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 develops a consumer theory of information when information is cheap. I propose a generalized notion of precision—the negative-log of the efficiency index used by Moscarini and Smith (2002)—that measures how well an experiment 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 to costs. Additionally, I show iso-least-precision sets have finitely many kinks and are otherwise locally quasi-convex and consider implications for information demand in a linear constraint setting. Finally, I derive an upper bound, quadratic in the number of possible states, for the number of sources ever used in non-vanishing proportions: at most as many as there are state pairs.

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 such 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. Unfortunately, such an exercise presupposes an understanding of preferences—no mean feat when information values are notoriously ill-behaved.\(^1\) This paper seeks to provide an answer for the finite-state/finite-action world by developing an approximate consumer theory for information, valid when information is cheap (or equivalently when budgets large).

\(^1\) Most famously, the value of information often has non-concave regions. See, for example, (Radner and Stiglitz, 1984) and (Chade and Schlee, 2002).

* Email: gary.baker@wisc.edu
Figure 1: The geometry of iso-precision lines in a three-state environment with two available information sources. Iso-precision lines for a fixed pair of states (dashed lines) bow out, but the iso-least-precision line (solid line) has inward pointing kinks. Demand for information behaves as though indifference curves were similarly shaped.

My approach is inspired by Moscarini and Smith (2002) who apply a large-deviations method to develop an approximation for sample demand from a single information source in a quasilinear setting. I extend their approximation for a (discrete) sample demand in a single-source quasilinear setting to a multi-source continuous setting and explore implications for the general consumer theory problem.

In particular, using this approximation, 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 information 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 prices.

Using this result, I then explore properties of demand by treating maximin precision as if it were a utility function. 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 it’s 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.

The kinked geometry additionally implies an upper bound, quadratic in the number of states, on the number of information sources any decision maker will use in non-vanishing proportions at low costs (Proposition 3). That is, at low costs, only at most as many information sources as there are dichotomies are ever used. Thus, in the simplest, two-state hypothesis testing world, only corners are ever optimal, with interior solutions only occurring with more complex decision problems.

\[ \text{Results for discrete sample demand are deferred to Appendix B.} \]
Finally, 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 aforementioned work of Moscarini and Smith (2002), and my Proposition 0 generalizes their main result to a setting with multiple available experiments and refines 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.

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. Finally, Section 5 examines the performance of the approximations numerically, and Section 6 concludes.

2 MODEL

2.1 The underlying decision problem

A decision-maker (DM) must choose an action \( a \) from a finite set, \( A \), under an uncertain state of the world \( \theta \) drawn from a finite set, \( \Theta \). 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/experiments, \( \mathcal{E}_1, \ldots, \mathcal{E}_J \). Each information source is a conditionally independent Blackwell experiment, \( \mathcal{E}_j = (\mathcal{X}, \langle \mu_{j\theta} \rangle_{\theta \in \Theta}) \), consisting of a collection of state-dependent distributions over some arbitrary space of realizations, \( \mathcal{X} \). 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 mutually absolutely continuous and thus have defined likelihood ratios (Radon-Nikodym derivatives) \( \frac{d\mu_{j\theta}}{d\mu_{j\theta'}} \).

Under this assumption, the \( \mathbb{R}^{|\Theta|-1} \) vector of log-likelihood ratios relative to some base state, \( \theta_0 \), is a sufficient statistic for each realization. We can thus without loss identify each realization by its own vector of log-likelihood ratios:

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

We thus can summarize conditionally independent realizations from multiple experiments 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.
2.2 Quantity of information

The DM will choose not just which information sources to use, but also how much to consume from each. Conditionally i.i.d. samples are the usual measure of quantity in this context; however, because samples are fundamentally discrete, I introduce a notion of information quantity that preserves many of the natural properties of (discrete) samples while allowing a standard (continuous) consumer theory treatment.

Say an experiment, \( \mathcal{E}_j \), is infinitely divisible if for any \( k \), there exists an experiment \( \mathcal{E}_{j}^{1/k} \) such that \( k \) conditionally i.i.d. samples from \( \mathcal{E}_{j}^{1/k} \) are equivalent to a single sample from \( \mathcal{E}_j \). Such experiments naturally admit a notion of fractional “samples.”

Thus, a rational quantity \( t_j = a/b \) of information from \( \mathcal{E}_j \) is equivalent to \( a \) conditionally i.i.d. samples from \( \mathcal{E}_{j}^{1/b} \). Formally, the state-\( \theta \) log-likelihood ratio distribution of rational quantity \( t = n/k \) of \( \mathcal{E}_j \), \( \mu_{\theta}^t \), is \( \star \prod_{i=1}^n \mu_{\theta}^{1/k} \), i.e. the \( n \)-fold convolution\(^3\) of the log-likelihood ratio distribution of \( \mathcal{E}_{j}^{1/k} \). Non-rational quantities are simply an appropriate limit—for example, the sum of the appropriate number of samples from \( \mathcal{E}_{j}^{1/10} \), \( \mathcal{E}_{j}^{1/100} \), and so on. 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 log-likelihood ratio moment-generating function of \( \mu_{\theta} \) is \( M_{\theta}(\zeta) \), then the moment-generating function of \( \mu_{\theta}^t \) is \( M_{\theta}(\zeta)^t \).

We can equivalently view quantity of information as a time spent observing a continuous-time, state-dependent process with conditionally independent increments.\(^4\) Given the ubiquity of time constraints, in my budget-constrained setting, time is perhaps the most natural interpretation; however, note the model is formally static: the DM chooses quantity, then observe the realizations all at once.

Example 1: Gaussian Signals Let \( \Theta \subset \mathbb{R} \) and \( \mathcal{E} \) reports the true state \( \theta \) plus \( \mathcal{N}(0, \sigma^2) \) noise. Because the sum of Gaussian distributions is Gaussian, \( k \) samples is equivalent to a single Gaussian with noise variance \( \sigma^2/k \). To wit, quantity of information is equivalent to choosing the signal’s precision. Equivalently, quantity in the Gaussian case can be viewed as the time to observe a Brownian motion with state-dependent drift (see, for example Keppo et al., 2008).

Example 2: Compound Poisson Signals Given any information source, we can always generate an infinitely divisible analog by Poissonizing it. That is, instead of choosing sample size directly, the decision maker chooses an expected number of samples and then receives a number of samples drawn from the appropriate distribution. Infinitely divisibility holds because the sum of Poisson draws is itself Poisson. An equivalent information source arises if the DM chooses how long to wait for sample draws that arrive according to a Poisson arrival process.

Compound Poisson signals admit easy generalization of many of the methods for discretely sampled experiments, and all infinitely divisible are almost compound Poisson—formally, all infinitely divisible experiments are weak limits of Poissonized ones (Ch. 9, Prop. 3 Le Cam, 1986). For this reason, I assume experiments are compound Poisson for Proposition 4 to simplify the proof; however all other proofs apply for infinitely divisible experiments more gen-

---

\(^3\) The convolution of two distributions is the distribution of their sum, i.e. 
\[
(\mu_1 * \mu_2)(S) = \int 1_S(x + y)\mu_1(dx)\mu_2(dy)
\]

\(^4\) This equivalence is guaranteed by Theorem 2.IX.5 of Feller (1970).
erally.

Note that I assume infinite divisibility merely to use standard consumer theory ideas such as indifference curves for the sake of exposition. Appendix B generalizes all result to the standard discretely sampled world.

2.3 The information consumer theory problem

The DM chooses a non-negative quantity, \( t = (t_1, ..., t_J) \in \mathbb{R}_+^J \), of information from each information source at costs, \( c = (c_1, c_2, ..., c_J) > 0 \) per unit quantity from each source subject to budget constraint, \( Y \).

After choosing the expected sample bundle, the DM observes the realizations, Bayes updates her beliefs appropriately, and chooses an action to maximize her expected payoff. The DM thus wants to choose the feasible bundle that minimizes her expected loss\(^5\) from acting after observing the realizations:

\[
\text{minimize } L(t) = \sum_{\theta} p_0 \int_{x \in \Theta_{\theta}} (u(a^*(\theta), \theta) - u(a(x), \theta)) \mu_0^t(dx) \tag{1}
\]

subject to \( c \cdot t \leq Y \)

where \( a^*(\theta) \) is the optimal action if the state were known to be \( \theta \), \( a(x) \) is the expected-payoff-maximizing decision after observing realized log-likelihood ratios \( x \), and \( \mu_0^t \) is the log-likelihood ratio distribution of realizations for the chosen expected sample bundle, \( \mu_0^t \equiv \star_{j=1}^J \mu_{j0}^t \).

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 (\( a = R \)) or accept (\( a = A \)) the null. Naturally, the DM wants to reject when the null is false and vice-versa, so the payoffs satisfy \( u(R, \theta_1) > u(A, \theta_1) \) and \( u(A, \theta_0) > u(R, \theta_0) \). Assume the DM has prior \( p \) that the alternative (\( \theta_1 \)) is true, and the belief that makes the DM exactly indifferent between the two actions as \( p \).

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

\[
L(t) = (1 - p)\alpha_l(t)(u(A, \theta_0) - u(R, \theta_0)) + p\alpha_{II}(t)(u(R, \theta_1) - u(A, \theta_1))
\]

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

\(^5\) Equivalent to maximizing the usual value of information in a budget-constrained setting.
Bayes’s rule is a sum when written in terms of log-likelihood ratios:

\[ \alpha_I(t) = P \left( I + s_t > \tilde{l} | \theta_0 \right) \]  

(2)

where \( I = \log(p/(1-p)) \) is the prior log-likelihood ratio, \( \tilde{l} = \log(\bar{p}/(1-\bar{p})) \) 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 \( \theta_0 \) is true must be negative—i.e. when \( \theta_0 \) is true, on average the realizations should push the log-likelihood posterior down towards stronger beliefs that \( \theta_0 \) is true.

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

\[ \alpha_I(t) = P \left( \frac{s_t}{t} > \frac{\tilde{l} - l}{t} | \theta_0 \right) \]

Here we can see why we can’t use a more familiar asymptotic approach such as a central limit theorem: \( E(s/t) = \bar{s} < 0 \) 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 deviation 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:

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

where \( M \) is the state-\( \theta_0 \) 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 \(^6\) of information source \( \mathcal{E}_j \) for the \( \theta, \theta' \) dichotomy as the minimized value of the moment-generating function for the \( \{\theta, \theta'\} \) log-likelihood ratio conditional on true state \( \theta' \):

\[ \rho_j(\{\theta, \theta'\}) \equiv \min_{\zeta} M_j(\zeta ; \{\theta, \theta'\}) = \min_{\zeta} \left\{ \int_{x \in \mathbb{R}^{|\theta|}} \mu_{\theta}(dx)^\zeta \mu_{\theta'}(dx)^{1-\zeta} \right\} \]

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.\(^7\)

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 quan-

---

\(^6\) I follow Moscarini and Smith (2002) here. Torgersen (1991) refers to this as the Chernoff number.

\(^7\) If \( \mathcal{E}_1 \) Blackwell dominates \( \mathcal{E}_2 \) then \( \mathcal{E}_1 \) has lower efficiency indices for all dichotomies.
tity, \( T \equiv \sum t_j \) and the proportions of total quantity from each experiment as \( r = (r_1, \ldots, r_J) \equiv (t_1/T, \ldots, t_j/T) \).

Then define the \( r \)-composite experiment as one with quantities \( r \) from each information source. This constructed source then has log-likelihood rate moment-generating functions, 
\[
M_r(\zeta; \{\theta, \theta'\}) \equiv \prod M_j(\zeta; \{\theta, \theta'\})^{r_j}.
\]
By construction, quantity \( T \) from the \( r \)-composite experiment is equivalent to the bundle \( t \) because 
\[
M_T = \prod M_j^{t_j}.
\]
Define the \( r \)-composite efficiency index 
\[
\rho_r(\{\theta, \theta'\}) \equiv \min_{\zeta} \left\{ \prod_{j=1}^J M_j(\zeta; \{\theta, \theta'\})^{r_j} \right\}
\]
We can then further break \( r \) into contributions from each source by defining the marginal efficiency index of \( \mathcal{E}_j \) as 
\[
\rho_{r_j}(\{\theta, \theta'\}) \equiv M_j(\zeta^*_r; \{\theta, \theta'\}),
\]
so 
\[
\rho_r(\{\theta, \theta'\}) = \prod \rho_{r_j}(\{\theta, \theta'\})
\]
With only two states, Moscarini and Smith (2002) show that each mistake probability, and thus the expected loss itself, is proportional to \( \rho' / \sqrt{T} \) for \( T \) large. 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 \( \mathcal{D} \) be the collection of dichotomies. Then, when the worst-case dichotomy, 
\[
\arg \max_D \rho_r(D),
\]
is unique, the expected loss from consuming quantities \( t = [t_1, \ldots, t_j] \) from \( \mathcal{E}_1, \ldots, \mathcal{E}_J \) is
\[
L(t) = A(r) \frac{\max_{D \in \mathcal{D}} \left\{ \prod_{j=1}^J \rho_{r_j}(D)^{t_j} \right\}}{\sqrt{T}} \left( 1 + O\left( \frac{1}{T} \right) \right)
\]
where \( A(r) \) 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 and 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 \( \delta \) from \( \mathcal{E}_j \) reduces expected losses by roughly factor of \( \rho_{r_j}(D)^{\delta} \).

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

---

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

9 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, \( |f(x)| < Cx \)
Corollary (Chernoff’s asymptotic relative efficiency). If $\mathcal{C}_1$ and $\mathcal{C}_2$ are two information sources with efficiency indices $\rho_1(\cdot)$ and $\rho_2(\cdot)$ respectively, then if quantity $t_1$ from $\mathcal{C}_1$ has the same expected loss as $t_2$ from $\mathcal{C}_2$ then, for $t_1$ large

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

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 sufficient to characterize demand if corners were always optimal, but as we’ll shortly see, interior solutions are quite common.

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 $-\log(L(t))$, so we can use such as “utility” for information. I thus denote $U(t) \equiv -\log(L(t))$.

Define the precision of an information source as $\beta_j(D) \equiv -\log(\rho_j(D))$.

Example 1: Gaussian Signals A signal with realizations $\theta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2/t)$, has precision $(1/8)(\theta_1 - \theta_0)^2 t/\sigma^2$ for the $\{\theta_1, \theta_0\}$ dichotomy. Thus, for Gaussian signals, precision in my generalized sense corresponds with the natural measure of precision for Gaussians: the signal-to-noise ratio.

Example 2: Compound Poisson Signals Suppose $\mathcal{E}$ is a Poissonization of $\hat{\mathcal{E}}$ with efficiency index $\hat{\rho}$ for a single sample. Then the efficiency index of quantity $t$ of $\hat{\mathcal{E}}$ ($t$ expected samples from $\hat{\mathcal{E}}$) is simply the expected efficiency index of the realized draw: $\sum_k \hat{\rho}^k e^{t^k/k!} = \exp(t(\hat{\rho} - 1))$. Hence, the precision of a compound Poisson signal is $t(1 - \hat{\rho})$. Notice that $1 - \hat{\rho} < -\log(\hat{\rho})$, so $t$ samples in expectation has a lower precision than the same number of samples for certain. Intuitively, at large samples, the expected loss is convex, so randomizing over the number of samples is less preferred.

Similarly, define the marginal precision $\beta_{jr}(D) = -\log(\rho_{jr}(D))$. Since $\beta$ 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. Using Proposition 0, we can then write utility for information in terms of the worst-case total precision:

$$U(t) = \min_{D \in \mathcal{D}} \left\{ \sum_{j=1}^{I} t_j \beta_{jr}(D) \right\} \left( 1 + O\left( \frac{\log(T)}{T} \right) \right)$$ (4)

That is, at large quantities, the DM prefers bundles with higher total precision for their worst-case dichotomy. More strongly, maximizing worst-case precision yields an approxima-
tion for the utility-maximizing (loss-minimizing) information bundle whose percent error van-
ishes with costs:

**Proposition 1 (Maximin precision).** Let costs be \( c = (\varepsilon, \varepsilon\kappa_1, \ldots, \varepsilon\kappa_J) \). As \( \varepsilon \) goes to 0, if \( t^* \) maximizes utility, \( U \), (minimizes expected losses), subject to \( c \cdot t \leq Y \), then \( t^* = t(1 + O(\varepsilon)) \) where \( t \) maximizes the worst-case total precision, \( \min_D \sum n_j \beta(D) \), subject to \( c \cdot t \leq Y \).

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

\[
\frac{\max_D \left\{ \rho(D)^Y / (\bar{r} \cdot c) \right\}}{\sqrt{Y / (\bar{r} \cdot c)}} \leq \frac{\max_D \left\{ \rho(D)^Y / (r^*_Y \cdot c) \right\}}{\sqrt{Y / (r^*_Y \cdot c)}} \leq \frac{\max_D \left\{ \rho(D)^Y / (r \cdot c) \right\}}{\sqrt{Y / (r \cdot c)}}
\]

where \( A < \infty \) and \( A > 0 \) are upper and lower bounds, uniform in across all values of \( r \), on \( L(t)(\max_D \rho(D)^T)^{-1/2} \sqrt{T} \) 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 \( r \) 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 the 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, \( t \). The worst-case total precision is thus homothetic and locally quasiconvex\(^{10} \) (typically, strictly so) 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 precision. To see quasiconvexity, recall that \( \rho(D) \) 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, \( \prod \rho(D)^{r_j} \geq \prod \rho(D)^{r_j} \), or equivalently

\[
\sum_{j=1}^{J} r_j \beta(D) \leq \sum_{j=1}^{J} r_j \beta_j(D) \tag{5}
\]

---

\(^{10}\) Say a function is *locally quasiconvex* at \( t \) if, for a small enough open ball, the intersection of the lower contour set of \( t \) and the ball around \( t \) is convex.
(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, \( r \), are income-independent, so we can equivalently solve the maximin precision problem by finding proportions that maximize the worst-case precision per dollar, \( \beta_r / c \cdot r \). Put another way, at low enough prices, there’s no such thing as an inferior or luxury source of information:

**Corollary 1** (Unit income elasticity). If the maximin precision per dollar bundle is unique, the arc\(^{11}\) income elasticity of demand for all information sources given a fixed change in income is \( 1 + O(\epsilon) \) as costs \( (\epsilon) \) 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 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).\(^{12}\)

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, 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, \( r \), that maximize worst-case precision per dollar, \( \min_D \sum r_j \beta_f(D)/(c \cdot r) \), 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.

Thus, for small enough price changes, the change in demand is purely income effect: information sources are locally perfect complements.

---

\(^{11}\) The arc formula avoids technical issues around differentiating an approximation and applies equally in the discrete sampling environment.

\(^{12}\) See the supplementary material for an example.
Corollary 2 (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, $\delta$, of $c_i$ is

$$\frac{\partial c_i}{\partial r} \cdot c_i \cdot (1 + O(\varepsilon + \delta))$$

However, around costs where the maximin precision is non-unique, information demand exhibits massive substitution effects as the demands jumps between kinks/corners. Put another way, Hicksian demand for information should be approximately a step function.

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. For generic information sources, the proportions, $r^*$ maximizing worst-case precision per dollar have support on at most $|\Theta|(|\Theta| - 1)/2$ (i.e. the number of dichotomies) distinct information sources.

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

Thus, if the true optimal information bundle has support on more information sources, the excess sources must have make up a vanishing proportion of the total bundle.

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, sources are roughly substitutable in proportion to their relative marginal precision. The following proposition confirms this intuition:

Proposition 4 (Marginal rate of substitution). Let $\xi_1$ and $\xi_2$ be compound Poisson experiments.\footnote{\textsuperscript{13} Compound Poisson experiments are assumed for technical simplicity. The general result can be proved using a variation of the method in Appendix A.1, but is highly technical for $|\Theta| > 2$.} At any bundle with unique worst-case dichotomy, $D$, the marginal rate of substitution

\textsuperscript{13} Compound Poisson experiments are assumed for technical simplicity. The general result can be proved using a variation of the method in Appendix A.1, but is highly technical for $|\Theta| > 2$.\textsuperscript{13}
between them is
\[
\frac{\partial L}{\partial t_1} = \frac{\beta_1(r(D))}{\beta_2(r)} + O(T^{-1})
\]

Proof. Let \( \hat{\mathcal{E}}_1 \) be a Poissonization of \( \hat{\mathcal{E}}_1 \). Then we can write the expected loss from quantity \( t_1 \) from \( \mathcal{E}_1 \) (\( t_1 \) expected samples from \( \hat{\mathcal{E}}_1 \)) by averaging the loss from each possible draw of samples from \( \hat{\mathcal{E}}_1 \):

\[
L(t_1, t_2, \ldots, t_J) = \sum_{k=0}^{\infty} \frac{t_1^k e^{t_1}}{k!} \hat{L}(k, t_2, \ldots, t_J)
\]

where \( \hat{L}(k, t_2, \ldots, t_J) \) is the expected loss, but with \( k \) samples from \( \hat{\mathcal{E}}_1 \) replacing quantity \( t_1 \) from \( \mathcal{E}_1 \). Because expected losses are bounded, by a standard dominated convergence argument, \( L \) is differentiable\(^\text{14}\) in \( t_1 \):

\[
\frac{\partial L(t_1, t_2, \ldots, t_J)}{\partial t_1} = \sum_{k=0}^{\infty} \frac{t_1^k e^{t_1}}{k!} \left( \hat{L}(k + 1, t_2, \ldots, t_J) - \hat{L}(k, t_2, \ldots, t_J) \right)
\]

In words, \( \partial L / \partial t_1 \) is the reduction in expected loss from one for-sure extra sample from \( \hat{\mathcal{E}}_1 \). Because this single additional sample of \( \hat{\mathcal{E}}_1 \) only affects the marginal efficiency index up to a \( O(T^{-1}) \) term, we can apply Proposition 0 to write

\[
\frac{\partial L(t_1, t_2, \ldots, t_J)}{\partial t_1} = (\hat{\rho}_1(D) - 1)A(r)\left(\frac{t_1\prod_{j=2}^{J} \rho_j(D)}{\sqrt{T}}\right)^{\frac{1}{2}} \left( 1 + O\left(\frac{1}{T}\right) \right)
\]

where \( \hat{\rho}_1 \) is the marginal efficiency index of \( \hat{\mathcal{E}}_1 \) and \( D_r \) is the (unique at \( r \)) worst-case dichotomy. As discussed in section 4.1, the precision of a compound Poisson experiment is one less the efficiency index of the underlying experiment, so

\[
\frac{\partial L(t_1, t_2, \ldots, t_J)}{\partial t_1} = -\beta_1(D)L(t_1, t_2, \ldots, t_J) \left( 1 + O\left(\frac{1}{T}\right) \right)
\]

Completing the proof requires only application of the definition of marginal rate of substitution.

Note that in contrast to the relative approximations seen thus far, the marginal rate of substitution approximation has an arbitrarily small absolute error. Similarly, in a discrete setting this approach yields an exact-at-large-samples expression for the number of samples from one experiment required to compensate for a loss of a sample from another. See Appendix B for a more thorough discussion of the substitutability of samples in the general non-infinitely divisible setting.

\(^{14}\) Note that the expected loss is differentiable everywhere for compound Poisson experiments, even though demand behaves as though preferences are non-differentiable at sample ratios where the worst-case dichotomy is non-unique.
5 NUMERICAL PERFORMANCE

From a purely theoretical perspective, the $O(T^{-1})$ convergence rate of the large deviations approximations is quite good. By comparison, central limit theorems—a staple for estimating standard errors in applied settings—converges at rate $O(T^{-1/2})$. Although these approximations apply in different settings and are thus not substitutes, the large deviations approximation should in principle be no more suspect than a central limit theorem one.

Nonetheless, $O$ merely implies that some (possibly very large) $R > 0$ exists such that the error is eventually smaller than $R/T$. To justify practical application, we must rely on numerical simulation. Because numerical analysis of the underlying large-deviations approximations can be found in multiple sources (e.g. Lugannani and Rice, 1980), I will focus on numerical analysis of the consumer theory implications.\footnote{In these simulations, infinite divisibility is achieved by Poissonization.}

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 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. 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 illustrate the behavior of demand for price ratios where the maximin precision bundle occurs at a kink. For high enough budgets, the true optimal bundle is within a constant (and thus vanishing percent error) of the kink.
Figure 3: Illustration of substitution effects for information: Hicksian demand for the same information sources used in Figure 2 as a function of the cost of Experiment 1, holding the cost of Experiment 2 fixed. The solid line is a true Hicksian demand curve, and the dashed line is the equivalent curve predicted by the maximin precision rule.

Even with these smooth curves around the kinks, demand still exhibits the predicted large substitution effects, jumping straight to the corners for steeper cost ratios. To illustrate this, I plot in Figure 3 a Hicksian (compensated)\textsuperscript{16} demand for information for the two information sources used in Figure 2. I use Hicksian, rather than the usual Marshallian, demand for two reasons: (1) the Hicksian demand predicted by the maximin precision rule takes a very simple form (a step function) and it is comparatively easy to visually evaluate the approximation quality, and (2) by holding expected losses fixed, rather than budget, the quality of the approximation does not vary too much over the range of costs because total quantity does not vary much over the cost range as it would with a Marshallian demand curve. 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, it’s worth noting 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.

Put together, 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.

6 CONCLUSION AND FUTURE RESEARCH

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 and Smith (2002), I have shown that infor-

\textsuperscript{16} Hicksian demand holds expected losses constant rather than budget and thus illustrates demand absent any income effects.
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, the 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 OMMITTED PROOFS

A.1 Proof of Proposition 0

Similar to Moscarini and Smith (2002), I start with the simple hypothesis testing problem with a single information source. That is, we have two states, $\theta_0$ and $\theta_1$, with prior that $\theta_1$ is true given by $p$, and two actions, accept and reject, where reject is optimal when $\theta_1$ 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 $\theta_0$ is true) and Type-II errors respectively, and $\alpha_I$ and $\alpha_{II}$ 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 that the posterior log-likelihood ratio is above the rejection threshold when the true state is $\theta_0$:

$$\alpha_I(t) = P \left( \log \frac{d\mu_{\theta_1}}{d\mu_{\theta_0}}(x) > \bar{l} \mid \theta_0 \right)$$

Moscarini and 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 and 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 and 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 $\phi$, then the survivor function of $Y$ is

$$P(Y \geq y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iyu} \phi(u) \frac{du}{iu}$$

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 and Rice, 1980).** Let $Y$ be a real-valued random variable with $\mathbb{E}(Y) < 0$, and $\hat{Y}_t$ be the sample average of $t$ i.i.d. draws of $Y$. If $Y$’s characteristic function $\phi(\zeta)$, analytic through a strip, $\{\zeta : -\Im(\zeta) \in (\zeta^* - \epsilon, \zeta^* + \epsilon)\}$, where $\zeta^* \equiv \arg\min_{\zeta} \mathbb{E}(e^{\zeta Y})$, then for $\zeta$ close enough to zero

$$P(\hat{Y}_t \geq \zeta) = e^{t(K(\zeta^*)(\zeta) - \zeta \zeta^*)} \frac{1}{\zeta^*(\zeta) \sqrt{2\pi t K''(\zeta^*)}} \left(1 + O\left(\frac{1}{t}\right)\right)$$

(7)

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

Note that although Lugannani and Rice work with discretely sampled distributions, they use the method of steepest descents to approximate the inversion of $M(i\zeta)$ given by the previous lemma. Thus, because the method of steepest descents does not require integer powers, (7) holds in the infinitely divisible setting as well.

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

$$M(\zeta; \theta_1, \theta_0) = \int \mu_{\theta_1}(dx)^\zeta \mu_{\theta_0}(dx)^{1-\zeta}$$

$$= M(1 - \zeta; \theta_0, \theta_1)$$

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 $\phi$ 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 $\xi_t = \tilde{1}/t$ and $\zeta_t^*$ the minimizer of $K(\xi_t) - \zeta \zeta_t^*$, and apply Taylor’s theorem:

---

17 $M(\zeta)$ is guaranteed to be a valid moment-generating function only for positive integer $t$ for non-infinitely divisible distributions.
Let \( \zeta^* \) be the minimizer of \( M(\zeta) \) (equivalently, of \( K(\zeta) \)). By applying Taylor’s theorem and the FOC, \( K'(\zeta^*) = 0 \), we can write

\[
\zeta_t^* = \zeta^* + O(1/t)
\]

\[
K(\zeta_t^*) = K(\zeta^*) + \frac{1}{2t^2} K''(\zeta^*) + O(1/t^3)
\]

\[
K''(\zeta_t^*) = K''(\zeta^*) + O(1/t)
\]

Note that by definition of precision and the efficiency index we have, \( K(\zeta^*) = -\beta \) so \( e^{K(\zeta^*)} = \rho^l \). We can then plug each of these into Equation (7) and apply Taylor’s theorem again. Breaking it down into parts we have

\[
e^{tK(\zeta_t^*)} = \rho^l (1 + O(1/t))
\]

\[
e^{\zeta_t \zeta^*} = e^{\zeta^*} (1 + O(1/t))
\]

\[
\frac{\zeta_t \sqrt{2\pi t K''(\zeta_t^*)}}{\sqrt{2\pi t K''(\zeta^*)}} = (\zeta^* + O(1/t)) \sqrt{2\pi t K''(\zeta^*)} + O(1)
\]

\[
= \zeta^* \sqrt{2\pi t K''(\zeta^*)} (1 + O(1/t))
\]

Plugging each of the above parts into Equation (7) we have that the error probability is

\[
\alpha_l(t) = \frac{e^{\zeta^*} \rho^l}{\zeta^* \sqrt{2\pi K''(\zeta^*)}} \sqrt{I} \left(1 + O\left(\frac{1}{t}\right)\right)
\]

(8)

Repeating this process gives a similar expression for \( \alpha_h \). Because \( M(\zeta; \theta_1, \theta_0) = M(1 - \zeta; \theta_0, \theta_1) \), we need only replace \( \zeta^* \) 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 \( L(t)(\max_D \rho_T(D)^T)^{-1/2} \) has (finite) upper and (strictly positive) lower bounds uniform over all \( r \), including when the worst-case dichotomy is non-unique. First, write the expected loss as

\[
L(t) = A(r, T) \frac{\max_D \{\rho_T(D)^T\}}{\sqrt{T}}
\]

By Proposition 0, we have \( A(r, T) = \Omega(1) \) for each fixed \( r \) with unique worst-case dichotomy. To show the same for \( r \) 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 \( L(t) \) 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 \( L_D(t) \). 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 and Smith, 2002, p. 2363):

17
Lemma (Moscarini and Smith (2002) Claim 3). For an experiment with state-dependent distributions, \( \mu_0 \), then for \( \varepsilon > 0 \) and weights \( b_0 \), the following holds for quantity \( t \) (\( t \) samples if non-infinitly-divisible)

\[
P\left( \sum_{\theta \neq 0} b_0 \frac{\mu_0'(x)}{\mu_0'(\theta)} > \varepsilon \mid \theta_0 \right) = O\left( \frac{\max_{\theta \neq 0} \rho(\theta, \theta_0)^T}{\sqrt{t}} \right)
\]

(9)

Because each mistake probability takes the form of (9), we have that

\[
L(t) = O\left( \max_D \rho_T(D)^T / \sqrt{T} \right)
\]

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

\[
A_1 \frac{\max_D \rho_T(D)^T}{\sqrt{T}} \leq L(t) \leq A_2 \frac{\max_D \rho_T(D)^T}{\sqrt{T}}
\]

so \( A(r, t) = O(1) \) even if \( r \) 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 \( r \) 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 \( r_0 \) has no upper bound. Then for any \( \alpha \) and any \( \delta > 0 \), we can find \( T \) such that \( A(r, T) > \alpha \) for some \( r \) within \( \delta \) of \( r_0 \). But then we have a contradiction, because we can choose \( \alpha \) as high as we like—in particular we can choose \( \alpha > \sup_T A(r, T) \)—so for any \( \delta \) no matter how small, we can find \( A(r_0 + \delta, T) > \sup_T A(r_0, T) \) which violates continuity of \( A \) (both \( L \) and \( \rho \) 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 \( r^*_Y \) approach the relative proportions of the maximin precision bundle \( \bar{r} \) (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:

\[
\frac{\min_D \beta_T^r(D)}{r_Y \cdot c} - \frac{\min_D \beta_T^r(D)}{\bar{r} \cdot c} < O\left( \frac{1}{Y} \right)
\]

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

\[
r^*_Y \in \left\{ \bar{r} + O(Y^{-1}) : \bar{r} \in \arg \min \left\{ \min_D \beta_T^r(D)/(r \cdot c) \right\} \right\}
\]

(10)

Because precision per dollar is differentiable for each dichotomy, we must have by Taylor’s
theorem that for some element, \( \tilde{r} \), of the argmax of worst-case precision:

\[
\min_{D} \left\{ \frac{\beta_Y(D)}{r_Y \cdot c} \right\} = \min_{D} \left\{ \frac{\beta_Y(D)}{\tilde{r} \cdot c} + O(\tilde{r} - r_Y^*) \right\} = \min_{D} \left\{ \frac{\beta_Y(D)}{\tilde{r} \cdot c} \right\} + O(\tilde{r} - r_Y^*)
\]

But from before we had that

\[
\min_{D} \left\{ \frac{\beta_Y(D)}{r_Y \cdot c} \right\} = \min_{D} \left\{ \frac{\beta_Y(D)}{\tilde{r} \cdot c} \right\} + O\left( \frac{1}{Y} \right)
\]

We thus must have that \( \tilde{r} - r_Y^* = O(Y^{-1}) \) as required.\(^{18}\)

Q.E.D.

A.3 Proof of Proposition 2

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

First suppose \( t^* \) 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 \( t' \). By construction, we must then have \( c \cdot t^* \leq c \cdot t' \).

Now consider the point \( t_\lambda = \lambda t^* + (1 - \lambda)t' \) for \( \lambda > 1 \). For \( \lambda \) close enough to 1, the constraints slack at \( t^* \) 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 \( t_\lambda \) is at most as much as that of \( t^* \). We then have a contradiction: \( t^* \) cannot be cost minimizing because we can find a strictly lower cost bundle satisfying all constraints by consuming \( \epsilon \) less than \( t_\lambda \).

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 \( t^* \) 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 \( t^* \). Suppose for contradiction that the number of dichotomies with equal precision plus binding non-negativity constraints at \( t^* \) is strictly less than \( J \), and thus that there are fewer than \( J \) of such surfaces. These surfaces intersect at \( t^* \) 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 \( t^* \).

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

\(^{18}\) Note that when the argmax is non-unique, \( r_Y^* \) may not converge, but its accumulation points will be a subset of the worst-case precision argmax.
at $t^*$. Further, for $t$ close enough to $t^*$ the iso-precision lines on the outer contour will be a subset of those on the outer contour at $t^*$. Finally, this $S$ intersects the budget line at $t^*$, so there must be bundles other than $t^*$ on $S$ with cost at most that of $t^*$. But then we have a contradiction because there are cheaper bundles with higher least-precision close to $t^*$. Q.E.D.

APPENDIX B  DISCRETE SAMPLING

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:

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

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 stated in terms of maximin precision bundle and thus need no modification to their proofs. The proof of Proposition 1 only needs to 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. Nonetheless, at high budgets, the proportions of the true optimal bundle must be within $O(\gamma^{-1})$ of proportions that would make the budget constraint bind, so this merely adds another $O(\gamma^{-1})$ term.

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 $\mathcal{E}_2$ samples to at least compensate for a loss of $k_1$ samples from $\mathcal{E}_1$.

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

$$\begin{bmatrix} k_1 \beta_{1r}(D_r) \\ k_2 \beta_{2r}(D_r) \end{bmatrix}$$
Proof. Fix $r$ 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 $\mathcal{E}_2$ that just compensates for a loss of $k_1$ samples from $\mathcal{E}_1$. Then we have

\[
n_1 \beta_{1r} + n_2 \beta_{2r} + \sum_{j=3}^{J} n_j \beta_{jr} + \log(A(r)) + O(N^{-1}) \leq (n_1 - k_1) \beta_{1r'} + (n_2 + k_2) \beta_{2r'} + \sum_{j=3}^{J} n_j \beta_{jr'} + \log(A(r'))
\]

(11)

where $r'$ 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 $r$ doesn’t change much. Specifically, $r' - r = O(N^{-1})$. Applying this fact with Taylor’s theorem to the FOC for $\beta_r$, we have that $\tau_{r'} - \tau_r = O(N^{-1})$ as well. We can then apply Taylor’s theorem (remember precision is defined on the reals, even for non-divisible experiments) to write

\[
(n_1 - k_1) \beta_{1r'} + (n_2 + k_2) \beta_{2r'} + \sum_{j=3}^{J} n_j \beta_{jr'}
\]

\[
= (n_1 - k_1) \beta_{1r} + (n_2 + k_2) \beta_{2r} + \sum_{j=3}^{J} n_j \beta_{jr} + \left[ k_2 \frac{M'_{2r}(\tau_r)}{M_{2r}(\tau_r)} - k_1 \frac{M'_{1r}(\tau_r)}{M_{1r}(\tau_r)} \right] (\tau_{r'} - \tau_r) + O((\tau_{r'} - \tau_r)^2)
\]

Further, because $A(r)$ is a differentiable function of $\tau_r$ (see (8) in the last part of the proof of Proposition 0), we have that $\log(A(r')) - \log(A(r)) = O(N^{-1})$. Plugging all of this into (11) and rearranging gives

\[
k_2 \geq k_1 \frac{\beta_{1r}}{\beta_{2r}} + O(N^{-1})
\]

(12)

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

\[
k_2 \leq k_1 \frac{\beta_{1r}}{\beta_{2r}} + 1 + O(N^{-1})
\]

(13)

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

\[
k_2 = \left[ k_1 \frac{\beta_{1r}}{\beta_{2r}} \right]
\]

as claimed. Q.E.D.
BIBLIOGRAPHY


