = \bar{l}/n$, and apply Taylor's theorem:

Let τ be the minimizer of $\mathrm{K}(t)$ (and thus the minimizer of the MGF). By applying Taylor's theorem and the FOC, $\mathrm{K}'(\tau) = 0$, we can the write

$$\begin{aligned}
\tau_n &= \tau + O(1/n) \\
\mathrm{K}(\tau_n) &= \mathrm{K}(\tau) + \frac{1}{2n^2}\mathrm{K}''(\tau) + O(1/n^3) \\
\mathrm{K}''(\tau_n) &= \mathrm{K}''(\tau) + O(1/n)
\end{aligned}$$

Note that by definition of the Chernoff precision and the efficiency index we have, $\mathrm{K}(\tau) = -\beta$ so $e^{n\mathrm{K}(\tau)} = \rho^n$. We can then plug each of these into Eq. 12 and apply Taylor's theorem again. Breaking it down into parts we have

$$\begin{aligned}
e^{n\mathrm{K}(\tau_n)} &= \rho^n(1 + O(1/n)) \\
e^{-n\tau_n\xi_n} &= e^{\tau(\bar{l})}(1 + O(1/n)) \\
\tau_n\sqrt{2\pi n\mathrm{K}''(\tau_n)} &= (\tau + O(1/n))\sqrt{2\pi n\mathrm{K}''(\tau) + O(1)} \\
&= \tau\sqrt{2\pi n\mathrm{K}''(\tau)}(1 + O(1/n))
\end{aligned}$$

Plugging each of the above parts into Eq. 12 we have that the error probability is

$$\frac{e^{\tau\bar{l}}}{\tau\sqrt{2\pi\mathrm{K}''(\tau)}}\frac{\rho^n}{\sqrt{n}}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{13}$$

Plugging this into the original equation for expected loss gives the claimed result for two-state/two-action decision problems. Completing the proof for general finite-state/finite-action decision problems follows from application of Thm. 4 of Moscarini & Smith. □

A.2 Proof of Cor. 3

We proceed in two parts: first, show that for any N we can find $\bar{A} > 0$ such that for N high enough, the remainder on the negative-log expected loss approximation is smaller than $\bar{A}\log(N)/N$ *for all* possible sample proportions, then show that a bundle with maxi-min precision per dollar beats any bundle with proportions more than ϵ away with epsilon falling inversely to income.

In order to show that such an $\bar{A}$ upper bound for the error uniformly exists, it will be useful to close the choice space. Because samples can only be consumed discretely, only rational $\boldsymbol{\omega}$ are ever correspond with a true sample bundle. Note however that the composite log-likelihood ratio moment generating functions, $M_{\boldsymbol{\omega}}$, are defined *for any* $\boldsymbol{\omega}$. We can thus apply the characteristic formula inversion formula from the previous proof to get a "distribution" that we can use to compute an "expected loss." We can thus extend the expected loss function to all positive real "sample sizes." Furthermore, because $M_{\boldsymbol{\omega}}$ is pointwise-continuous in $\boldsymbol{\omega}$, this extended loss function will be as well. Finally, notice that M_{ω}, still has the required properties to apply the saddlepoint approximation—namely, it is analytic on unit interval (and thus the characteristic function is analytic on the strip with imaginary part in the unit interval), and has a unique minimum within that interval.

Thus, we can write the negative-log of the extended expected loss as

$$-\log(L(\mathbf{n})) = \min_{D}\{N\beta_{\omega}(D)\}\left(1 + \frac{A(\boldsymbol{\omega}, N)\log(N)}{N}\right)$$

where $\mathbf{n} \in \mathbb{R}_{+}^{J}$ and A is some continuous function that is $O(1)$ for all ω fixed. Note that because L and β_{ω} are continuous[21], this expression must hold even for ω with non-unique worst-case dichotomies. Lastly $\bar{A}(\omega) = \sup_N |A(\omega, N)| < \infty$ is continuous in ω by the theorem of the maximum, and thus have a finite upper bound, $\bar{A}$, since the set of sample proportions in compact.

Now to show that the optimal sample proportions are within ϵ of the maxi-min precision, I show that the maxi-min precision bundle must eventually beat any bundle more than ϵ away.

First, samples must be chosen discretely, so we'll first let $\boldsymbol{\omega}_N^*$ be the sample proportions as close as possible to $\boldsymbol{\omega}^*$ with sample size N. Note that the distance between the two is $O(N^{-1})$.

Now let $N^* = Y/(\boldsymbol{\omega}^* \cdot \mathbf{c})$ be the total sample size of spending the entire budget on samples in proportions $\boldsymbol{\omega}^*$. Suppressing dependence on the dichotomy, we can then write an *underestimate* of the negative-log expected loss achieved by spending the entire budget on a bundle with sample proportions as close as possible to ω^*:

$$-\log(L^*) \geq Y\frac{\beta_{\omega^*}}{\boldsymbol{\omega}^* \cdot \mathbf{c}}\left(1 + O\left(\frac{\log(Y)}{Y}\right) - \frac{\bar{A}\log(Y/(\boldsymbol{\omega}^* \cdot \mathbf{c}))}{Y/(\boldsymbol{\omega}^* \cdot \mathbf{c})}\right)$$

Where the $O(\log(Y)Y^{-1})$ arises from replacing the closest possible sample proportions to $\boldsymbol{\omega}^*$ with exactly $\boldsymbol{\omega}^*$.[22] Now we want to find an *overestimate* of the same for the all points more than ϵ from ω^*. First pick ϵ small enough that highest precision sample proportions are on the boundary of the ϵ ball around ω^*. Repeating the above procedure, we can then write

$$-\log(L^{\epsilon}) \leq Y\frac{\beta_{\omega^*}}{\boldsymbol{\omega}^* \cdot \mathbf{c}}\left(1 + O(\epsilon) + \frac{\bar{A}\log(Y/(\boldsymbol{\omega}^* \cdot \mathbf{c}))}{Y/(\boldsymbol{\omega}^* \cdot \mathbf{c})}\right)$$

where L^{ϵ} is the best possible expected loss for proportions more than ϵ away from $\boldsymbol{\omega}^*$. Rearranging and canceling terms, we have $L^* < L^{\epsilon}$ if and only if a $O(\epsilon)$ term is smaller than a $O(\log(Y)Y^{-1})$ term. Thus the optimal sample proportions must be within $\epsilon = O(\log(Y)Y^{-1})$ of ω^*. $\square$.

21 We must use the log expected loss here, because the multiplicative constant on the expected loss approximation jumps discontinuously—and thus so too would the remainder—when the worst-case dichotomy changes (see Eq. 13).

22 Follows from Taylor's theorem. Even though the worst-case precision has a kink ω^*, the derivatives along paths into the kink are uniformly bounded=.

A.3 Proof of Prop. 4

Fix ω such that the worst-case dichotomy is unique. Since the worst-case dichotomy will always be the same throughout the proof, I supress any dependence on it. Let k_2 be the minimum number of samples of $\mathscr{E}_2$ that just compensates for a loss of k_1 samples from $\mathscr{E}_1$. Then we have

$$\left(n_1\beta_{1\omega} + n_2\beta_{2\omega} + \sum_{j=3}^{K} n_j\beta_{j\omega}\right) + \log(C_\omega) + O(N^{-1}) \leq (n_1 - k_1)\beta_{1\omega'} + (n_2 + k_2)\beta_{2\omega'} + \sum_{j=3}^{K} n_j\beta_{j\omega'} + \log(C_{\omega'}) \quad (14)$$

where ω' is the composite factor associated with the new sample bundle and C_ω is the proportionality constant for the expected loss approximation. Start with N high enough that this substitution doesn't change the worst-case dichotomy. Then notice that $\omega' - \omega = O(N^{-1})$. Applying this fact with Taylor's theorem to the FOC for β_ω, we have that $\tau_{\omega'} - \tau_\omega = O(N^{-1})$ as well. We can then apply Taylor's theorem to write

$$\begin{aligned}
&(n_1 - k_1)\beta_{1\omega'} + (n_2 + k_2)\beta_{2\omega'} + \sum_{j=3}^{K} n_j\beta_{j\omega'} \\
&\quad = (n_1 - k_1)\beta_{1\omega} + (n_2 + k_2)\beta_{2\omega} + \sum_{j=3}^{K} n_j\beta_{j\omega} + \left[k_2\frac{M'_{2\omega}(\tau_\omega)}{M_{2\omega}(\tau_\omega)} - k_1\frac{M'_{1\omega}(\tau_\omega)}{M_{1\omega}(\tau_\omega)}\right](\tau_{\omega'} - \tau_\omega) + O((\tau_{\omega'} - \tau_\omega)^2) \\
&\quad = (n_1 - k_1)\beta_{1\omega} + (n_2 + k_2)\beta_{2\omega} + \sum_{j=3}^{K} n_j\beta_{j\omega} + O(N^{-1})
\end{aligned}$$

Further, because C_ω is a differentiable function of τ_ω (see the last part of the proof of Prop. 0), we have that $\log(C_{\omega'}) - \log(C_\omega) = O(N^{-1})$. Plugging all of this into (14) and rearranging gives

$$k_2 \geq k_1\frac{\beta_{1\omega}}{\beta_{2\omega}} + O(N^{-1}) \quad (15)$$

Repeating this procedure for the substitution of k_1 of $\mathscr{E}_1$ for $(k_2 - 1)$ of $\mathscr{E}_2$ (which does just worse than the original bundle) gives

$$k_2 \leq k_1\frac{\beta_{1\omega}}{\beta_{2\omega}} + 1 + O(N^{-1}) \quad (16)$$

Together, by squeezing k_2 between (15) and (16) we have for N large enough

$$k_2 = \left\lceil k_1\frac{\beta_{1\omega}}{\beta_{2\omega}}\right\rceil$$

To complete the proof, let $k_1(N)$ grow to infinity but at rate oN. We can then write:

$$\frac{\beta_{1\omega}}{\beta_{2\omega}} + O\left(\frac{N^{-1}}{k_1(N)}\right) \leq \text{AMRS} \leq \frac{\beta_{1\omega}}{\beta_{2\omega}} + \frac{1}{k_1(N)} + O\left(\frac{N^{-1}}{k_1(N)}\right)$$

Which by squeezing gives

$$\text{AMRS} = \frac{\beta_{1\omega}}{\beta_{2\omega}}$$

as claimed. □.

BIBLIOGRAPHY

Azrieli, Y. (2014), 'Comment on "The Law of Large Demand for Information"', *Econometrica* **82**(1), 415–423.

Blackwell, D. (1951), Comparison of experiments, *in* 'Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability', University of California Press, pp. 93–102.

Butler, R. W. (2017), *Saddlepoint Approximations with Applications*, Cambridge University Press.

Chade, H. & Schlee, E. E. (2002), 'Another look at the Radner-Stiglitz nonconcavity in the value of information', *Journal of Economic Theory* **107**(2), 421–452.

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations', *The Annals of Mathematical Statistics* **23**(4), 493–507.

Cramér, H. (1938), 'Sur un nouvaeu théorème-limite de la théorie des probabilités', *Actualités Scientifiques et Industrielles* (736).

Daniels, H. E. (1954), 'Saddlepoint approximations in statistics', *The Annals of Mathematical Statistics* **25**(4), 631–650.

Dembo, A. & Zeitouni, O. (1998), *Large Deviations Techniques and Applications*, 2nd edn, Springer Berlin Heidelberg.

Gil-Pelaez, J. (1951), 'Note on the inversion theorem', *Biometrika* **38**(3/4), 481.

Jeffreys, H. & Jeffreys, B. (1956), *Methods of Mathematical Physics*, 3rd edn, Cambridge University Press.

Lugannani, R. & Rice, S. (1980), 'Saddle point approximation for the distribution of the sum of independent random variables', *Advances in Applied Probability* **12**(2), 475–490.

Montgomery, D. C. (2012), *Design and Analysis of Experiments*, 8th edn, John Wiley & Sons.

Moscarini, G. & Smith, L. (2002), 'The law of large demand for information', *Econometrica* **70**(6), 2351–2366.

Pomatto, L., Strack, P. & Tamuz, O. (2018), 'The cost of information'.

Radner, R. & Stiglitz, J. E. (1984), 'A nonconcavity in the value of information'.

Torgersen, E. (1991), *Comparison of Statistical Experiments*, Cambridge University Press.