CONSUMER THEORY FOR CHEAP INFORMATION

Gary Baker

This version: 11 October 2021

CLICK HERE FOR THE MOST RECENT VERSION

ABSTRACT

Classic comparisons—e.g. Blackwell efficiency—of information sources tell us little about tradeoffs between different sources, especially when they differ in cost. This paper seeks to fill that gap for finite-state environments by describing the consumer theory of information when information is cheap (or budgets large). Motivated by large deviations theory, I propose a generalized notion of precision that measures how well an information source distinguishes two possible states. I then show that maximizing the precision of the worst-case state pair yields an approximation for information demand with percent error vanishing proportionally with costs. I further show iso-least-precision sets have finitely many kinks and are otherwise bowed *out*, and thus information demand exhibits vanishing substitution effects at almost all cost ratios. This kinked geometry additionally implies an upper bound on the number of information sources that will ever be used in non-vanishing proportions: at most as many as there are state pairs. Finally, because precision is independent of prior and payoffs, all decision-makers roughly agree on the optimal bundle. In sum, demand for information starkly deviates from the benchmark convex preferences model of standard consumer theory.

KEYWORDS: Demand for information, value of information, Bayesian decision theory, comparison of experiments, large deviations theory

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 news 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. With the explosive growth of the internet and the availability of not just near-unlimited quantities of information, but also information from highly varied and distinct sources, questions such as these are more relevant than ever.

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 standard consumer theory exercise: set the marginal rate of substitution 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]

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).

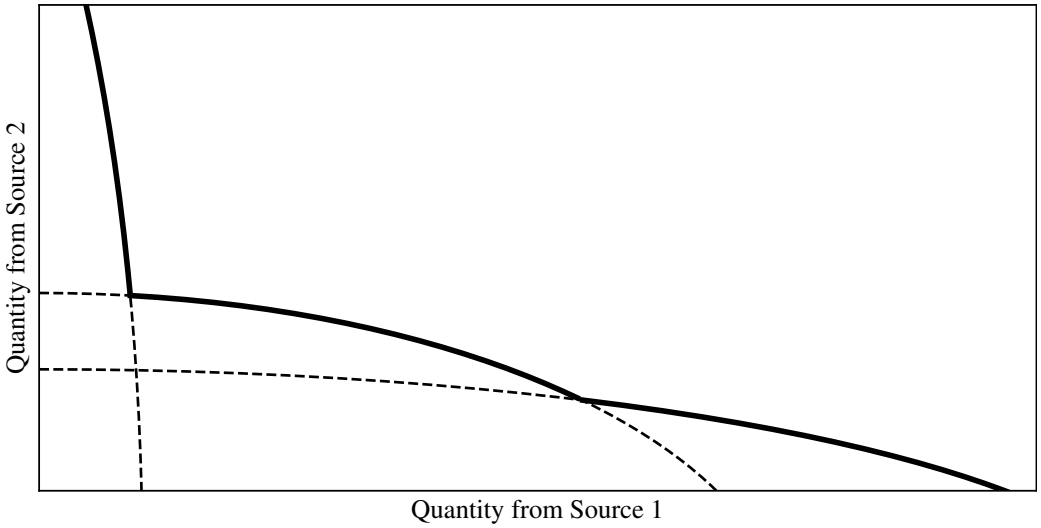

Figure 1: An example geometry of iso-precision lines—and thus the approximate geometry of preferences—in a three-state environment with two available information sources. Iso-precision lines for a fixed dichotomy (dashed lines) bow out, but the iso-least-precision line (solid line) has inward pointing kinks.

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).

In particular, drawing on large-deviations methods, I define a generalized notion of *precision* that measures how well an information source (Blackwell experiment) discriminates between a given pair of possible states (a dichotomy). I then show that a maximin rule—maximize the total precision of the bundle for the worst-case state pair—yields an approximation for demand, with percent error vanishing proportionally with costs (Proposition 1). Furthermore, because precision is a purely statistical property, independent of prior and payoffs, all decision-makers roughly agree on the optimal bundle at low costs.

I then explore properties of demand by treating maximin precision *as if* it were a utility function. In particular, for a fixed-dichotomy precision is homothetic and iso-precision-lines bow *out*—that is, precision of composite sources is less than the sum of its part—so iso-*least*-precision lines exhibit inward-pointing kinks (Figure 1).

Maximin precision bundles thus only occur in finitely-many possible relative proportions (Proposition 2), corresponding with iso-least-precision corners and kinks, and thus demand for information behaves as if sources were perfect complements for small price changes, with large substitution effects only near costs where the maximin precision bundle jumps between kinks/corners. Furthermore, homotheticity of precision guarantees all information sources have income elasticities approach unity, and thus no information source is ever an inferior good.

The kinked geometry additionally implies a relationship between the number of possible states number of information sources ever consumed in non-vanishing proportions (Proposition 3). As a consequence, at low costs, no decision-maker will ever consume more information sources (in non-vanishing proportions) than there are pairs of states. Thus, in the simplest, two-state hypothesis testing world, only corners are ever optimal, with interior solutions only occurring with more complex decision problems.

Although the as-if-kinked nature of preferences would seemingly render discussion of the marginal rate of substitution away from the kinks pointless, information is not a typical good.

Data sets might have hard upper bounds on the number of available samples and the non-rival nature of information consumption often leads to non-linear (e.g. subscription) pricing schemes. Such is to say: even with kinks, a complete consumer theory should describe the marginal rate of substitution elsewhere. To that end, I show that, when the worst-case dichotomy is unique, the information from distinct sources is substitutable roughly in proportion to each source's *marginal* precision (Proposition 4).

This paper is most closely related to the work of Moscarini & Smith (2002) who draw on large-deviations methods to write an approximation for information value, and thus information demand, in a quasilinear setting with one available source of information. My Proposition 0 generalizes their main result to a setting with multiple available sources and improves upon their estimate of the approximation's convergence rate.

Additionally, this result fits into the statistical literature on asymptotic relative efficiency, which traditionally asks how many samples from one statistical test are required to perform as well as a given number from another—that is, relative efficiency typically only considers the corners. In particular, these results generalize Chernoff's (1952) notion of relative efficiency to a setting with multiple hypotheses (more than two states) and demonstrates the existence of non-corner solutions in such a setting.

These results lend themselves to multiple applications. First and foremost, they provide a general approach for analysis of information markets, and suggest caution when dealing with information as a good in applied settings. In particular, the benchmark consumer theory model (smooth, convex preferences) is fails rather starkly.

On a more practical note, the maximin precision rule also suggests a criterion for optimal experiment design, accounting not just for differing statistical properties of the different options, but also their costs.[2]

The remainder of this paper is structured as follows: Section 2 lays out the formal model assumptions, including the relevant notion of information "quantity" used throughout this paper. Section 3 covers the necessary large-deviations background. Section 4 defines precision, states the maximin approximation rule, and explores features of the implied consumer theory. Lastly, Section 5 explores the performance of the approximations in practice. Finally, Section 6 concludes.

2 MODEL

A decision-maker (DM) must choose an action a from a finite set, A, under an uncertain state of the world θ drawn from a finite set, Θ. The DM has Bernoulli utility $u(a, \theta)$ and a full-support prior $p \in \Delta\Theta$. For simplicity, assume the optimal action for each state, $a^*(\theta) \equiv \arg\max_a u(a, \theta)$, is unique and distinct for all states. The DM chooses her action to maximize expected payoffs.

Prior to acting, the DM may purchase information about the state of the world from J distinct sources, $\mathscr{E}_1, \dots, \mathscr{E}_J$. Each information source is a conditionally independent Blackwell experiment, $\mathscr{E}_j = \langle \mathbb{X}, \langle \mu_{j\theta} \rangle_{\theta \in \Theta} \rangle$, consisting of a collection of state-dependent distributions over some arbitrary space of realizations, $\mathbb{X}$. Assume each information source is informative, but no realization perfectly rules in or any strict subset of the world. Formally, $\mu_{j\theta}$ and $\mu_{j\theta'}$ are mu-

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

tually absolutely continuous and thus have defined likelihood ratios (Radon-Nikodym derivatives) $d\mu_{j\theta}/d\mu_{j\theta'}$.

Under this assumption, the $\mathbf{R}^{|\Theta|-1}$ vector of log-likelihood ratios relative to some base state, θ_0, is a sufficient statistic for each realization. We can thus without loss of generality identify each realization by its own vector of log-likelihood ratios:

$$\mathbf{x} \equiv (x_\theta)_{\theta \neq \theta_0} = \left(\log \left(\frac{d\mu_{j\theta}}{d\mu_{j\theta_0}}(\mathbf{x}) \right) \right)$$

Because Bayes's rule is a sum when expressed in terms of log-likelihood ratios, we thus can summarize conditionally independent realizations from multiple sources simply by adding their log-likelihood ratios. To avoid technicalities, assume all log-likelihood ratio distributions are thin-tailed in the sense that they have well-defined moment-generating functions on an open set containing the origin.

The DM chooses a *quantity*, $\mathbf{t} = (t_1, \dots, t_J) \in \mathbb{R}^J$, of information from each information source at costs, $\mathbf{c} = (c_1, c_2, \dots, c_J) = (\varepsilon, \varepsilon\kappa_2, \dots, \varepsilon\kappa_J)$ per unit quantity from each source subject to budget constraint, Y. The key approximation will hold as costs become small, which I capture by setting ε to be small.

Quantity of information is usually measured by conditionally i.i.d. samples—a fundamentally discrete object, but to avoid discrete-optimization technicalities, I will assume for exposition[3] that sources are *infinitely divisible*. Formally, $\mathscr{E}$ is infinitely divisible if for any k, there exists an experiment $\mathscr{E}^{1/k}$ such that k conditionally i.i.d. samples from $\mathscr{E}^{1/k}$ are equivalent to a single sample from $\mathscr{E}$.

Thus, a rational quantity $t_j = a/b$ of information from $\mathscr{E}_j$ is equivalent to a conditionally i.i.d. samples from $\mathscr{E}_j^{1/b}$. Formally, the state-θ log-likelihood ratio distribution of rational quantity $t = n/k$ of $\mathscr{E}_j$, $\mu_{j\theta}^t$, is $\star_{i=1}^n \mu_{j\theta}^{1/k}$, i.e. the n-fold convolution[4] of the log-likelihood ratio distribution of $\mathscr{E}^{1/k}$. Non-rational quantities are simply the appropriate limit—for example, the sum of the appropriate number of samples from $\mathscr{E}_j^{1/10}$, $\mathscr{E}_j^{1/100}$, and so on.[5] Because the moment-generating function of a sum is simply the product of moment-generating functions, the quantity-t conditional distributions of log-likelihood ratios are perhaps more easily described in terms of their moment-generating functions: if the moment-generating function of $\mu_{j\theta}$ is $M_{j\theta}(\zeta)$, then the moment-generating function of $\mu_{j\theta}^t$ is $M_{j\theta}(\zeta)^t$.

We can equivalently view quantity of information as a time spent observing a continuous-time, state-dependent process with conditionally independent increments.[6] For example, a DM might choose how long to watch a given news channel, where state-dependent new stories arrive according to a (state-independent) Poisson process. Given the ubiquity of time constraints, in my budget-constrained setting, time is perhaps the most natural interpretation; however, remember the model is formally static: the DM chooses quantity, then observes the realizations all at once.

After choosing an information bundle, the DM observes the realizations, Bayes updates her beliefs appropriately, and chooses an action to maximize her expected payoff. The DM

3 All results hold when quantity is measured by non-divisible samples. See Appendix B.

4 $(\mu_1 \star \mu_2)(S) \equiv \int \mathbf{1}_S(x+y)\mu_1(dx)\mu_2(dy)$, i.e. the distribution of the sum.

5 Theorem 3, Chapter 9 of Le Cam (1986) guarantees that such a limit is a valid experiment.

6 See Theorem 2.IX.5 of Feller (1970) for a proof of this claim.

thus wants to choose the feasible bundle that minimizes her expected loss[7] from acting after observing the realizations:

$$L(\mathbf{t}) = \sum_{\theta} p_\theta \int_{\mathbf{x}\in\mathbb{R}^{|\Theta|-1}} \big(u(a^*(\theta),\theta) - u(a(\mathbf{x}),\theta)\big)\mu_\theta^{\mathbf{t}}(d\mathbf{x}) \tag{1}$$

where $a(\mathbf{x})$ is the expected-payoff-maximizing decision[8] after observing realized log-likelihood ratios $\mathbf{x}$, and $\mu_\theta^{\mathbf{t}}$ is the log-likelihood ratio conditional distribution of realizations for the chosen expected sample bundle, $\mu_\theta^{\mathbf{t}} \equiv \bigstar_{j=1}^{J} \mu_{j\theta}^{t_j}$.

Except under restrictive functional-form assumptions on the available information sources, Equation (1) has no convenient closed form, necessitating the application of a large-sample approximation.

3 LARGE DEVIATIONS

Because large-deviations methods are relatively uncommon in economics, I will introduce the approach first in the two-state/two-action world. Here we can pose the decision problem as a classic statistical dichotomy: there are two states $\Theta = \{\theta_0, \theta_1\}$, corresponding with *null* and *alternative* hypotheses, and the DM must either choose to *reject* ($a = \mathscr{R}$) or *accept* ($a = \mathscr{A}$) the null. Naturally, the DM wants to reject when the null is false and vice-versa so the payoffs satisfy $u(\mathscr{R},\theta_1) > u(\mathscr{A},\theta_1)$ and $u(\mathscr{A},\theta_0) > u(\mathscr{R},\theta_0)$. Assume the DM has prior p that the alternative (θ_1) is true, and the belief that makes the DM exactly indifferent between the two actions as $\bar{p}$.

Following Moscarini & Smith (2002), we can write the expected loss from quantity $\mathbf{t}$ in this environment in terms of the Type-I and Type-II error probabilities (respectively, α_I and α_{II}):

$$L(\mathbf{t}) = (1-p)\alpha_I(\mathbf{t})(u(\mathscr{A},\theta_0) - u(\mathscr{R},\theta_0)) + p\alpha_{II}(\mathbf{t})(u(\mathscr{R},\theta_1) - u(\mathscr{A},\theta_1))$$

As the quantity of information gets large, the error probabilities should fall to zero. With a single source of information we can write the Type-I error probability, leveraging the fact that Bayes's rule is a sum when written in terms of log-likelihood ratios:

$$\alpha_I(t) = \mathbb{P}\left(l + s_t > \bar{l} \,|\, \theta_0\right) \tag{2}$$

where $l = \log(p/(1-p))$ is the prior log-likelihood ratio, $\bar{l} = \log(\bar{p}/(1-\bar{p}))$ is the log-likelihood ratio of the indifference belief, and s_t is the log-likelihood ratio of the realization of quantity t of information. Note that the expected value of the log-likelihood ratio when θ_0 is true must be negative—i.e. when θ_0 is true, on average the realizations should push the log-likelihood posterior *down* towards stronger beliefs that θ_0 is true.

Notice we could have equivalently written (2) in terms of the sample average log-likelihood

7 Equivalent to maximizing the usual value of information in a budget-constrained setting.

8 Although for simplicity, I work with a Bayesian optimal decision rule—i.e. the DM uses the optimal decision thresholds for a log-likelihood ratio test—similar to Chernoff (1952), my approach applies for general decision rules based on additive statistical test under suitable regularity conditions.

ratio as follows:

$$\alpha_I(t) = \mathbb{P}\left(\frac{s_t}{t} > \frac{\bar{l} - l}{t} \,\middle|\, \theta_0\right)$$

Here we can see why we can't use a more familiar asymptotic approach such as a central limit theorem: $\mathbb{E}(s_t/t) \equiv \bar{s} < 0$ but mistakes happen roughly only when the sample average is positive—i.e. *far* from its mean. This contrasts with the central limit theorem which describes the distribution of a sample average *near* the mean (roughly, within $1/\sqrt{t}$ of the mean).

Cramér (1938) canonically showed that the probability of such a large deviations is falling exponentially fast with rate given by a minimized moment-generating function. We can see a basic version of this by an application of Markov's inequality:

$$\alpha_I(t) \approx \mathbb{P}\left(\frac{s_t}{t} > 0 \,\middle|\, \theta_0\right) = \mathbb{P}\left(\exp\left(\zeta \frac{s_t}{t}\right) > 1 \,\middle|\, \theta_0\right) < \min_{\zeta}\{M(\zeta)\}^t$$

where M is the state-θ_0 log-likelihood ratio moment-generating function for a single unit of information. Motivated by this approach, we can then turn back to the general finite-state problem.

Define the *efficiency index*[9] of information source $\mathscr{E}_j$ for the θ, θ' dichotomy as the minimized value of the moment-generating function for the θ, θ' log-likelihood ratio conditional on true state θ':

$$\rho_j(\theta, \theta') \equiv \min_{\zeta} M_j(\zeta\,; \theta, \theta') = \min_{\zeta} \left\{ \int_{\mathbf{x} \in \mathbb{R}^{|\Theta|-1}} \mu_{j\theta}(d\mathbf{x})^{\zeta} \mu_{j\theta'}(d\mathbf{x})^{1-\zeta} \right\}$$

Note that $\rho_j(\theta, \theta') = \rho_j(\theta', \theta)$ because $M_j(\zeta\,; \theta, \theta') = M_j(1 - \zeta\,; \theta', \theta)$ so the index is unique for a given dichotomy, independent of order. This efficiency index—a special case of the index developed by Chernoff (1952) for evaluating the asymptotic relative efficiency of two statistical tests—describes the exponential rate at which the Type-I or Type-II error problems fall in a simple testing problem. Efficiency indices are always between 0 and 1 and are multiplicative for i.i.d. samples. Furthermore, *lower* efficiency index indicate *better* large sample performance.[10]

Unlike the previous literature, however, I am not just interested in the behavior of individual information sources in isolation, but also how they interact when used together. To generalize the efficiency index to a setting with multiple sources, first denote the total quantity, $T \equiv \sum t_j$ and the proportions of total quantity from each experiment as $\mathbf{r} = (r_1, \ldots, r_J) \equiv (t_1/T, \ldots, t_j/T)$.

Then define the $\mathbf{r}$-composite experiment as one with quantities $\mathbf{r}$ from each information source. This constructed source then has log-likelihood rate moment-generating functions, $M_{\mathbf{r}}(\zeta; \theta, \theta') \equiv \prod M_j(\zeta; \theta, \theta')^{r_j}$. By construction, quantity T from the $\mathbf{r}$-composite experiment is equivalent to the bundle $\mathbf{t}$ because $M_{\mathbf{r}}^T = \prod M_j^{t_j}$. Define the *composite* efficiency index $\rho_{\mathbf{r}}(\theta, \theta')$ analogously.

$$\rho_{\mathbf{r}}(\theta, \theta') \equiv \min_{\zeta} \left\{ \prod_{j=1}^{J} M_j(\zeta; \theta, \theta')^{r_j} \right\}$$

9 I follow Moscarini & Smith (2002) here. Torgersen (1991) refers to this as the *Chernoff number*.

10 If $\mathscr{E}_1$ Blackwell dominates $\mathscr{E}_2$ then $\mathscr{E}_1$ has lower efficiency indices for all dichotomies.

We can then further break $\rho_{\mathbf{r}}$ into contributions from each source by defining the *marginal* efficiency index of $\mathscr{E}_j$ as $\rho_{j\mathbf{r}}(\theta,\theta') \equiv M_j(\zeta^*_{\mathbf{r}};\theta,\theta')$ where $\zeta^*_{\mathbf{r}}$ is the minimizer of $M_{\mathbf{r}}(\cdot;\theta,\theta')$, so

$$\rho_{\mathbf{r}}(\theta,\theta') = \prod \rho_{j\mathbf{r}}^{r_j}(\theta,\theta')$$

With only two states, Moscarini & Smith (2002) show that each mistake probability, and thus the expected loss itself, is proportional to $\rho^t/\sqrt{t}$ for t large.[11] With more than two states, they further show that, because each mistake probability is exponentially falling, the expected loss is eventually dominated by the most likely mistake—that is, the *largest* efficiency index. Using marginal efficiency indices, I can now state a generalized version of their main result:

Proposition 0. *Let $\mathscr{D}$ be the collection of dichotomies. Then when the worst-case dichotomy, $\arg\max_D \rho_{\mathbf{r}}(D)$, is unique, the expected loss from consuming quantities $\mathbf{t} = [t_1,\dots,t_J]$ from $\mathscr{E}_1,\dots,\mathscr{E}_J$ is*[12]

$$L(\mathbf{t}) = A(\mathbf{r})\frac{\max_{D\in\mathscr{D}}\left\{\prod_{j=1}^{J}\rho_{j\mathbf{r}}(D)^{t_j}\right\}}{\sqrt{T}}\left(1+O\left(\frac{1}{T}\right)\right) \tag{3}$$

where $A(\mathbf{r})$ depends only on the relative proportions of each information source.

Proof. See Appendix A.1.

A version of Proposition 0 follows by direct application of Theorems 1 and 4 of Moscarini and Smith (with slight modification to allow for infinite divisibility); however, such would give a $O(T^{1/2})$ remainder. By contrast, I apply a different proof technology—a saddlepoint approximation due to Lugannani & Rice (1980)—to show that the remainder is actually the tighter $O(T^{-1})$.

It's worth emphasizing that, efficiency indices are purely properties of the information sources, not the decision maker's prior or payoffs. Thus, at large enough quantities, all decision makers agree that an additional small δ from $\mathscr{E}_j$ reduces expected losses by a roughly factor of $\rho_{j\mathbf{r}}(D)^{\delta}$.

Note that from Proposition 0 we immediately get a multi-state generalization of the main result of Chernoff (1952) for log-likelihood ratio tests.

Corollary (Chernoff's asymptotic relative efficiency). *If $\mathscr{E}_1$ and $\mathscr{E}_2$ are two information sources with efficiency indices $\rho_1(\cdot)$ and $\rho_2(\cdot)$ respectively, then if quantity t_1 from $\mathscr{E}_1$ has the same expected loss as t_2 from $\mathscr{E}_2$ then, for t_1 large*

$$\frac{t_2}{t_1} \approx \frac{\log(\max_{D\in\mathscr{D}}\{\rho_1(D)\})}{\log(\max_{D\in\mathscr{D}}\{\rho_2(D)\})}$$

The above result illustrates the importance of the *log* efficiency index for substitutability of different sources, but tells us little about preferences away from the corners. Of course, this would be sufficiency to characterize demand if corners were always optimal, but as we'll shortly see, interior solutions are quite common.

11 Moscarini & Smith (2002) don't assume infinite divisibility, so quantity for them is simply number of conditionally i.i.d. samples

12 Say a function, $f(x)$ is $O(x)$ as x goes to zero, if there exists some positive constant function such that for x small enough, $|f(x)| < Cx$

To get a complete consumer theory, we need to take a closer look at the approximation given by Proposition 0.

4 RESULTS

4.1 Precision and information demand

Recall that the DM's objective is to *minimize* the expected loss, but the multiplicative form of Equation (3) begs to be transformed by logs. Under a budget constraint, the DM will equivalently choose her information bundle to maximize $-\log(L(\mathbf{t}))$, so we can use such as "utility" for information. I thus denote $U(\mathbf{t}) \equiv -\log(L(\mathbf{t}))$.

Define the *precision* of an information source as

$$\beta_j(D) \equiv -\log(\rho_j(D))$$

Similarly, define the marginal precision $\beta_{j\mathbf{r}}(D) = -\log(\rho_{j\mathbf{r}}(D))$. Since β is additive for i.i.d. samples and higher for more informative (lower efficiency index) experiments, one can view it as a generalization of the classic notion of precision. In fact, for a standard Gaussian signal which reveals the true state plus mean-zero noise with variance $1/\gamma$—that is, with (classical) precision γ—the generalized precision is proportional to γ.

Using Proposition 0 we can then write utility for information as

$$U(\mathbf{t}) = \min_{D \in D} \left\{ \sum_{j=1}^{J} t_j \beta_{j\mathbf{r}}(D) \right\} \left(1 + O\left(\frac{\log(T)}{T} \right) \right) \tag{4}$$

That is, at large quantities, the DM prefers bundles with higher total precision for their *worst-case* dichotomy. The following proposition guarantees the stronger claim that the maximin precision bundle is close to the true optimal, loss-minimizing one:

Proposition 1 (Maximin precision). *As all costs go to zero ($\varepsilon \to 0$), the loss-minimizing information bundle is $\mathbf{t}^* = \bar{\mathbf{t}}(1 + O(\varepsilon))$ where $\bar{\mathbf{t}}$ maximizes the worst-case total precision, $\min_D \sum t_j \beta_{j\mathbf{r}}(D)$, subject to $\mathbf{c} \cdot \mathbf{t} \leq Y$.*

Proof. Making costs (ε) small is equivalent to making the budget large, so fix sample cost vector $\mathbf{c}$. Let $\mathbf{r}_Y^*$ be the relative proportions of the loss-minimizing bundle at budget Y and $\bar{\mathbf{r}}$ the same for the least-precision per dollar bundle (not necessarily unique). Then we must have

$$\underline{A} \frac{\max_D \left\{ \rho_{\mathbf{r}_Y^*}(D)^{Y/(\mathbf{r}_Y^* \cdot c)} \right\}}{\sqrt{Y/(\mathbf{r}_Y^* \cdot c)}} < \overline{A} \frac{\max_D \left\{ \rho_{\mathbf{r}^*}(D)^{Y/(\mathbf{r}^* \cdot c)} \right\}}{\sqrt{Y/(\mathbf{r}^* \cdot c)}}$$

where $\overline{A} < \infty$ and $\underline{A} > 0$ are upper and lower bounds on $L(\mathbf{t})(\max_D \rho_{\mathbf{r}}(D)^T)^{-1}\sqrt{T}$ for $T \geq 1$ (shown to exist, even when the worst-case dichotomy is non-unique, in Appendix A.2). Taking logs and rearranging terms we have that the precision per dollar at the true optimal proportions approaches the maximal least-precision per dollar at rate $O(Y^{-1})$. Application of Taylor's theorem completes the proof, because precision is differentiable in $\mathbf{r}$ for each dichotomy. Technical details are deferred to Appendix A.2.

In short, Proposition 1 allows us to analyze preferences over information bundles by treating worst-case total precision *as though* it were the DM's utility function.

Interestingly, this implies that our risk-neutral DM behaves as though she were extremely risk-*averse*—choosing her information bundle to minimize the probability of the most likely mistake. Further, because precision is independent of DM specifics, whenever the maxi-min precision is unique, all DMs will agree on the optimal bundle, up to a vanishing (percent) remainder.

By this approach, we can see that information demand will have some peculiar properties following from the unusual nature of least total precision. For example, the following lemma implies low-cost information demand deviates starkly from the benchmark differentiable, quasiconcave utility model:

Lemma 1 (Properties of precision). *For a fixed dichotomy, precision is homothetic and quasiconvex in quantity,* $\mathbf{t}$. *The worst-case total precision is thus homothetic and* locally *quasiconvex*[13] *around any bundle where the least-precision dichotomy is unique.*

Proof. Homotheticity follows immediately: scaling up an information bundle does not change the relative proportions of each information source in the bundle, and thus does not affect marginal precisions. To see quasiconvexity, recall that $\rho_{j\mathbf{r}}(D)$ is the moment-generating function for the D-dichotomy log-likelihood ratio evaluated at the minimizer for the whole bundle. We thus have for a fixed dichotomy, $\prod \rho_{j\mathbf{r}}(D)^{r_j} \geq \prod \rho_j(D)^{r_j}$, or equivalently

$$\sum_{j=1}^{J} r_j \beta_{j\mathbf{r}}(D) \leq \sum_{j=1}^{J} r_j \beta_j(D) \tag{5}$$

(with equality if and only if all sources with $r_j > 0$ have the same log-likelihood ratio moment-generating function minimizer). Put another way, precision of a bundle is less than the sum of its parts for a fixed dichotomy. Minimization does not preserve quasiconvexity, but because there are only finitely-many dichotomies, local behavior is preserved around any point where the worst-case dichotomy is unique (typically almost everywhere).

Lemma 1 has two important implications. First, homotheticity implies optimal proportions, $\mathbf{r}$, are income-independent, so we can equivalently solve the maximin precision problem by finding proportions that maximize the worst-case precision *per dollar*, $\beta_{\mathbf{r}}/\mathbf{c}\cdot\mathbf{r}$. Put another way, at low enough prices, there's no such thing as an inferior or luxury source of information:

Corollary (Unit income elasticity). *If the maximin precision per dollar bundle is unique, the arc*[14] *income elasticity of demand for all information sources given a fixed change in income is* $1 + O(\varepsilon)$ *as costs go to zero.*

Second, the local quasiconvexity imply that most information bundles cannot be optimal with a linear budget. One would be forgiven for thinking (5) simply implies that loss-minimizing bundles must lie near corners; however, because only the least-precision dichotomy matters for big bundles, total *least* precision will be non-quasiconvex whenever different sources have different worst-case dichotomies. Intuitively, interior solutions may arise because distinct

13 Say a function is *locally quasiconvex* at $\mathbf{t}$ if, for a small enough open ball, the intersection of the lower contour set of $\mathbf{t}$ and the ball around $\mathbf{t}$ is convex.

14 Differentiating O terms is only possible under limited circumstances—hence, the use of the arc formula.

information sources can cover for the others' weaknesses. In fact, so long as the information sources differ in their worst case dichotomy, interior solutions may arise even when one source has higher precision for *all* dichotomies (or even more strongly, is Blackwell-dominant).[15]

The nature of these interior solutions is perhaps easiest understood geometrically by treating iso-least-precision curves as though they were the DM's indifference curves. The iso-precision curve for each fixed dichotomy bows out, but the iso-*least*-precision curve (the outer contour) will have inward pointing kinks when the worst-case dichotomy is non-unique (Figure 1).

We thus clearly have that for a generic pair of sources (whose iso-precision lines strictly bow out), information will be consumed in at most finitely many possible ratios corresponding with the corner solutions and kinks where iso-precision lines intersect. Proposition 2 generalizes this observation to generic, finite collections of information sources.

Proposition 2 (Iso-least-precision kinks). *For generic information sources—i.e. ones for which (5) holds strictly—across all costs, there are finitely many relative proportions,* $\mathbf{r}$*, that maximize worst-case precision per dollar,* $\min_D \sum r_j \beta_{j\mathbf{r}}(D)/(\mathbf{c}\cdot\mathbf{r})$*, and at almost all costs, the maximin-precision-per-dollar proportions are unique and invariant to small cost changes.*

Proof (sketch). Almost-everywhere local quasiconvexity guarantees that maximin precision guarantees that maximin precision proportions can only occur on a measure-zero set. Showing that only finitely many proportions are ever optimal is left to Appendix A.3.

Because interior maximin precision bundles are always at iso-least-precision kinks, information from different sources are locally perfect complements. For small enough price changes, the change in demand is purely income effect: information sources are locally perfect complements.

Corollary (Price elasticities). *If the maximin precision bundle is unique (true for almost all cost vectors), the (arc) price elasticity of demand for all sources given a small enough percent change,* δ*, of* c_i *is* $(r_i c_i/\mathbf{r}\cdot\mathbf{c})\,(1+O(\varepsilon+\delta))$

However, around costs where the maximin precision is non-unique, information demand exhibits massive substitution effects as the demands jumps between kinks/corners.

4.2 Complexity of optimal sample bundles

Because iso-precision lines bow out, in the simplest, binary state, decision problems optimal information bundles only have a single information source (in non-vanishing proportions) at low costs. Interior solutions can only occur in environments with at least three possible states. This suggests that "sophisticated" bundles (more distinct sources) require more "complicated" decision problems (more possible states of the world).

In fact, because of the kinked geometry of the maxi-min precision problem, there is a sharp relationship between the number of distinct information sources in a bundle and the number of dichotomies:

Proposition 3. *If proportions* $\mathbf{r}^*$ *maximize the worst-case precision per dollar and has support on* K *distinct generic information sources, then the* $\mathbf{r}^*$*-composite experiment has equal precision for at least (generically, exactly)* K *dichotomies.*

15 See the supplementary material for an example.

Proof (sketch). By Proposition 2, interior maxi-min precision bundles occur at kinks where multiple dichotomies have equal precision. If $\mathbf{r}^*$ has support on K distinct sources, this kink must be K-dimensional, and thus be the intersection of K iso-precision surfaces. See Appendix A.4 for a formal proof.

We can thus put an upper bound on the number of information sources that will ever be consumed in non-vanishing proportions at large budgets: at most as many as there are dichotomies ($|\Theta|(|\Theta|-1)/2$).

4.3 Tradeoffs between sources of information

In the linear-constraint setting, the kinked nature of solutions renders the marginal rate of substitution between sources effectively irrelevant. However, information is a peculiar good, and non-linear costs arise quite naturally. For example, data sources will have a fixed upper bound on available samples (at least in the short term). Further, the non-rival nature of information consumption lends itself to non-linear pricing schemes such as subscriptions and bundling.

In such cases, the maxi-min precision solution might not be feasible, and the marginal rate of substitution might prove useful.

Equation (4) suggests that, provided the worst-case dichotomy is locally unique, the marginal utility for information from a given source ought to be roughly to that source's marginal precision. Thus sources should be roughly substitutable in proportion to their relative marginal precision. Normally, however, we should be cautious of heuristics that rely on differentiating O terms.

Fortunately, a variation of the same saddlepoint method used to prove the expected loss approximation Proposition 0 works equally well to on the derivative of expected loss:

Proposition 4 (Marginal rate of substitution). *At any bundle with unique worst-case dichotomy, D, the marginal rate of substitution between them is*

$$\frac{\partial L/\partial t_1}{\partial L/\partial t_2} = \frac{\beta_{1\mathbf{r}}(D)}{\beta_{2\mathbf{r}}(D)} + O(T^{-1})$$

Proof. See Appendix A.5.

5 HOW WELL DOES THE APPROXIMATION PERFORM IN PRACTICE?

It's worth considering whether any of these approximations are even useful in practice. After all, in principal, the approximation holds only when information purchases are large, and thus when error probabilities are vanishingly small. In such a world, does it really matter whether or not the DM chooses a bundle of information optimally? I claim that, despite this, these approximations are still useful for understanding preferences. To that end, I briefly discuss both the theoretical and numerical performance of these results.

5.1 Theoretical performance

First, recall that these approximations have percent errors falling proportional with $1/T$, a respectable convergence rate by asymptotic approximation standards. By comparison, the central limit theorem—a staple for estimating standard errors in applied work—converges with error falling with rate $1/\sqrt{T}$. Although these approximations are not substitutes, based on

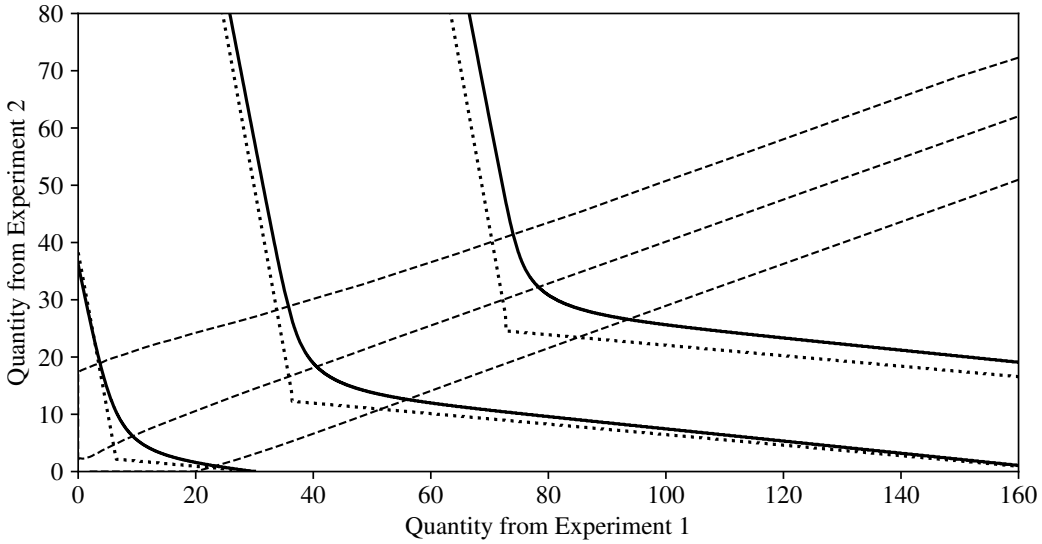

Figure 2: Numerical simulations for a three-state decision problem with two available sources of information. Dotted lines are the iso-least-precision lines, solid lines are true indifference curves, and dashed lines are income expansion paths for three different cost ratios.

convergence rates alone, the large deviations approach should be no more suspect than the central limit theorem one.

Second, from a practical standpoint, even though mistakes are very rare, they are an astronomically more rare when using the maximin precision rule:

Corollary. *Let* $\mathbf{t}^*(Y)$ *be the maximal bundle with proportions* $\mathbf{r}^*$ *maximizing worst-case precision per dollar and* $\mathbf{t}(Y)$ *be the same but for some other fixed proportions* $\mathbf{r}$*. Then the loss satisfies*

$$\lim_{Y \to \infty} \frac{L(\mathbf{t}(Y))}{L(\mathbf{t}^*(Y))} = \infty$$

where Y^* *is the required budget from purchasing information in proportions* $\mathbf{r}^*$ *and* Y *the required budget purchasing in proportions* $\mathbf{r}$*.*

More strongly, this ratio not only grows without bound, but grows exponentially fast. Thus, especially in environments where mistakes are extremely costly, choosing a bundle far from the maximin precision one can have extremely high costs.

5.2 Numerical performance

We can't entirely justify these approximations by theory alone. I conclude with a few observations of the approximation's performance in simulation.[16]

Figure 2 illustrates some of the main features of numerical simulations of information demand.[17]

First, recall that there are two main sources of approximation error: the pure large deviations error, and the error from ignoring all but the most likely mistake. In theory, the error from ignoring less likely mistakes is exponentially falling and thus negligible in comparison

16 Click here to download the Jupyter notebooks.

17 In these simulations, infinite divisibility is achieved by Poissonization. That is, instead of choosing number of samples directly, the DM chooses *expected* samples and then receives a Poisson distributed draw of samples.

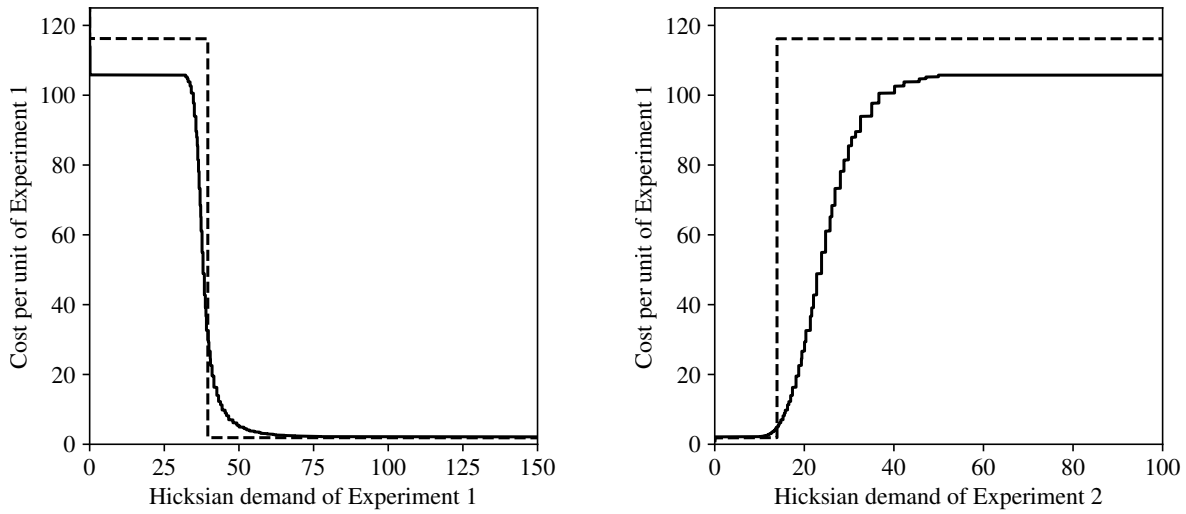

Figure 3: Hicksian demand for the same information sources used in Figure 2. The solid line is a true Hicksian demand curve, and the dashed line is the equivalent curve predicted by the maximin precision rule.

to the $O(T^{-1})$ error of the large deviations approximation. However, near iso-least-precision kinks, the second-most-likely mistake is nearly as likely as the most likely and thus the approximation performs comparatively poorly in those regions.[18] Graphically, this manifests as indifference curves taking a smooth curve rather than the sharp turn the iso-least-precision lines take.

This might seem like an issue given that optimal solutions are predicted to be at kinks, but because the approximation is very tight away from the kink, optimal bundles are still constrained to be near kinks.

The income expansion paths plotted in Figure 2 show this. For high enough budgets, the true optimal bundle is within a constant (and thus vanishing percent error) of the kink. Demand still illustrates the predicted large substitution effects, jumping straight to the corners for steeper cost ratios.

To illustrate these substitution effects, Figure 3 shows the Hicksian[19] demand for information for the two information sources used in Figure 2. Although the true Hicksian demand (solid) is not quite the step function predicted by the maximin precision rule (dashed line), it tracks very closely and illustrates the predicted small substitution effects everywhere except near the jump between kinks.

Finally, notice that even at low budgets when the approximation is poor on the *intensive* margin, the maximin precision approximation typically performs well on the *extensive* margin, correctly predicting the set of sources used in positive quantities.

All told, these results suggest that the approximations perform best when the number of states is relatively small, as many states can generate many kinks and thus may have relatively smooth, quasiconcave. indifference curves at realistic information quantities.

Interested readers can further explore the approximations for different sources of information in the simulations in the supplemental online materials.

18 There is also a second order issue that because kinks point inward, total sample size tends to be lower than it would be at a corner on the same iso-precision-line.

19 Hicksian demand holds expected losses constant rather than budget. I use Hicksian demand here because it holds total information quantity roughly constant, so the quality of the approximation does not vary much over the cost range.

6 CONCLUSION

This paper has provided a general approach for understanding the consumer theory for information in finite-state/finite-action environments. Specifically, using an improved version of the large-deviations approximation of Moscarini & Smith (2002), I have shown that information demand is, up to a percent error vanishing with costs, given by a maximin precision rule.

Notably, in contrast to a purely statistical approach to comparing information sources, my consumer-theoretic approach explicitly allowed consideration of costs and complementarities between different signals, and demonstrated the salience of interior solutions, typically neglected by the relative efficiency literature.

By treating worst-case precision as if it were utility, I was able to describe a variety of consumer-theoretic quantities of interest. In particular, worst-case precision is kinked, and thus information demand behaves as if perfect complements at most costs. Additionally, this approach allowed approximated the substitutability of different information sources whenever the worst-case dichotomy is unique. Finally, the kinked nature of precision implied a novel relationship between the number of information sources used and the number of states: no more information sources will ever be consumed than there are state pairs.

These results ought to have application both to the economic theory of information, but also—more practically—to the optimal design of experiments.

All told, these results suggest caution when dealing with information as good: the standard assumptions (smooth, convex) fail rather spectacularly.

APPENDIX A OMITTED PROOFS

A.1 Proof of Proposition 0

Similar to Moscarini & Smith (2002), I start with the simple hypothesis testing problem with a single information source. 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

$$L(t) = (1-p)\alpha_I(t)L_I + p\alpha_{II}(t)L_{II}$$

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:

$$\alpha_I(n) = \mathbb{P}\left(\log\left(\frac{d\mu^t_{\theta_1}}{d\mu^t_{\theta_0}}(x)\right) > \bar{l} \,\middle|\, \theta_0\right)$$

Moscarini & Smith (2002) apply a classic change-of-measure approach similar to Cramér (1938) to prove their main result for discretely sampled information sources, where the log-likelihood ratio is an i.i.d. sum. In contrast, I use a *saddlepoint* approach which more cleanly applies to infinitely divisible source and 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 an 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*

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

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

We can then approximate the survivor function for the log-likelihood ratio by applying the method of steepest descents to its characteristic function:

Lemma (Lugannani & Rice, 1980). *Let Y be a real-valued random variable with $\mathbb{E}(Y) < 0$, and $\bar{Y}_n$ be the sample average of n i.i.d. draws of Y. If Y's characteristic function $\phi(\zeta)$, analytic through a strip, $\{\zeta : -\mathfrak{I}(\zeta) \in (\zeta^* - \epsilon, \zeta^* + \epsilon)\}$, where $\zeta^* \equiv \arg\min_\zeta \mathbb{E}(e^{\zeta Y})$, then for ξ close enough to zero*

$$\mathbb{P}(\bar{Y}_n \geq \xi) = \frac{e^{n(K(\zeta^*(\xi)) - \zeta^*(\xi)\xi)}}{\zeta^*(\xi)\sqrt{2\pi n K''(\zeta^*(\xi))}}\left(1 + O\left(\frac{1}{n}\right)\right) \tag{6}$$

where $K(\zeta) \equiv \log(M(\zeta))$ *is the cumulant generating function of* Y *and* $\zeta^*(\xi)$ *is the minimizer of* $K(\zeta) - \zeta\xi$.

Note that although Lugannani & Rice work with discretely sampled distributions, they use the method of steepest-descents to approximate the inversion given by the previous lemma, and thus works for any real-valued n, (though ϕ^n will only be a valid characteristic function when infinite divisibility applies).

A reader interested in the logic of the proof should see the proof of Proposition 4 in Appendix A.5 where I derive a formula for the derivative of the mistake probability using the same method.

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

$$\begin{aligned} M(\zeta;\theta_1,\theta_0) &= \int \mu_{\theta_1}(d\mathbf{x})^{\zeta}\mu_{\theta_0}(d\mathbf{x})^{1-\zeta} \\ &= M(1-\zeta;\theta_0,\theta_1) \end{aligned}$$

Because of this symmetry, simplify notation by writing $M(\zeta) \equiv M(\zeta;\theta_1,\theta_0)$. Now, recall we assumed that $M(\zeta;\theta_1,\theta_0)$ is defined on an interval around zero. M must then be infinitely differentiable at $\zeta = 0$ and $\zeta = 1$, and thus so too must be all points between (by dominated convergence and convexity of $e^{\zeta x}$). Because the characteristic function is $\phi(u) = M(-iu)$, 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 $M(0) = M(1) = 1$.

To finish the proof, we need now only let $\xi_t = \bar{l}/t$ and ζ_t^* the minimizer of $K(\xi_t) - \zeta\xi_t$, and apply Taylor's theorem:

Let ζ^* be the minimizer of $M(\zeta)$ (equivalently, of $K(\zeta)$). By applying Taylor's theorem and the FOC, $K'(\zeta^*) = 0$, we can the write

$$\begin{aligned} \zeta_t^* &= = \zeta^* + O(1/t) \\ K(\zeta_t^*) &= K(\zeta^*) + \frac{1}{2t^2}K''(\zeta^*) + O(1/t^3) \\ K''(\zeta_t^*) &= K''(\zeta^*) + O(1/t) \end{aligned}$$

Note that by definition of precision and the efficiency index we have, $K(\zeta^*) = -\beta$ so $e^{tK(\zeta^*)} = \rho^t$. We can then plug each of these into Equation (6) and apply Taylor's theorem again. Breaking it down into parts we have

$$\begin{aligned} e^{tK(\zeta_t^*)} &= \rho^t(1 + O(1/t)) \\ e^{\zeta_t^*\zeta_t^*\xi_t} &= e^{\zeta^*\bar{l}}(1 + O(1/t)) \\ \zeta_t^*\sqrt{2\pi n K''(\zeta_t^*)} &= (\zeta^* + O(1/t))\sqrt{2\pi t K''(\zeta^*) + O(1)} \\ &= \zeta^*\sqrt{2\pi t K''(\zeta^*)}(1 + O(1/t)) \end{aligned}$$

Plugging each of the above parts into Equation (6) we have that the error probability is

$$\alpha_I(t) = \frac{e^{\zeta^*\bar{l}}}{\zeta^*\sqrt{2\pi K''(\zeta^*)}}\frac{\rho^t}{\sqrt{t}}\left(1 + O\left(\frac{1}{t}\right)\right) \tag{7}$$

Repeating this process gives a similar expression for α_{II}. Because $M(\zeta;\theta_1,\theta_0) = M(1-\zeta;\theta_0,\theta_1)$, we need only replace ζ^* with $1-\zeta^*$ and change the cutoff log-likelihood ratio appropriately. Plugging this into the original equation for expected loss gives the claimed result for two-state/two-action decision problems.

Application of Moscarini and Smith's Theorem 4 completes the proof for the general finite-state/finite-action case. *Q.E.D.*

A.2 Omitted Parts of the Proof of Proposition 1

Part 1: Uniform bounds

We need to show that $L(\mathbf{t}(\max_D \rho_{\mathbf{r}}(D)^T)^{-1}\sqrt{T}$ has (finite) upper and (strictly positive) lower bounds uniform over all $\mathbf{r}$, including when the worst-case dichotomy is non-unique. First, write the expected loss as

$$L(\mathbf{t}) = A(\mathbf{r},T)\frac{\max_D\{\rho_{\mathbf{r}}(D)^T\}}{\sqrt{T}}$$

By Proposition 0, we have $A(\mathbf{r},T) = O(1)$ as T gets large for each fixed $\mathbf{r}$ with unique worst-case dichotomy. To show the same for $\mathbf{r}$ with non-unique worst-case dichotomy, we can apply a similar logic to Moscarini and Smith's proof of their Theorem 4.

First note that $L(\mathbf{t})$ is higher than the expected loss if the DM additionally received a signal that perfectly reveals the state unless it's in the worst-case dichotomy. Call this loss $L_D(\mathbf{t})$. Because the loss is positive only when the state is in D, L_D can be written using Proposition 0. To get an upper bound, use Claim 3 of (Moscarini & Smith 2002, p. 2363):

Lemma (Moscarini & Smith (2002) Claim 3). *For an experiment with state-dependent distributions, μ_θ, then for $\varepsilon > 0$ and weights b_θ, the following holds for quantity t (t samples if non-infinitely-divisible)*

$$\mathbb{P}\left(\sum_{\theta\neq\theta_0} b_\theta \frac{\mu_\theta^t}{\mu_{\theta_0}^t}(\mathbf{x}) > \varepsilon \;\middle|\; \theta_0\right) = O\left(\frac{\max_{\theta\neq\theta_0}\{\rho(\theta,\theta_0)^t\}}{\sqrt{t}}\right) \tag{8}$$

Because each mistake probability takes the form of (8), we have that

$$L(\mathbf{t}) = O(\max_D\{\rho_{\mathbf{r}}(D)^T\}/\sqrt{T})$$

Putting everything together, we have that for some constants, $A_1 > 0$ and $A_2 < \infty$

$$A_1\frac{\max_D\{\rho_{\mathbf{r}}(D)^T\}}{\sqrt{T}} \leq L(\mathbf{t}) \leq A_2\frac{\max_D\{\rho_{\mathbf{r}}(D)^T\}}{\sqrt{T}}$$

so $A(\mathbf{r},t) = O(1)$ even if $\mathbf{r}$ has non-unique worst-case dichotomy.

We now want to show that there are A_1 and A_2 such that the above holds *uniformly* for all $\mathbf{r}$ and T bounded away from zero. Because the space of sample proportions is compact, it suffices to show that every open neighborhood has a finite bound.

To prove the upper bound, suppose otherwise—i.e. that every neighborhood around some $\mathbf{r}_0$ has no upper bound. Then for any α and any $\delta > 0$, we can find T such that $A(\mathbf{r},T) > \alpha$ for some $\mathbf{r}$ within δ of $\mathbf{r}_0$. But then we have a contradiction, because we can choose α as high we like—in particular we can choose $\alpha > \sup_T A(\mathbf{r},T)$—so for any δ no matter how small, we can

find $A(\mathbf{r}_0 + \delta, T) > \sup_T A(\mathbf{r}_0, T)$ which violates continuity of A (both L and ρ are continuous, so A must be as well). By similar reasoning, we can guarantee a uniform, strictly positive lower bound.

Part 2: Convergent precision implies convergent proportions

To formally complete the proof, we must justify the claim that precision per dollar approaching the optimum at rate $O(Y^{-1})$ implies that the relative proportions of optimal demand $\mathbf{r}_Y^*$ approach the relative proportions of the maximin precision bundle $\bar{\mathbf{r}}$ (not necessarily unique).

From before, we have that the least-precision per dollar of the loss-minimizing sample bundle is within $O(Y^{-1})$ of the maximum:

$$\frac{\min_D\{\beta_{\mathbf{r}_Y^*}(D)\}}{\mathbf{r}_Y \cdot c} - \frac{\min_D\{\beta_{\bar{\mathbf{r}}}(D)\}}{\bar{\mathbf{r}} \cdot c} < O\left(\frac{1}{Y}\right)$$

It remains to show that $\mathbf{r}_Y^*$ is within $O(Y^{-1})$ of $\bar{\mathbf{r}}$—i.e., that

$$\mathbf{r}_Y^* \in \left\{\bar{\mathbf{r}} + O(Y^{-1}) \,:\, \bar{\mathbf{r}} \in \arg\min_{\mathbf{r}} \left\{\min_D \beta_{\mathbf{r}}(D)/(\mathbf{r}\cdot\mathbf{c})\right\}\right\} \tag{9}$$

Because precision per dollar is differentiable for each dichotomy, we must have by Taylor's theorem that for some element, $\bar{\mathbf{r}}$, of the argmax of worst-case precision:

$$\begin{aligned}\min_D\left\{\frac{\beta_{\mathbf{r}_Y^*}(D)}{\mathbf{r}_Y^*\cdot\mathbf{c}}\right\} &= \min_D\left\{\frac{\beta_{\bar{\mathbf{r}}}(D)}{\bar{\mathbf{r}}\cdot\mathbf{c}} + O(\bar{\mathbf{r}} - \mathbf{r}_Y^*)\right\}\\ &= \min_D\left\{\frac{\beta_{\bar{\mathbf{r}}}(D)}{\bar{\mathbf{r}}\cdot\mathbf{c}}\right\} + O(\bar{\mathbf{r}} - \mathbf{r}_Y^*)\end{aligned}$$

But from before we had that

$$\min_D\left\{\frac{\beta_{\mathbf{r}_Y^*}(D)}{\mathbf{r}_Y^*\cdot\mathbf{c}}\right\} = \min_D\left\{\frac{\beta_{\bar{\mathbf{r}}}(D)}{\bar{\mathbf{r}}\cdot\mathbf{c}}\right\} + O\left(\frac{1}{Y}\right)$$

We thus must have that $\bar{\mathbf{r}} - \mathbf{r}_Y^* = O(Y^{-1})$ as required.[20] *Q.E.D.*

A.3 Proof of Proposition 2

Consider the dual, cost-minimization problem: choose $\mathbf{t}$ to minimize total costs, such that the total precision for each dichotomy is at least B.

First suppose $\mathbf{t}^*$ solves this problem. I claim that no other point has the same set of binding constraints (including non-negativity constraints). To see this, suppose the same constraints bind at $\mathbf{t}'$. By construction, we must then have $\mathbf{c}\cdot\mathbf{t}^* \leq \mathbf{c}\cdot\mathbf{t}'$.

Now consider the point $\mathbf{t}_\lambda = \lambda\mathbf{t}^* + (1-\lambda)\mathbf{t}'$ for $\lambda > 1$. For λ close enough to 1, the constraints slack at $\mathbf{t}^*$ remain slack, but by strict convexity of precision for generic sets of experiments, the previously binding precision constraints must become slack (binding non-negativity constraints still bind). But notice that the total cost of $\mathbf{t}_\lambda$ is at most as much as that

20 Note that when the argmax is non-unique, $\mathbf{r}_Y^*$ may not converge, but its accumulation points will be a subset of the worst-case precision argmax.

of $\mathbf{t}^*$. We then have a contradiction: $\mathbf{t}^*$ cannot be cost minimizing because we can find a *strictly* lower cost bundle satisfying all constraints by consuming ε less than $\mathbf{t}_\lambda$.

Because there are finitely many constraints, there are finitely many combinations of constraints, and thus the set of sample bundles that are ever cost-minimizing for a given precision level is finite.

Finally, because precision is homothetic, the cost-minimizing sample proportions are independent of the target precision level. Thus, there equally must be finitely many possible sample proportions that ever solve the primal problem. *Q.E.D.*

A.4 Proof of Proposition 3

It suffices to show that in an environment with J available experiments, if $\mathbf{t}^*$ maximizes precision for a given budget and cost vector, then the number of dichotomies with equal precision plus the number of binding non-negativity constraints equals J.

Consider the collection of surfaces defined by binding non-negativity constraints or tangent to an iso-precision line on the outer contour at $\mathbf{t}^*$. Suppose for contradiction that the number of dichotomies with equal precision plus binding non-negativity constraints at $\mathbf{t}^*$ is strictly less than J, and thus that there are fewer than J of such surfaces. These surfaces intersect at $\mathbf{t}^*$ by construction, but also along a lower-dimensional affine surface, S. By construction, S is tangent to all iso-precision lines on the outer contour at $\mathbf{t}^*$.

Now recall that precision is strictly convex for each dichotomy for generic experiments, so by moving along S, we can increase the precision for all dichotomies that had equal precision at $\mathbf{t}^*$. Further, for $\mathbf{t}$ close enough to $\mathbf{t}^*$ the iso-precision lines on the outer contour will be a subset of those on the outer contour at $\mathbf{t}^*$. Finally this S intersects the budget line at $\mathbf{t}^*$ so there must be bundles other than $\mathbf{t}^*$ on S with cost at most that of $\mathbf{t}^*$. But then we have a contradiction because there are cheaper bundles with higher least-precision close to $\mathbf{t}^*$. *Q.E.D.*

A.5 Proof of Proposition 4

Similar to the proof of Proposition 0, I first show the result for a two-state/two-action world with a single source of information.

Recall that in this case we can write the mistake probability as

$$\alpha_I(t) = \mathbb{P}\left(l + s_t > \bar{l} \,|\, \theta_0\right) = \mathbb{P}\left(\frac{s_t}{t} > \frac{\bar{l} - l}{t} \,\middle|\, \theta_0\right)$$

Ignoring issues of differentiability for now, we can apply Leibniz's rule to the characteristic function inversion formula to write the mistake probability as

$$\frac{\partial \alpha_I(t)}{\partial t} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-iu\xi_n} \log(\phi(u))\phi(u/t)^t}{iu} du$$

where $\phi(u/t)^t$ is the characteristic function of the sample average log-likelihood ratio, and $\xi_t = (\bar{l} - l)/t$. Changing variables $iu = v$, we can rewrite this in terms of the cumulant generating function $K(\zeta) = M(\zeta)$:

$$\frac{\partial \alpha_I(t)}{\partial t} = \frac{1}{2\pi} \lim_{L \to \infty} \int_{-iL+c}^{iL+c} \frac{K(v/t) e^{t(K(v/t) - v\xi_t)}}{iv} dv \tag{10}$$

where, by Cauchy's theorem, c can be whatever I like provided v remains in the analytic strip of ϕ. In particular, I can pick $c = \bar{\zeta}_t$ where $\bar{\zeta}_t = \arg\min\{K(\zeta/t) - \zeta\xi_t\}$ so the path of integration passes through the real line at the minimizer of the exponential term along the reals. Further, by the Cauchy-Riemann equations, a critical point of an analytic function must be a saddlepoint, thus, although $\bar{\zeta}_t$ corresponds with a minimum travelling along the real line, it corresponds with the *maximum* travelling along the path of integration perpendicular to the real line.

We can then apply a Laplace approximation to the previously given integral—that is, for t large the value of the integral is dominated by the region around $\bar{\zeta}_t$.

Changing variables again, we have

$$\frac{\partial \alpha_I(t)}{\partial t} = \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} \frac{K(v/t) e^{t(K(\bar{\zeta}_t + ix) - (\bar{\zeta}_t + ix)\xi_t)}}{\bar{\zeta}_t + ix} dx$$

Because the right-hand side of (10) is real, it will be useful first to split up the integrand into its real and imaginary parts, cancelling any imaginary parts. Simplifying notation by writing $g(x) \equiv K(x) - x\xi_t$ write

$$\begin{aligned}
\frac{\partial \alpha_I(t)}{\partial t} &= \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} K(\bar{\zeta}_t + ix) \frac{e^{t\Re(g(\bar{\zeta}_t + ix))}(\bar{\zeta}_n \cos(t\Im(g(\bar{\zeta}_t + ix))) + x \sin(t\Im(g(\bar{\zeta}_t + ix))))}{\bar{\zeta}_t + ix} dx \\
&= \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} K(\bar{\zeta}_t + ix) \frac{e^{t\Re(g(\bar{\zeta}_t + ix))}(\bar{\zeta}_n \cos(t\Im(g(\bar{\zeta}_t + ix))) + x \sin(t\Im(g(\bar{\zeta}_t + ix))))}{\bar{\zeta}_t^2 - x^2} dx \\
&= \frac{1}{2\pi} \lim_{L\Rightarrow\infty} \int_{-L}^{L} e^{t\hat{g}(x)} h(x) dx
\end{aligned}$$

where the first line follows from Euler's formula, the second from multiplying the numerator and denominator by $\bar{\zeta}_t - ix$ and cancelling the purely imaginary terms, and the third from appropriately defining $\hat{g}$ and h to simplify the remaining terms.

What remains is a purely real-valued integrand and can thus be attacked with standard real analysis.

Now applying the Laplace approximation—effectively approximating the integral by a Gaussian locally around the max of $\hat{g}$ (normalized to $x = 0$ above)—we can write[21]

$$\frac{1}{2\pi} \lim_{L\Rightarrow\infty} \int_{-L}^{L} e^{t\hat{g}(x)} h(x) dx = e^{t\hat{g}(0)} \frac{1}{-n\hat{g}''(0)\sqrt{2\pi}} h(0)(1 + O(t^{-1}))$$

To complete the approximation, we need only use the definitions of h and $\hat{g}$:

$$\begin{aligned}
\hat{g}(0) &= K(\bar{\zeta}_t) - \bar{\zeta}_t \xi_t \\
-\hat{g}''(0) &= K''(\bar{\zeta}_t) \\
h(0) &= \frac{K(\bar{\zeta}_t)}{\bar{\zeta}_t}
\end{aligned}$$

where the first and third follow by definition and the second follows by the Cauchy-Riemann equations because the second derivative of $\hat{g}$ at zero is along the complex axis, thus the second derivative along the real axis is sign-flipped.

21 See Ch. 17 in Jeffreys & Jeffreys (1956)

Plugging all this in, we get a formula similar to the original formula of Lugannani & Rice (1980) from Appendix A.1.

$$\frac{\partial \alpha_I(t)}{\partial t} = K(\bar{\zeta}_t)\frac{e^{n(K(\bar{\zeta}_t)-\bar{\zeta}_t\bar{\xi}_t)}}{\bar{\zeta}_t\sqrt{2\pi n K''(\bar{\zeta}_t)}}\left(1 + O\left(\frac{1}{n}\right)\right)$$

Lastly, we need only repeat the Taylor approximation method used in the proof of subsection A.1 to get

$$\frac{\partial \alpha_I(t)}{\partial t} = -\beta\alpha_I(n)(1 + O(t^{-1}))$$

which implies that for a two-state/two-action environment, the marginal loss is $\partial L/\partial t = -\beta L(t)(1 + O(t^t))$ From this we can immediately derive the same formula for the many state case:

$$\frac{\partial L(\mathbf{t})}{\partial t_j} = -\beta_{j\mathbf{r}} L(\mathbf{t})(1 + O(T^{-1}))$$

Q.E.D.

Typically in these settings, quantity of information is measured in conditionally i.i.d. *samples*, which are fundamentally discrete.

Because the log-likelihood ratio of n draws from a given experiment is simply the sum of log-likelihood ratios, we still have that the log-likelihood ratio moment-generating function is $M(\zeta)^n$. The only difference is that, in general, $M(\zeta)^n$ is a valid moment-generating function only for whole-numbered n, whereas for infinitely-divisible sources all positive powers were valid moment-generating functions.

We can thus define total worst-case precision exactly as we did before:

$$B(n_1, \dots, n_J) \equiv \min_D \left\{ \max_\zeta \left\{ -\sum_{j=1}^{J} n_j \log(M(\zeta; D)) \right\} \right\}$$

Notice, however, that the above function is well-defined for any real n, so we can still draw iso-precision lines in $\mathbb{R}^J$, even though true indifference curves are generically singletons defined only on $\mathbb{N}^J$. We can thus find a maximin precision "bundle" exactly as we did before, though now it may not correspond with any actual available bundle of information.

All propositions except Proposition 4 (which fundamentally depends on differentiability) thus apply. Propositions 2 and 3 are state in terms of maximin precision bundle and thus need no modification to their proofs. The proof of Proposition 1 needs only be modified slightly since the true optimal bundle may not use the entire budget, though it can't differ by more than the cost of the cheapest source.

At high budgets, the proportions of the true optimal bundle must be within $O(Y^{-1})$ of proportions that would make the budget constraint bind, so this merely adds another $O(Y^{-1})$ term leaving the overall result unchanged.

To generalize Proposition 4, requires a bit more work, as we first need to define what we even mean by *marginal rate of substitution* in a discrete setting.

B.1 A discrete version of Proposition 4

Without infinite divisibility, indifference sets are typically singletons, so we there's no exact rate at which samples can be substituted. Instead, we can look at minimum compensating substitutions—that is, what is the minimum number of $\mathscr{E}_2$ samples to at least compensate for a loss of k samples from $\mathscr{E}_1$.

Proposition 4A (Sample substitutability). *Consider a sample bundle with sample proportions* $\mathbf{r}$ *and unique least-precision dichotomy* $D_{\mathbf{r}}$*. Then the minimum number of additional samples,* k_2*, of* $\mathscr{E}_2$ *to compensate for a loss of* k_1 *samples of* $\mathscr{E}_1$ *is, for* N *high enough, exactly*

$$\left\lceil k_1 \frac{\beta_{1\mathbf{r}}(D_{\mathbf{r}})}{\beta_{2\mathbf{r}}(D_{\mathbf{r}})} \right\rceil$$

Proof. Fix $\mathbf{r}$ such that the worst-case dichotomy is unique. Since the worst-case dichotomy will always be the same throughout the proof, I suppress 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

$$\begin{aligned} n_1\beta_{1\mathbf{r}} + n_2\beta_{2\mathbf{r}} + \sum_{j=3}^{J} n_j\beta_{j\mathbf{r}} + \log(A(\mathbf{r})) + O(N^{-1}) \\ \leq (n_1 - k_1)\beta_{1\mathbf{r}'} + (n_2 + k_2)\beta_{2\mathbf{r}'} + \sum_{j=3}^{J} n_j\beta_{j\mathbf{r}'} + \log(A(\mathbf{r}')) \end{aligned} \tag{11}$$

where $\mathbf{r}'$ is the composite factor associated with the new sample bundle. Start with N high enough that this substitution doesn't change the worst-case dichotomy. Then notice that for this fixed size substitution $\mathbf{r}$ doesn't change much. Specifically, $\mathbf{r}' - \mathbf{r} = O(N^{-1})$. Applying this fact with Taylor's theorem to the FOC for $\beta_{\mathbf{r}}$, we have that $\tau_{\mathbf{r}'} - \tau_{\mathbf{r}} = O(N^{-1})$ as well. We can then apply Taylor's theorem (remember precision is defined on the reals, even for non-divisible experiments) to write

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

Further, because $A(\mathbf{r})$ is a differentiable function of $\tau_{\mathbf{r}}$ (see Equation (7) in the last part of the proof of Proposition 0), we have that $\log(A(\mathbf{r}')) - \log(A(\mathbf{r})) = O(N^{-1})$. Plugging all of this into (11) and rearranging gives

$$k_2 \geq k_1\frac{\beta_{1\mathbf{r}}}{\beta_{2\mathbf{r}}} + O(N^{-1}) \tag{12}$$

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

$$k_2 \leq k_1\frac{\beta_{1\mathbf{r}}}{\beta_{2\mathbf{r}}} + 1 + O(N^{-1}) \tag{13}$$

Together, by squeezing k_2 between (12) and (13) we have for N large enough

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

as claimed. *Q.E.D.*

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.

Feller, W. (1970), *An Introduction to Probability Theory and Its Applications, Vol. II*, 2nd edn, John Wiley & Sons.

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.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer New York.

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.

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

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