CONSUMER THEORY FOR CHEAP INFORMATION

By

Gary Baker

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: 17 May 2022

The dissertation is approved by the following members of the Final Oral Committee:

Lones Smith, Professor, Economics

Daniel Quint, Associate Professor, Economics

Marek Weretka, Associate Professor, Economics

Dmitry Orlov, Assistant Professor, Finance

ACKNOWLEDGMENTS

First, I thank my advisor, Lones Smith, for providing mentorship and guidance throughout the past eight years. I further thank Daniel Quint and Marek Weretka for their support and advice throughout the process.

Additionally, this dissertation was only possible with the support of my peers, particularly Moshi Alam, Amrita Kulka, Dennis McWeeny, and Elan Segarra.

Finally, I'm thankful to my parents for their support and patience throughout my (unexpectedly long) studies.

ABSTRACT

This dissertation consists of two chapters on the approximating the demand for information from multiple sources.

Chapter 1 develops a consumer theory for multiple cheap sources of information in a finite-action/finite-state setting. I show that demand for information in this setting is well approximated by a *precision* maximization problem. Isoprecision curves—the approximate indifference curves—are the upper envelope of finitely many quasiconvex curves and hence exhibit kinks. Hicksian demand for information is thus approximately locally constant and discontinuous. 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.

Chapter 2 generalizes the Chapter 1 to a class of monotone decision problems with one-dimensional continuous states. I show that the decision maker's expected loss is approximately the loss she would have from taking the optimal action implied by the maximum likelihood estimate, and thus is a weighted average of the reciprocal of the Fisher information. In contrast to the discrete case, approximate indifference curves are smooth and exhibit monotonically increasing rates of substitution. Large sample demand in such settings this behaves much closer to the benchmark consumer theory model, in contrast to the results of Chapter 1.

Contents

List of Figures

CHAPTER ONE

CHEAP INFORMATION CONSUMER THEORY FOR FINITE-STATE DECISION PROBLEMS

ABSTRACT

Classic comparisons—e.g. Blackwell efficiency—of information sources tell us little about trade-offs between different sources, especially when they differ in cost. This chapter develops a consumer theory for multiple cheap sources of information for finite-state decision problems, building on Moscarini and Smith (2002). I show that demand for information in finite-state decision problems is well approximated by a *precision* maximization problem. Isoprecision curves—the approximate indifference curves—are the upper envelope of finitely many quasiconvex curves and hence exhibit kinks. Hicksian demand for information is thus approximately locally constant and discontinuous. 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

1.1 INTRODUCTION

Often a decision-maker wishes to acquire information prior to making a decision under uncertainty and must not only decide how much information to purchase, but also from where to acquire it. For example, a newsreader must decide which news sites to read, or a researcher must decide which research designs to use. With the explosive growth of the internet and the availability of near-unlimited quantities of information from highly varied and distinct sources, questions such as these are more relevant than ever.

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 chapter provides an answer for the finite-state/finite-action world by developing an *approximate* consumer theory for information, valid when information is cheap.

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[2] setting and explore implications for the multiple-good consumer theory problem.

In particular, 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 1.1). Furthermore, because precision is a purely statistical property, independent of prior and payoffs, all decision-makers roughly agree on the optimal

0 This chapter has been adapted from my job market paper by the same name as this dissertation.

1 Most famously, the value of information is typically non-concave. See, for example, (Radner and Stiglitz, 1984) and (Chade and Schlee, 2002).

2 Results for discrete sample demand are deferred to Appendix 1.B.

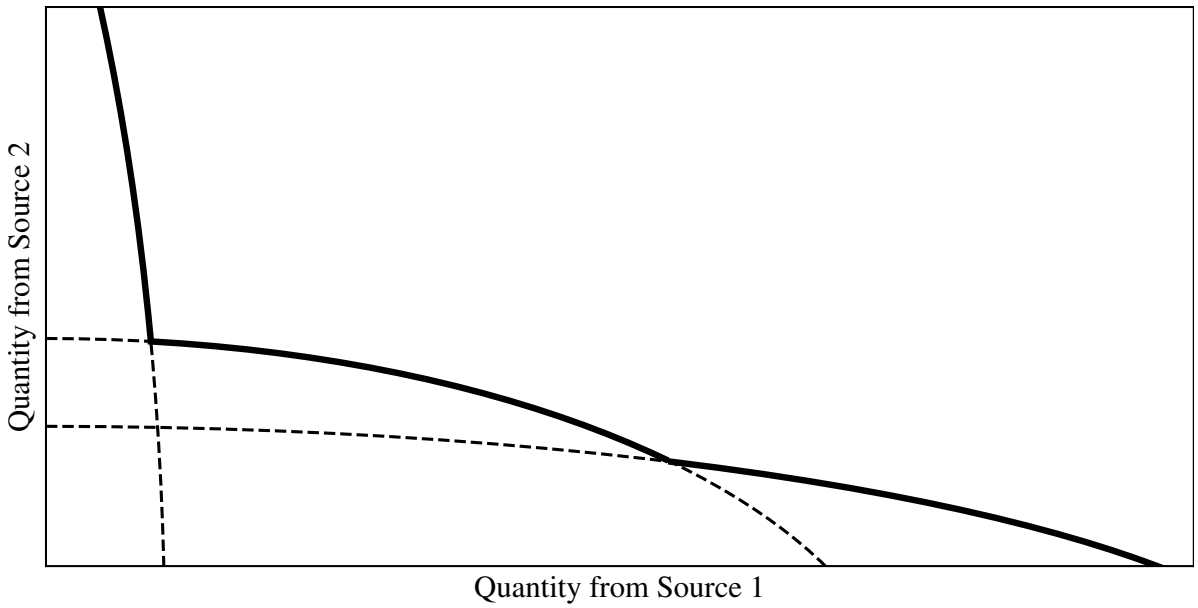

Figure 1.1: The geometry of isoprecision lines in a three-state environment with two available information sources. Isoprecision lines for a fixed pair of states (dashed lines) bow out, but the iso-least-precision line (solid line) has inward pointing kinks. Demand for information behaves as though indifference curves were so shaped.

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 isoprecision lines bow *out*—that is, precision of a composite source is less than the sum of its parts—so iso-*least*-precision lines exhibit inward-pointing kinks (Figure 1.1). Maximin precision bundles thus only occur in finitely-many possible relative proportions (Proposition 1.2), corresponding with iso-least-precision corners and kinks. Demand for information thus 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 1.3). Specifically, optimal bundles at low costs consist of at most as many information sources as dichotomies. Thus, in the simplest, two-state hypothesis testing world, only corners are ever optimal, with interior solutions only occurring with more complex

decision problems.

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 1.4).

The work in this chapter is most closely related to Moscarini and Smith (2002), and my Proposition 1.0 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 chapter is structured as follows: Section 1.2 lays out the formal model assumptions, including the relevant notion of information "quantity" used throughout this chapter. Section 1.3 covers the necessary large-deviations background. Section 1.4 defines precision, states the maximin approximation rule, and explores features of the implied consumer theory. Finally, Section 1.5 examines the numerical performance of the approximations, and Section 1.6 concludes. Except where otherwise noted, all proofs are deferred to the appendix.

1.2 MODEL

1.2.1 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 θ drawn from a finite set, Θ. The DM has Bernoulli utility $u(a, \theta)$ and a full-support prior p over Θ. 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 information sources, $\mathscr{E}_1, \dots, \mathscr{E}_J$. Each information source is a conditionally independent Blackwell experiment, $\mathscr{E}_j = \langle \mathbb{X}, \langle \mu_{j\theta} \rangle_{\theta \in \Theta} \rangle$: a collection of state-dependent distributions over an arbitrary space of realizations, $\mathbb{X}$. Each information source is imperfectly informative. More specifically, $\mu_{j\theta}$ and $\mu_{j\theta'}$ are mutually absolutely continuous and thus have finite likelihood ratios $d\mu_{j\theta}/d\mu_{j\theta'}$.

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

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

Assume all log-likelihood ratio distributions are thin-tailed, with finite moment-generating functions on an open set containing the origin.

1.2.2 Quantity of information

The DM chooses not only the information sources to use, but also *how much* to consume from each. I introduce a notion of information quantity that preserves many of the natural properties of (discrete) conditionally independent sample sizes while allowing a standard (continuous) consumer theory treatment.

Say an experiment, $\mathscr{E}_j$, is *infinitely divisible* if for any k, there exists an experiment $\mathscr{E}_j^{1/k}$ such that k conditionally i.i.d. samples from $\mathscr{E}_j^{1/k}$ are Blackwell equivalent to a single sample from $\mathscr{E}_j$. Such experiments naturally admit a notion of fractional "samples."

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[3] of the log-likelihood ratio distribution of $\mathscr{E}^{1/k}$. Non-rational quantities are simply an appropriate limit. Because the moment-generating

3 The convolution of two distributions is the distribution of their sum, i.e.

$$(\mu_1 \star \mu_2)(S) \equiv \int \mathbf{1}_S(x+y)\mu_1(dx)\mu_2(dy)$$

function of a sum is simply the product of moment-generating functions, the quantity-t conditional distributions of log-likelihood ratios can also be described in terms of their moment-generating functions: if the log-likelihood ratio 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.[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 $\mathscr{E}$ report the true state θ 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 σ^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 observing 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).[5]

4 By Theorem 2.IX.5 of Feller (1970), infinitely divisible distributions are equivalent to continuous processes with i.i.d. increments.

5 Proposition 1.4 assumes compound Poisson experiments to simplify the proof; however, all other proofs apply for infinitely divisible experiments more generally.

I assume infinite divisibility merely to use standard consumer theory ideas such as indifference curves for the sake of exposition. Appendix 1.B generalizes all results to the standard discretely sampled world.

1.2.3 The information consumer theory problem

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

After choosing her information 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 relative to perfect information[6] from acting after observing the realizations:

$$\min_{\mathbf{t}\geq 0} \; L(\mathbf{t}) = \sum_{\theta} p_\theta \int_{\mathbf{x}\in\mathbb{R}^{|\Theta|-1}} \big(u(a^*(\theta),\theta) - u(a(\mathbf{x}),\theta)\big)\mu^{\mathbf{t}}_\theta(d\mathbf{x}) \quad \text{subject to } \; \mathbf{c}\cdot\mathbf{t} \leq Y \tag{1.1}$$

where $a^*(\theta)$ is the optimal action in state θ, $a(\mathbf{x})$ is the expected-payoff-maximizing decision after observing realized log-likelihood ratios $\mathbf{x}$, and $\mu^{\mathbf{t}}_\theta$ is the log-likelihood ratio distribution of realizations for the chosen expected sample bundle, $\mu^{\mathbf{t}}_\theta \equiv \bigstar_{j=1}^{J} \mu_{j\theta}^{t_j}$.

Except under restrictive functional-form assumptions on the available information sources, $L(\mathbf{t})$ has no convenient closed form, necessitating the application of a large-sample approximation.

1.3 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.

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

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 and 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:

$$\alpha_I(t) = \mathbb{P}\left(l + s_t > \bar{l} \,|\, \theta_0\right) \tag{1.2}$$

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 (1.2) in terms of the sample average log-likelihood 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 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:

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

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*[7] of information source $\mathscr{E}_j$ for the θ, θ' dichotomy as the minimized value of the moment-generating function for the $\{\theta, \theta'\}$ 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.[8]

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,

$$M_{\mathbf{r}}(\zeta; \{\theta, \theta'\}) \equiv \prod M_j(\zeta; \{\theta, \theta'\})^{r_j}$$

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

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

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.

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

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 and Smith (2002) show that each mistake probability, and thus the expected loss itself, is proportional to $\rho^t/\sqrt{t}$ for t large.[9] 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 1.0. *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* $\mathbf{t} = [t_1, \ldots, t_J]$ *from* $\mathscr{E}_1, \ldots, \mathscr{E}_J$ *is*[10]

$$L(\mathbf{t}) = A(\mathbf{r}) \frac{\max_{D \in \mathscr{D}} \left\{ \prod_{j=1}^{J} \rho_{j\mathbf{r}}(D)^{t_j} \right\}}{\sqrt{T}} \left(1 + O\left(\frac{1}{T}\right)\right) \tag{1.3}$$

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

Proof. See Appendix 1.A.1.

A version of Proposition 1.0 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

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

10 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, $|f(x)| < Cx$

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 $O(T^{-1})$.

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

Note that from Proposition 1.0 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 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 1.0.

1.4 RESULTS

1.4.1 Precision and information demand

Recall that the DM's objective is to *minimize* the expected loss, but the multiplicative form of Equation (1.3) 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 $\beta_j(D) \equiv -\log(\rho_j(D))$.

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 $\mathscr{E}$ is a Poissonization of $\hat{\mathscr{E}}$ with efficiency index $\hat{\rho}$ for a single sample. Then the efficiency index of quantity t of $\mathscr{E}$ (t expected samples from $\hat{\mathscr{E}}$) is simply the expected efficiency index of the realized draw: $\sum_k \hat{\rho}^k e^t t^k/k! = \exp(t(\hat{\rho} - 1))$. 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 $\beta_{j\mathbf{r}}(D) = -\log(\rho_{j\mathbf{r}}(D))$. 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. Using Proposition 1.0, we can then write utility for information in terms of the *worst-case* total precision:

$$U(\mathbf{t}) = \underbrace{\min_{D \in \mathscr{D}} \left\{ \sum_{j=1}^{J} t_j \beta_{j\mathbf{r}}(D) \right\}}_{\text{worst-case total precision}} \left(1 + O\left(\frac{\log(T)}{T}\right)\right) \tag{1.4}$$

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 approximation for the utility-maximizing (loss-minimizing) information bundle whose percent error vanishes with costs:

Proposition 1.1 (Maximin precision). *Let costs be* $\mathbf{c} = (\varepsilon, \varepsilon\kappa_2, \ldots, \varepsilon\kappa_J)$. *As* ε *goes to* 0, *if* $\mathbf{t}^*$ *maximizes*

utility, U, (minimizes expected losses), subject to $\mathbf{c}\cdot\mathbf{t}\leq Y$, then $\mathbf{t}^{*}=\bar{\mathbf{t}}(1+O(\varepsilon))$ where $\bar{\mathbf{t}}$ maximizes the worst-case total precision, $\min_{D}\sum n_{j}\beta_{j\mathbf{r}}(D)$, 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

$$\underline{A}\frac{\max_{D}\left\{\rho_{\bar{\mathbf{r}}}(D)^{Y/(\bar{\mathbf{r}}\cdot c)}\right\}}{\sqrt{Y/(\bar{\mathbf{r}}\cdot c)}}\leq\underline{A}\frac{\max_{D}\left\{\rho_{\mathbf{r}_{Y}^{*}}(D)^{Y/(\mathbf{r}_{Y}^{*}\cdot c)}\right\}}{\sqrt{Y/(\mathbf{r}_{Y}^{*}\cdot c)}}\leq\overline{A}\frac{\max_{D}\left\{\rho_{\bar{\mathbf{r}}}(D)^{Y/(\bar{\mathbf{r}}\cdot c)}\right\}}{\sqrt{Y/(\bar{\mathbf{r}}\cdot c)}}$$

where $\overline{A}<\infty$ and $\underline{A}>0$ are upper and lower bounds, uniform in across all values of $\mathbf{r}$, 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 1.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 1.A.2.

In short, Proposition 1.1 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, quasi-concave utility model:

Lemma 1.1 (Properties of precision). *For a fixed dichotomy, precision is homothetic and quasi*convex *in quantity, $\mathbf{t}$. The worst-case total precision is thus homothetic and* locally *quasiconvex*[11] *(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 $\rho_{j\mathbf{r}}(D)$ 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

$$\sum_{j=1}^{J} r_j \beta_{j\mathbf{r}}(D) \leq \sum_{j=1}^{J} r_j \beta_j(D) \tag{1.5}$$

(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 1.1 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*[12] *income elasticity of demand for all information sources given a fixed change in income is* $1 + O(\varepsilon)$ *as costs* (ε) *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 (1.5) simply implies that loss-minimizing bun-

11 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.

12 The arc formula avoids technical issues around differentiating an approximation and applies equally in the discrete sampling environment.

dles 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).

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 isoprecision 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 1.1).

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 isoprecision lines intersect. Proposition 1.2 generalizes this observation to generic, finite collections of information sources.

Proposition 1.2 (Iso-least-precision kinks). *For generic information sources—i.e. ones for which (1.5) 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 1.A.3.

Thus, 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. Put another way, Hicksian demand for information should be approximately a step function.

1.4.2 Complexity of optimal sample bundles

Because isoprecision 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 1.3. *For generic information sources, the proportions,* $\mathbf{r}^*$ *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 1.2, 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 isoprecision surfaces. Thus, $\mathbf{r}^*$ can have support on at most as many sources as there are state pairs. See Appendix 1.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.

1.4.3 Trade-offs 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 (1.4) 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 1.4 (Marginal rate of substitution). *Let $\mathscr{E}_1$ and $\mathscr{E}_2$ be compound Poisson experiments.*[13] *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. Let $\mathscr{E}_1$ be a Poissonization of $\hat{\mathscr{E}}_1$. Then we can write the expected loss from quantity t_1 from $\mathscr{E}_1$ (t_1 *expected* samples from $\hat{\mathscr{E}}_1$) by averaging the loss from each possible draw of samples from $\hat{\mathscr{E}}_1$:

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

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

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

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

13 Compound Poisson experiments are assumed for technical simplicity. The general result can be proved using a variation of the method in Appendix 1.A.1, but is highly technical for $|\Theta| > 2$.

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.

we can apply Proposition 1.0 to write

$$\frac{\partial L(t_1, t_2, \dots, t_J)}{\partial t_1} = (\hat{\rho}_{1\mathbf{r}}(D_{\mathbf{r}}) - 1)A(\mathbf{r})\frac{\{\rho_{1\mathbf{r}}^{t_1}\prod_{j=2}^{J}\rho_{j\mathbf{r}}(D_{\mathbf{r}})^{t_j}\}}{\sqrt{T}}\left(1 + O\left(\frac{1}{T}\right)\right)$$

where $\hat{\rho}_{1\mathbf{r}}$ is the marginal efficiency index of $\hat{\mathscr{E}}_1$ and $D_{\mathbf{r}}$ is the (unique at $\mathbf{r}$) worst-case dichotomy. As discussed in section 1.4.1, the precision of a compound Poisson experiment is one less the efficiency index of the underlying experiment, so

$$\frac{\partial L(t_1, t_2, \dots, t_J)}{\partial t_1} = -\beta_{1\mathbf{r}}(D)L(t_1, t_2, \dots, t_J)\left(1 + O\left(\frac{1}{T}\right)\right) \tag{1.6}$$

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 1.B for a more thorough discussion of the substitutability of samples in the general non-infinitely divisible setting.

1.5 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

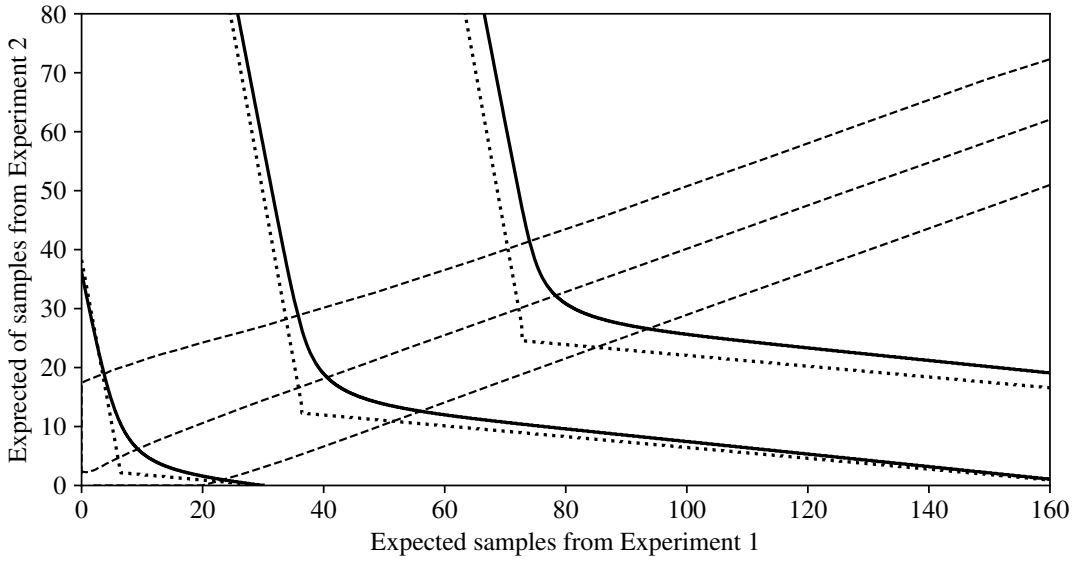

Figure 1.2: Illustration of income effects: income-expansion paths of demand in a three-state environment two available sources of information. Dotted lines are the iso-least-precision lines, solid lines are numerically computed true indifference curves, and dashed lines are numerically computed true income expansion paths for three different cost ratios.

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.[15]

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.

15 In these simulations, infinite divisibility is achieved by Poissonization.

The income expansion paths plotted in Figure 1.2 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.

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 1.3 a Hicksian (compensated)[16] demand for information for the two information sources used in Figure 1.2. 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.

---

16 Hicksian demand holds expected losses constant rather than budget and thus illustrates demand absent any income effects.

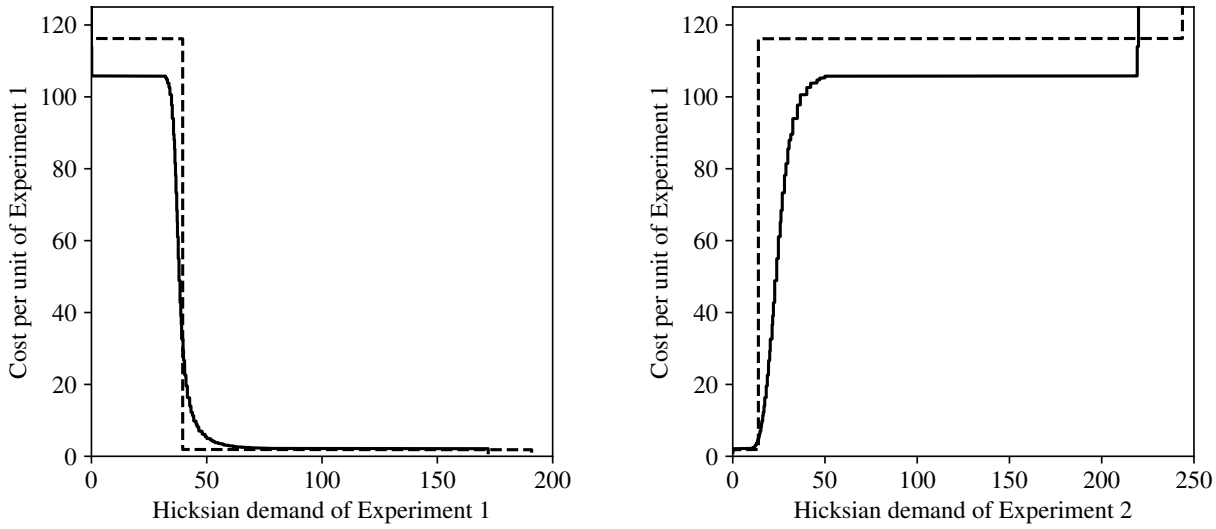

Figure 1.3: Illustration of substitution effects for information: Hicksian demand for the same information sources used in Figure 1.2 as a function of the cost of Experiment 1, holding the cost of Experiment 2 fixed. The solid line is a true Hicksian demand curve, and the dashed line is the equivalent curve predicted by the maximin precision rule.

1.6 CONCLUSION

This chapter 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 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.

Note, however, that many real-world problems don't fit neatly into the finite-state/finite-action paradigm. For example, researchers most commonly wish to *estimate* the value of some real-valued parameter.

In such settings, we cannot easily directly apply these results because a "worst-case" efficiency index will not exist. Specifically, for a dichotomy composed of two states arbitrarily close to one another, the efficiency index will be arbitrarily close to 1. In such settings, the large sample expected loss will be dominated by "small deviations" of the estimator around the truth, whereas in the finite-state case considered here, losses are zero except after large deviations of the realizations.

The following chapter will explore the consumer theory for information in such a setting with real-valued states.

1.A OMITTED PROOFS

1.A.1 Proof of Proposition 1.0

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, 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 that the posterior log-likelihood ratio is above the rejection threshold when the true state is θ_0:

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

Moscarini 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*

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

$$\mathbb{P}(\bar{Y}_t \geq \xi) = \frac{e^{t(K(\zeta^*(\xi)) - \zeta^*(\xi)\xi)}}{\zeta^*(\xi)\sqrt{2\pi t K''(\zeta^*(\xi))}} \left(1 + O\left(\frac{1}{t}\right)\right) \tag{1.7}$$

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

I provide a sketch of the proof of this lemma in Appendix 1.A.5. Note that although Lugannani and Rice work with discretely sampled distributions, they use the method of steepest descents to approximate the inversion of $M(i\zeta)^t$ given by the previous lemma. Thus, because the method of steepest descents does not require integer powers,[17] (1.7) 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

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

17 $M(\zeta)^t$ is guaranteed to be a valid moment-generating function only for positive integer t for non-infinitely divisible distributions.

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 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 (1.7) and apply Taylor's theorem again. Breaking it down into parts we have

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

Plugging each of the above parts into Equation (1.7) we have that the error probability is

$$\alpha_I(t) = \frac{e^{\zeta^* \bar{l}}}{\zeta^* \sqrt{2\pi K''(\zeta^*)}} \frac{\rho^t}{\sqrt{t}} \left(1 + O\left(\frac{1}{t}\right)\right) \tag{1.8}$$

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. □

1.A.2 Omitted Parts of the Proof of Proposition 1.1

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 1.0, we have $A(\mathbf{r},T) = O(1)$ 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 1.0. To get an upper bound, use Claim 3 of (Moscarini and Smith, 2002, p. 2363):

Lemma (Moscarini and 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)*

$$\mathbb{P}\left(\sum_{\theta\neq\theta_0} b_\theta \frac{\mu_\theta^t}{\mu_{\theta_0}^t}(\mathbf{x}) > \varepsilon \;\middle|\; \theta_0\right) = O\left(\frac{\max_{\theta\neq\theta_0}\rho(\theta,\theta_0)^t}{\sqrt{t}}\right) \tag{1.9}$$

Because each mistake probability takes the form of (1.9), 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

$$\mathbf{r}_Y^* \in \left\{ \bar{\mathbf{r}} + O(Y^{-1}) \,:\, \bar{\mathbf{r}} \in \arg\min_{\mathbf{r}} \left\{ \min_D \beta_{\mathbf{r}}(D)/(\mathbf{r}\cdot\mathbf{c}) \right\} \right\} \tag{1.10}$$

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.[18] □

1.A.3 Proof of Proposition 1.2

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 of $\mathbf{t}^*$. We then have

18 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.

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. □

1.A.4 Proof of Proposition 1.3

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 isoprecision 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 isoprecision 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 isoprecision 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 $\mathbf{t}^*$. □

1.A.5 Proof sketch of the Lugannani and Rice (1980) Saddlepoint Approximation

Lugannani and Rice (1980) approximate the CDF by applying a saddlepoint approximation to

a particular characteristic function inversion formula. The full formal proof relies a fair bit on some advanced complex analysis, and so I provide a sketch of the approach that attempts to minimize reliance on complex analysis beyond the standard Cauchy-Riemann conditions for analytic functions.

Recall that the characteristic function of a random variable, X is $\phi(t) = \mathbb{E}(e^{itX})$. Equivalently we can think of the characteristic function as a rotation of the MGF in the complex plane, $\phi(t) = M(it)$, so it shares with the MGF the property that the characteristic function of an independent sum of random variables is the product of each characteristic function.

The following lemma establishes a useful property about the characteristic function in the context of LLR distributions:

Lemma (Differentiability of the characteristic function). *Suppose ϕ is the characteristic function for a LLR distribution whose MGF is finite in an open interval containing the origin. Then if $\mathfrak{I}(t) \in (-1, 0)$, then ϕ is analytic at t, and thus is infinitely differentiable there.*[19]

Proof. Assuming ϕ is differentiable, we can use Leibniz rule to write

$$\phi^{(k)}(t) = \int \left(\log \frac{\mathrm{d}F(r \mid \theta')}{\mathrm{d}F(r \mid \theta)}\right)^k \mathrm{d}F(r \mid \theta')^{it} \mathrm{d}F(r \mid \theta)^{1-it}$$

Thus, so long as this integral is finite, we have differentiability. For t such that $\mathfrak{I}(t) \in (-1, 0)$ this follows immediately from dominated convergence and the fact that the LLR distribution has all its moments (because the MGF is defined on an open interval containing the origin). ▽

In particular, the previous lemma also implies that in a sufficiently small interval around any t in the analytic strip, ϕ is well approximated by its Taylor series—a tool that will prove useful later.

Now, because characteristic functions uniquely define their distribution, we can invert it to get the distribution function. Typically, one would use the standard Fourier inversion formula

19 $\mathfrak{R}(t)$ and $\mathfrak{I}(t)$ respectively denote the real and imaginary parts of t

to get the density of the distribution; however, we don't generally have a density, so we instead must use a less commonly known inversion formula:

Lemma (Characteristic function inversion). *Let F be the CDF for some distribution Y on $\mathbb{R}$, with characteristic function, ϕ. If ξ is a continuity point of F then*

$$1 - F(\xi) = \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L-ci}^{L-ci} \frac{e^{-it\xi}\phi(t)}{it} \mathrm{d}t \tag{1.11}$$

for arbitrary real constant $c > 0$ such that the MGF of F is differentiable at c.

Now, recall that the distribution in question is an i.i.d. sum of n samples, and we want to know when the sample average exceeds ξ (so when the sum exceeds $n\xi$). Thus, we can write the characteristic function as $\phi(t) = M(it)^n = e^{nK(it)}$. Using this and changing variables to $T = it$, rewrite the inversion formula as

$$1 - F(\xi) = \frac{1}{2\pi} \lim_{L\to\infty} \int_{-iL+c}^{iL+c} \frac{e^{n(K(T)-T\xi)}}{iT} \mathrm{d}T$$

Now, let $c = \tau(\xi)$ where $\tau(\xi)$ is the (real) argmin of $K(t) - t\xi$, so the path of integration crosses the real line perpendicular to the minimum (along the real line) of $K(t) - t\xi$. Recall that K is analytic, and thus by the Cauchy-Riemann equations, $\tau(\xi)$ is a saddlepoint. So, although $\tau(\xi)$ minimizes $K(t) - t\xi$ when traveling along the real axis, it *maximizes* it when traveling along the perpendicular complex axis.

Notice, the integrand in the above equation is complex, but the right-hand side is real, so it will be useful to separate the integrand into its complex part (which must integrate to zero since probabilities are real-valued) and its real part. First do another change of variables:

$$1 - F(\xi) = \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} \frac{e^{n(K(\tau(\xi)+ix)-(\tau(\xi)+ix)\xi)}}{\tau(\xi) + ix} \mathrm{d}x$$

Then split into real and complex parts:

$$\begin{aligned}
1 - F(\xi) &= \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} \frac{e^{n\Re(g(\tau(\xi)+ix))}(\cos(n\Im(g(\tau(\xi)+ix))) + i\sin(n\Im(g(\tau(\xi)+ix))))}{\tau(\xi)+ix} \mathrm{d}x \\
&= \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} \frac{e^{n\Re(g(\tau(\xi)+ix))}(\tau(\xi)\cos(n\Im(g(\tau(\xi)+ix))) + x\sin(n\Im(g(\tau(\xi)+ix))))}{\tau(\xi)^2 - x^2} \mathrm{d}x \\
&\equiv \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} e^{n\hat{g}(x)} h(x) \mathrm{d}x
\end{aligned}$$

where $g(T) \equiv K(T) - T\xi$. The first line follows from Euler's formula, the second follows from multiplying numerator and denominator by $\tau(\xi) - ix$ and then canceling any imaginary terms since the complex part of the integrand must integrate to zero, and the third follows from appropriately defining $\hat{g}$ and h to simplify notation.

The integrand of the last line is now mapping reals into reals and can thus be handled with standard real analysis tools. In particular, we can approximate the integral using a Laplace approximation. Roughly, the idea is that $\hat{g}(x)$ takes its max at $x = 0$, and on account of the exponentiation, the integrand will get almost all its mass from a vanishing interval around that max as n gets large.

Now write

$$\int_{-\infty}^{\infty} e^{n\hat{g}(x)} h(x) \mathrm{d}x = e^{n\hat{g}(0)} \int_{-\infty}^{\infty} e^{n(\hat{g}(x)-\hat{g}(0))} h(x) \mathrm{d}x$$

Using a second-order Taylor approximation, we can then write

$$e^{n\hat{g}(0)} \int_{-\infty}^{\infty} e^{n\hat{g}(x)-\hat{g}(0)} h(x) \mathrm{d}x \approx e^{n\hat{g}(0)} \int_{-\infty}^{\infty} e^{\frac{1}{2}n\hat{g}''(0)x^2} h(x) \mathrm{d}x$$

The LHS of the above now is a familiar form: a Gaussian integral. Thus, we can use the Taylor approximation for the expectation of $h(x)$ when x is Gaussian with small variance

$$\begin{aligned}
e^{n\hat{g}(0)} \int_{-\infty}^{\infty} e^{\frac{1}{2}n\hat{g}''(0)x^2} h(x) \mathrm{d}x &\approx e^{n\hat{g}(0)} \sqrt{\frac{2\pi}{-n\hat{g}''(0)}} (h(0) - (n\hat{g}''(0))^{-1} h''(0)) \\
&= e^{n\hat{g}(0)} \sqrt{\frac{2\pi}{-n\hat{g}''(0)}} h(0)(1 + O(n^{-1}))
\end{aligned}$$

Now, plugging things in and using the Cauchy-Riemann equations, we have

$$\hat{g}(0) = K(\tau(\xi)) - \tau(\xi)\xi$$
$$-\hat{g}''(0) = K''(\tau(\xi))$$
$$h(0) = \frac{1}{\tau(\xi)}$$

Plugging all this into our original equations, we have

$$1 - F(\xi) = \frac{e^{n(K(\tau(\xi))-\tau(\xi)\xi)}}{\tau(\xi)\sqrt{2\pi n K''(\tau(\xi))}}\left(1 + O\left(\frac{1}{n}\right)\right)$$

as claimed. □

1.B DISCRETE SAMPLING

Typically, in these settings, we would measure quantity of information by the number 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 isoprecision 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 1.4 (which fundamentally depends on differentiability) thus apply. Propositions 1.2 and 1.3 are stated in terms of maximin precision bundle and thus need no modification to their proofs. The proof of Proposition 1.1 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 $O(Y^{-1})$ of proportions that would make the budget constraint bind, so this merely adds another $O(Y^{-1})$ term.

To generalize Proposition 1.4, requires a bit more work, as we first need to define what we even mean by *marginal rate of substitution* in a discrete setting.

1.B.1 A discrete version of Proposition 1.4

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 1.4A (Sample substitutability). *Consider a sample bundle with sample proportions* $\mathbf{r}$ *and unique least-precision dichotomy* $D_{\mathbf{r}}$. *Then, for* $N = \sum_j n_j$ *high enough, 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 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

$$\begin{aligned} n_1\beta_{1\mathbf{r}} + n_2\beta_{2\mathbf{r}} + \sum_{j=3}^{J} n_j\beta_{j\mathbf{r}} + \log(A(\mathbf{r})) + O(N^{-1}) \\ \leq (n_1 - k_1)\beta_{1\mathbf{r}'} + (n_2 + k_2)\beta_{2\mathbf{r}'} + \sum_{j=3}^{J} n_j\beta_{j\mathbf{r}'} + \log(A(\mathbf{r}')) \end{aligned} \tag{1.12}$$

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

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

Further, because $A(\mathbf{r})$ is a differentiable function of $\tau_{\mathbf{r}}$ (see (1.8) in the last part of the proof of Proposition 1.0), we have that $\log(A(\mathbf{r}')) - \log(A(\mathbf{r})) = O(N^{-1})$. Plugging all of this into (1.12) and rearranging gives

$$k_2 \geq k_1\frac{\beta_{1\mathbf{r}}}{\beta_{2\mathbf{r}}} + O(N^{-1}) \tag{1.13}$$

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

$$k_2 \leq k_1 \frac{\beta_{1\mathbf{r}}}{\beta_{2\mathbf{r}}} + 1 + O(N^{-1}) \tag{1.14}$$

Together, by squeezing k_2 between (1.13) and (1.14) 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. □

CHAPTER TWO

CHEAP INFORMATION CONSUMER THEORY FOR ESTIMATION PROBLEMS

ABSTRACT

This chapter generalizes the approximate consumer theory for discrete decision problems developed in Chapter 1 to decision problems with real-valued unknown parameter and action (e.g. estimation problems). In this setting, I show expected loss is well-approximated by a weighted average of the inverse Fisher information of the sample bundle. From this, I derive a computationally-tractable expression for the marginal rate of substitution between two sources of information. In contrast to the discrete case, approximate indifference curves are smooth and exhibit monotonically increasing rates of substitution. Large sample demand in this setting thus behaves according to the benchmark consumer theory model. I apply this result to estimating the prevalence of a disease in a large population using tests of differing sensitivity and specificity.

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

2.1 INTRODUCTION

Chapter 1 developed an approximate consumer theory in finite-action/finite-state decision problems. In practice, however, uncertainty takes the form of a real-valued parameter. For example, a firm might want to estimate the slope of a demand curve before deciding whether to introduce a new product, or a policymaker might want to estimate the prevalence of some transmissible disease prior to deciding whether to implement some sort of mitigating policy.

In these settings, the approach in Chapter 1 does not easily apply: typical information sources struggle to distinguish between states close to one another. Formally, the efficiency index for a dichotomy with arbitrarily close state pairs is arbitrarily close to 1, so no "worst-case" dichotomy exists.

In this chapter, I show that for a general class of loss differentiable loss functions, expected loss for a Bayesian decision maker is well-approximated by a weighted average of the variance of the maximum likelihood estimator—i.e., the reciprocal of the Fisher information (Proposition 2.1).

The optimal information bundle is thus approximately one which minimizes the weighted Fisher information per dollar (Proposition 2.2). From this follows that demand behaves as though preferences are smooth and convex—in contrast to the somewhat pathological preferences from Chapter 1.

I then derive computationally tractable expression for the marginal rate of substitution between two sources of information (Proposition 2.4). As in Chapter 1, approximate marginal rate of substitution is scale invariant, and thus the implied optimal bundle has unit income elasticities.

Finally, I apply these results to a setting of designing a study to test for disease prevalence with multiple available tests. Given sensitivity and specificity of two tests, I derive the range of price ratios where interior solutions are optimal, and show that a test with relatively higher sensitivity is optimally used more when disease prevalence is expected to be high.

This work is most closely related to the extensive literature on experimental design (see Chaloner and Verdinelli, 1995, for an overview of experiment design in Bayesian settings). In

particular, this result generalizes the well-known *A-optimality* criterion[1] for constructing optimal experiments. Clyde (1993) previously derived a decision-theoretic foundation for this criterion with squared-deviation loss. My results generalize this to general (suitably-smooth) losses and consider the consumer-theoretic implications of this criterion.

The remainder of this chapter is structured as follows: Section 2.2 lays out the model and required regularity conditions, Section 2.3 describes the main results, Section 2.4 applies the results to the aforementioned disease prevalence estimation problem, and Section 2.5 concludes.

2.2 MODEL

2.2.1 States, Beliefs, and Information

A decision maker (DM) has a (full support) prior with continuously differentiable density μ over states of the world $\theta \in \Theta$ assumed to be a convex subset of the reals.[2]

The DM may purchase information from sources $\mathscr{E}_1, \dots, \mathscr{E}_J$ prior to acting, where an information source is a collection of realizations, X_j and a state-dependent distribution over those realizations, $F_j(x|\theta)$. For simplicity, let F_j either be a discrete or continuous distribution with likelihood function f_j.[3]

I will additionally assume that all experiments satisfy a number of regularity conditions, that require the experiments to be informative, but in a sense "thin-tailed," so not *too* informative. For simplicity, I defer the formal statement of these conditions to Appendix 2.A.1. Suffice it to say, experiments based on common distributions such as those in the exponential family satisfy these regularity conditions.

The DM may purchase a vector of conditionally independent samples $\mathbf{n} \equiv (n_1, \dots, n_J) \in \mathbb{N}^J$ at costs $\mathbf{c} \equiv (c_1, \dots, c_J) > 0$ per sample, up to her budget Y.

---

1 See Chapter 9, Pukelsheim (1993) for a textbook treatment of the alphabet of design criteria

2 For exposition, I focus on the one-dimensional (single-parameter) setting. The main results easily generalize to multidimensional settings. See Appendix 2.B.

3 To limit notation, I denote expectations over realizations as though f_j is a density.

After observing the realizations, $\mathbf{x_n}$, from the chosen bundle, the DM, updates her prior using Bayes rule to posterior $\mu(\theta|\mathbf{x_n})$.

2.2.2 *Actions and loss functions*

The DM has a *loss* function $l(a,\theta) \geq 0$, where for some $a^*(\theta)$, $l(a^*(\theta),\theta) = 0$ and $l(a,\theta) > 0$ otherwise.[4] After observing realizations from her chosen information bundle, the DM chooses an action $a \in A$ in order to minimize her expected loss $\int_{\theta\in\Theta} l(a,\theta)\mu(\theta|\mathbf{x_n})d\theta$.

Assume A is a convex subset of the reals, and without loss let $A = \Theta$, so the DM's problem may be viewed as one of *estimation*.

Let $l(\theta,\theta) = 0$ and $l(a,\theta) > 0$ for any $a \neq \theta$. As with the experiments, I will additionally assume a number of regularity conditions on the loss function. These conditions ensure the loss function is smooth and that the expected loss exists, but for the sake of brevity, the formal statements are deferred to Appendix 2.A.1. We will additionally require the loss function satisfy the following regularity conditions:

Note that the above are satisfied by common loss functions such as squared deviation.[5]

2.2.3 *Information demand*

As in Chapter 1, the DM chooses her bundle of information, observes the realizations, and then takes a once-and-for-all action. Her ex ante expected loss from a bundle $\mathbf{n}$ of information is

$$L(\mathbf{n}) = \int_{\theta\in\Theta}\int_{\mathbf{x_n}} l(a(\mathbf{x_n}),\theta) f_{\mathbf{n}}(\mathbf{x_n}|\theta) d\mathbf{x_n}\mu(\theta)d\theta$$

where $f_{\mathbf{n}}$ is the likelihood of the $\mathbf{n}$ bundle realization and $a(\mathbf{x_n})$ is the expected-utility-maximizing action after observing $\mathbf{x_n}$.

The DM's chooses her information bundle $\mathbf{n} \geq 0$ to minimize $L(\mathbf{n})$ subject to her budget

4 See Chapter 1 for how to convert an arbitrary state-dependent utility function to a loss function.

5 A notable exception are absolute deviation and check loss functions. The main approximation depends fundamentally on a Taylor approximation around the true state, and thus does not apply to loss functions non-differentiable around the truth.

constraint $\mathbf{n} \cdot \mathbf{c} \leq Y$.

2.3 RESULTS

2.3.1 Approximating loss for a single experiment

As in Chapter 1, we start by analyzing the expected loss from a single experiment $\mathscr{E}$. In particular, I will approximate the state-θ *risk* of n samples

$$R_n(\theta) \equiv \int L(a(\mathbf{x}_n), \theta) f_n(\mathbf{x}_n|\theta) d\mathbf{x}_n$$

where $a(\mathbf{x}_n)$ is the Bayes optimal action given realization $\mathbf{x}_n$. Notice that once we have an approximation for R_n, we can easily get one for $L(n)$ because the expected loss is simply the expectation of the risk:

$$L(n) = \int R_n(\theta)\mu(\theta)d\theta$$

At large samples, we should expect R_n to approach zero as Bayes optimal action should have very small variance with average approaching θ (Bayes estimators are consistent) To approximate R_n then we would like to understand the asymptotic distribution of $a(\mathbf{x}_n)$, then apply a second order Taylor approximation

$$\mathbb{E}(l(a, \theta)) \approx \frac{1}{2} l_{aa}(\theta, \theta)\mathbb{E}((a - \theta)^2)$$

Perhaps unsurprisingly, under the assumed regularity conditions, a central limit theorem applies to the optimal Bayes action:

Lemma 2.1 (Asymptotic distribution of the Bayes action). *If $a(\mathbf{x}_n)$ is the Bayes optimal action, and θ is the true state, then as $n \to \infty$*

$$\sqrt{n}(a(\mathbf{x}_n) - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$$

where I is the Fisher information of $\mathscr{E}$:

$$I(\theta) \equiv \int \left(\frac{\partial}{\partial\theta} \log f(x|\theta)\right)^2 f(x|\theta)dx$$

Proof heuristic. At large samples, the likelihood will be sharply peaked at the maximum likelihood estimate. At large samples, the sample likelihoods overwhelm the prior, so we should expect the posterior to be sharply peaked at the maximum likelihood estimate as well.[6] Thus, at large samples, the Bayes optimal action should approach to the maximum likelihood estimate, and thus should have the asymptotic distribution. See Appendix 2.A.2 for the formal proof.

Using this asymptotic distribution, we can approximate the risk as follows:

Lemma 2.2. *The state-θ risk is*

$$R_n(\theta) = l_{aa}(\theta,\theta)\frac{1}{nI(\theta)}\left(1 + o(1)\right) \tag{2.1}$$

as $n \to \infty$.

Proof heuristic. By Taylor's theorem, we have $L(\hat{\theta}_n, \theta) \approx \frac{1}{2} l_{aa}(\theta,\theta)(\hat{\theta}_n - \theta)^2$ Because $\hat{\theta}$ is asymptotically normal with variance $(nI(\theta))^{-1}$ by Lemma 2.1, we must then have $R_n(\theta) \approx l_{aa}(\theta,\theta)\frac{1}{nI(\theta)}$. See Appendix 2.A.3 for formal proof.

By taking expectations of Equation (2.1), we have that the expected loss is simply the expected variance suitably weighted by the loss function curvature:

Proposition 2.1 (Expected loss approximation). *The expected loss[7] from n samples is*

$$L(n) = \frac{1}{2n}\int \frac{1}{I(\theta)} l_{aa}(\theta,\theta)\mu(\theta)d\theta\left(1 + o(1)\right)$$

6 The Bernstein-von Mises Theorem (See Ch. 10 van der Vaart, 1998) roughly states that the posterior density at large samples is approximately $\mathcal{N}(\hat{\theta}_{MLE}, (nI(\hat{\theta}_{MLE})^{-1}))$ where $\hat{\theta}_{MLE}$ is the maximum likelihood estimate.

as $n \to \infty$.

Proposition 2.1 suggests that the optimal bundle should minimize a weighted average expected asymptotic variance. Under squared deviation (or any translationally symmetric loss function $l(a, \theta) = l(|a - \theta|)$), this "expected variance" criterion is known as the *A-criterion* (see p. 137 of Pukelsheim, 1993).

With this result, we can now generalize this criterion to account both for general loss functions and sample costs, and then formally show that demand is approximately given by this generalized A-criterion.

2.3.2 Optimal demand for information

Fisher information is additive in samples, so Proposition 2.1 easily extends to a multi-experiment bundle as we similarly did in Chapter 1.

Letting $N \equiv \sum_j n_j$ and $\mathbf{r} \equiv \mathbf{n}/N$ we define the *composite* Fisher information: $I_{\mathbf{r}}(\theta) \equiv \sum_j r_j I_j(\theta)$. The total Fisher information of a bundle is thus $NI_{\mathbf{r}}(\theta)$. We can then write the expected loss as

$$L(\mathbf{n}) = \frac{1}{2N} \int \frac{1}{I_{\mathbf{r}}(\theta)} l_{aa}(\theta, \theta) \mu(\theta) d\theta (1 + o(1))$$

as $N \to \infty$.

Notice that $(2N)^{-1} \int I_{\mathbf{r}}(\theta)^{-1} l_{aa}(\theta, \theta) \mu(\theta) d\theta$ is homothetic, so just as with finite-state model, the relative proportions of the optimal bundle should be invariant to budget. In particular, defining *utility*,

$$U(\mathbf{n}) \equiv L(\mathbf{n})^{-1} = N \left(\int \frac{1}{I_{\mathbf{r}}(\theta)} l_{aa}(\theta, \theta) \mu(\theta) d\theta \right)^{-1} (1 + o(1))$$

we can find optimal demand by finding the relative sample proportions $\mathbf{r}$ that maximize *utility per dollar*.

---

7 I conjecture that a saddlepoint-like approximation could tighten the convergence rate to $O(n^{-1})$ as in Chapter 1. Such an approximation is considerably more technical in this setting than it was in Chapter 1.

Proposition 2.2 (Optimal demand: generalized A-optimality). *Let costs be* $\mathbf{c} = (\varepsilon\kappa_1, \varepsilon\kappa_2, \ldots, \varepsilon\kappa_J)$*, and assume* $\bar{\mathbf{r}}$ *maximizing payoffs per dollar,*

$$\bar{\boldsymbol{r}} = \arg\max_{\mathbf{r}} \left\{ \left(\int \frac{1}{\sum_{j=1}^{J} r_j I_j(\theta)} l_{aa}(\theta,\theta)\mu(\theta)d\theta \right)^{-1} \Big/ \sum_{j=1}^{J} r_j\kappa_j \right\}$$

is unique. If $\mathbf{n}^*$ *minimizes expected losses, L, subject to* $\mathbf{c}\cdot\mathbf{n} \leq Y$*, then the optimal proportions*[8] $\mathbf{r}^* = \mathbf{n}^*/N$ *are*

$$\mathbf{r}^* = \bar{\mathbf{r}} + o(1)$$

as ε *goes to zero.*

Because the optimal relative proportions are constant in income at low prices (high incomes), we immediately have that samples from all experiments are normal goods:

Corollary (Unit income elasticities). *The arc income elasticity of demand for all information sources given a fixed change in income is* $1 + O(\varepsilon)$ *as costs (*ε*) go to zero.*

Notice that in contrast to the maximin optimization of the finite-state decision problem, defines a $\int (I_{\mathbf{r}}(\theta))^{-1} l_{aa}(\theta,\theta)\mu(\theta)d\theta$ much better behaved preference ordering. First, the approximation is clearly smooth $\mathbf{r}$, but more strongly, the approximate preferences are *convex*.

Proposition 2.3 (Expected loss is approximately convex at large samples). *Fix* $\mathbf{r}$*,* $\mathbf{r}'$ *and* $\lambda \in (0,1)$ *(all rational). For N high enough*

$$L(N(\lambda\mathbf{r} + (1-\lambda)\mathbf{r}') \leq \lambda L(N\mathbf{r}) + (1-\lambda)L(N\mathbf{r}')$$

Proof. $\left(\sum_j r_j I_j(\theta)\right)^{-1}$ is convex in $\mathbf{r}$, and the sum of convex functions is convex. The weighted average Fisher information is thus convex. Application of Proposition 2.1 completes the proof.

8 Here we can only say that optimal proportions converge. Because the actual optimal bundle is $N\mathbf{r}$, the distance between the "approximate" optimal bundle and the true optimal may diverge. If the approximation of Lemma 2.2 converges at a conjectured $O(n^{-1})$ rate as did the equivalent approximation in Chapter 1, the difference between the true and approximate bundles would be within a constant, as in Chapter 1.

□

In such settings, we thus have that information behaves essentially according to a benchmark consumer theory model with increasing rates of substitution, though corner solutions are still salient as they were in Chapter 1.

Note that the inequality in Proposition 2.3 is typically strict. For example, with two experiments, the weighted average inverse Fisher information is linear in $\mathbf{r}$ if and only if the two experiments have proportional Fisher information at (almost) *all* values of θ. Two homoskedastic Gaussian experiments will have this property, but a typical pair will not. In fact, in Section 2.4, I will show that *any* two distinct tests for a disease always have *strictly* convex weighted average inverse Fisher information.

2.3.3 Trade-offs between sources

Because the approximate preferences are smooth in this setting, understanding the substitutability of samples is useful even in settings with linear budget constraints.

We can then compute an exact-at-large-samples value for the rate at which samples from two sources are substitutable similar to Proposition 1.4A:

Proposition 2.4 (Sample substitutability). *Let k_2 be the minimum number of samples from $\mathscr{E}_2$ to compensate for a loss of k_1 samples from $\mathscr{E}_1$. Then for $n_1 + n_2$ large enough, k_2 is exactly*

$$k_2 = \left\lceil k_1 \left(\int \frac{I_1(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \Big/ \int \frac{I_2(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \right) \right\rceil$$

Proof heuristic. By application of Taylor's theorem, the increase in expected losses from giving up a sample of $\mathscr{E}_1$ is roughly

$$\int \frac{I_1(\theta)}{I_{\mathbf{r}}(\theta)} l_{aa}(\theta, \theta) \mu(\theta) d\theta$$

and similarly the decrease in expected losses from an additional sample of $\mathscr{E}_2$ is the same replacing I_1 by I_2. Thus, for the DM to be indifferent about trading off k_1 samples from $\mathscr{E}_1$ for k_2 from $\mathscr{E}_2$,

the ratio k_2/k_1 should equal the ratio of marginal losses. See Appendix 2.A.5 for the formal proof.

From this, and the fact that preferences are generically approximately convex at large samples, we can derive a condition for corner solutions:

Corollary (Corner solutions). *Suppose there are two available experiments, $\mathscr{E}_1$ and $\mathscr{E}_2$ with costs $c_1 = \varepsilon\kappa_1$ and $c_2 = \varepsilon\kappa_2$. Then as the costs go to zero, $\varepsilon \to 0$, the optimal bundle has proportion of samples from $\mathscr{E}_1$ approaching 1 whenever*

$$\frac{\kappa_2}{\kappa_1} > \int \frac{1}{I_1(\theta)} l_{aa}(\theta,\theta)\mu(\theta)d\theta \Bigg/ \int \frac{I_2(\theta)}{I_1(\theta)^2} l_{aa}(\theta,\theta)\mu(\theta)d\theta$$

Note that under the assumed regularity conditions, the left-hand side of the above inequality is always finite, and thus corners solutions always occur for sufficiently extreme price ratios, just as they did with the finite-state case.

2.4 APPLICATION

As an elementary application of the above results, consider the problem of a health policymaker who wishes the estimate the *prevalence*, $\theta \in [0,1]$ of some disease in a population. In this case, the policymaker must decide how many individuals from the population to sample, and which test for the disease to give to a given individual. In this setting, a test has two possible realizations—positive and negative—and is characterized by two values: its *sensitivity* (true positive rate) and *specificity* (true negative rate), hereafter denoted γ and δ respectively. The state dependent probability of observing a positive test result is then

$$p^+(\theta) = \theta\gamma + (1-\theta)(1-\delta)$$

with negative test result with complementary probability.

Note that we can without loss of generality assume $\gamma \geq 1-\delta$ because if γ (the true positive rate) were lower than $1-\delta$ (the false positive rate), it would be appropriate to swap which outcome we call "positive" because otherwise a positive result would be more likely if the individual is actually healthy. A test with $\gamma = 1-\delta$ is uninformative, returning positive with the same rate independent of actual health status.

For the sake of this example, assume there are two available distinct tests $\mathscr{E}_1$ and $\mathscr{E}_2$ with sensitivities and specificities γ_1, δ_1 and γ_2, δ_2 respectively.

In this case we can compute the Fisher information:

$$\begin{aligned} I_j(\theta) &= \left(\frac{d}{d\theta}\log p_j^+(\theta)\right)^2 p_j^+(\theta) + \left(\frac{d}{d\theta}\log(1-p_j^+(\theta))\right)^2 (1-p_j^+(\theta)) \\ &= \frac{(\gamma_j+\delta_j-1)^2}{\delta_j(1-\delta_j)+\theta(\gamma_j+\delta_j-1)(2\delta_j-1)-\theta^2(\gamma_j+\delta_j-1)^2} \end{aligned}$$

Observation 2.1. *If $\mathscr{E}_1$ and $\mathscr{E}_2$ have distinct sensitivity and specificity, the weighted average inverse Fisher information is strictly convex.*

To see this, note that the inverse fisher information is

$$I_j(\theta)^{-1} = \frac{\delta_j(1-\delta_j)}{(\gamma_j+\delta_j-1)^2} + \theta\frac{2\delta_j-1}{\gamma_j+\delta_j-1} - \theta^2$$

In order for the weighted average inverse Fisher information to be linear in samples, we must have some $B > 0$ such that $I_1(\theta) = BI_2(\theta)$ for all θ. Suppose then that $I_1(\theta) = BI_2(\theta)$ holds. Then we must have

$$\begin{aligned} 0 &= I_1(\theta)^{-1} - B^{-1}I_2(\theta) \\ &= \left(\frac{\delta_1(1-\delta_1)}{(\gamma_1+\delta_1-1)^2} - B^{-1}\frac{\delta_2(1-\delta_2)}{(\gamma_2+\delta_2-1)^2}\right) + \theta\left(\frac{2\delta_1-1}{\gamma_1+\delta_1-1} - B^{-1}\frac{2\delta_2-1}{\gamma_2+\delta_2-1}\right) + \theta^2(1-B^{-1}) \end{aligned}$$

But the above can only hold if $B = 1$. That is, the two tests are the same.

Two tests thus can only perfectly substitutable if they have identical sensitivity and specificity.

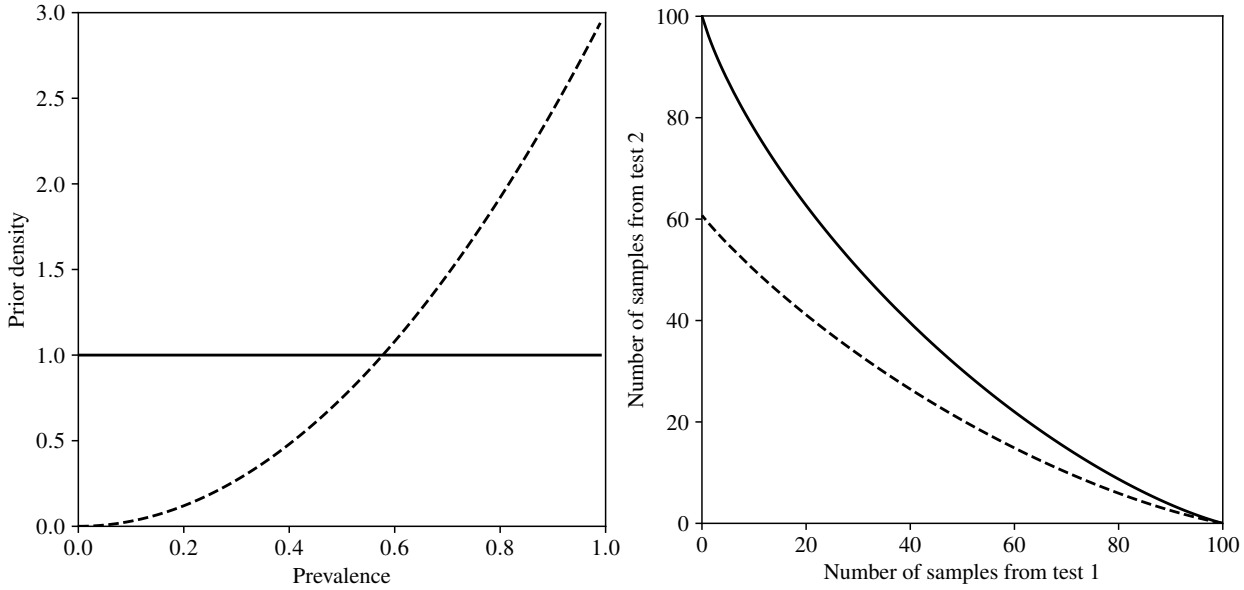

Figure 2.1: Higher sensitivity tests are more preferred when the prevalence is believed to be higher. Test 1 has sensitivity $\gamma = 0.5$ and specificity $\delta = 1$, and test 2 vice-versa. Solid line: Uniform prior and corresponding approximate indifference curve under squared-deviation loss. Dashed line: $B(3, 1)$ prior and corresponding approximate indifference curve.

Otherwise, there always exists prices such that the optimal bundle at large samples uses both tests.

Next, we can use the results to better understand how demand varies with prior. 2.4 illustrates how the approximate "indifference curves" vary as the prior becomes more weighted towards higher prevalences[9]

Observation 2.2 (More sensitive tests are better when prevalence is likely low). *Let $\mathscr{E}_1$ and $\mathscr{E}_2$ have sensitivities and specificities γ_1, δ_1 and γ_2, δ_2 respectively, with $\gamma_1 + \delta_1 = \gamma_2 + \delta_2$. For fixed costs per sample, if prior μ' dominates μ in the first-order stochastic dominance sense, then the limit optimal proportion of samples from $\mathscr{E}_1$ is higher under prior μ'.*

To understand why this holds, observe that higher sensitivity tests have their highest fisher

9 Because the inverse Fisher information is weighted by $l_{aa}(\theta, \theta)\mu(\theta)$, a prior with more weight on higher values of θ is equivalent to $l_{aa}(\theta, \theta)$ increasing at higher values of θ (mistakes are more costly at high prevalence).

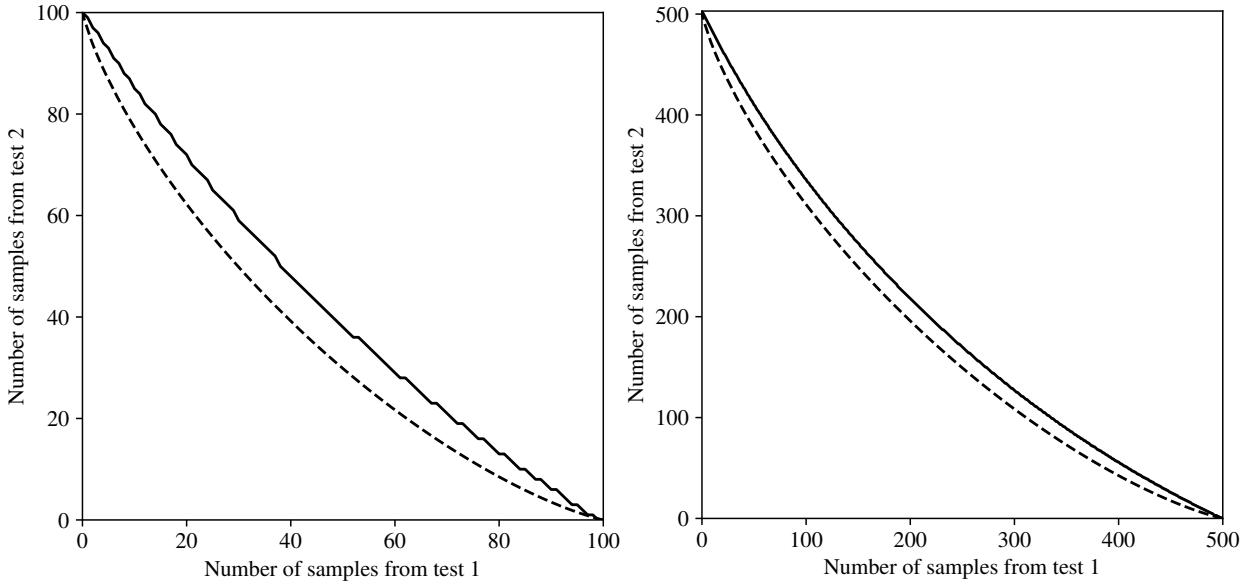

Figure 2.2: Numerical performance of the approximation under a uniform prior and squared-deviation loss at two sample sizes. Dashed line: Iso-expected variance line (approximate "indifference curve"). Solid line: Numerically computed upper contour boundary (bundles above the line are preferred to the lower corner bundle). Test 1 has sensitivity 0.5 and specificity 1, and test 2, vice-versa.

information at higher prevalence values:

$$I_j'(\theta) = (\gamma_j + \delta_j - 1)^3 \left[\frac{2p_j^+(\theta) - 1}{p_j^+(\theta)^2(1 - p_j^+(\theta))^2} \right]$$
$$> 0 \quad \text{if and only if } p_j^+(\theta) > 0.5$$

Holding fixed $\gamma + \delta - 1$, the probability of a positive test outcome is increasing with sensitivity, and thus the relatively higher sensitivity test performs relatively better when the true prevalence is higher.

Finally, we can use this testing example to illustrate the practical rate at which the approximation converges. 2.4 illustrates at two sample sizes the approximate indifference curves against a numerically computed boundary of an upper contour set (true indifference curves are generically singletons in the discrete setting). Note that as in Chapter 1, the approximation lies below the true boundary, suggesting the approximate loss underestimates the true loss.

In Chapter 1, this underestimation occurred primarily because the approximation neglected

all mistakes but the most common one. Here the approximation only considers small deviations of the Bayes action, and thus neglects loss that occur from large deviations of the action. In both cases, however, the neglected portion of expected loss is eventually vanishingly small.

Lastly, note that because the weighted average inverse Fisher information necessarily uses information about the prior and loss function, it may often perform practically better than the approximation for the finite-state model.

2.5 CONCLUSION

This chapter has shown that demand for information in continuous-state/continuous-action problems is well-approximated by minimizing an appropriate weighted average of the Fisher information. This result generalizes the well-known expected variance A-criterion from the experimental design literature, and considers implications in a consumer theory framework.

In particular, demand at large samples behaves as though preferences are smooth and convex, as in a standard consumer theory setting. Using the approximation, I then derive an exact-at-high-samples expression for the rate at which samples from tests may be substituted.

The results were then illustrated in a setting of estimating disease prevalence, showing that for distinct tests, interior solutions always exist at large sample for the appropriate choice of test costs.

In practice, the approximations here are likely to perform better than those from Chapter 1 because they account for preferences and prior—factors ignored by the large-deviations approach from before; however, future work is required to constrain the convergence rate of the asymptotic errors.

In sum, the results in both Chapters 1 and 2 develop a method for understanding the consumer theory of information from differing sources using methods from asymptotic statistics.

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

2.A OMITTED PROOFS

2.A.1 Regularity Conditions

In order to avoid pathological outcomes such as moments not existing, or distributions converging in the limit to ones with moments not existing, we need a number of regularity conditions. These conditions are sufficient conditions for the regularity conditions assumed by Strasser (1975), and will guarantee that the Bayes optimal action will asymptotically converge to a Gaussian (Lemma 2.1).

Assume all experiments satisfy the following regularity conditions:

(E1) (Experiments are minimally informative) For any compact K, $\inf_{\theta \in K} I_j(\theta) > 0$.

(E2) (Unique maximum likelihood estimator) $\log f_j(x|\theta)$ is concave

(E3) (Differentiability) $f_j(x|\theta)$ is continuously three-times differentiable in θ.

(E4) (Thin-tailed log-likelihood)[10] For any θ and compact K

$$\sup_{\vartheta \in K} \int (\log f_j(x|\theta))^2 f_j(x|\vartheta) dx < \infty$$

(E5) (Other integrability requirements) For every compact K,

$$\sup_{\theta \in K} \int \left| \frac{\partial}{\partial \theta} f(x|\theta) \right|^3 f(x|\theta) dx < \infty$$
$$\sup_{\theta \in K} \int \left| \frac{\partial^2}{\partial \theta^2} f(x|\theta) \right|^3 f(x|\theta) dx \leq \infty$$

Importantly, note that if two experiments satisfy the above conditions, any experiments composed of samples from those experiments satisfies it as well.

Similarly, we will require l to have a number of regularity conditions to guarantee that the

10 This necessarily requires that the set of outcomes that perfectly rule in or out any state is probability zero in all states.

asymptotically optimal Bayes action is unique, and that realized losses have finite moments under the posterior.

Assume l satisfies

(L1) (Correct estimates are uniquely optimal) l is convex and $l(\theta,\theta) = 0$, $l(a,\theta) > 0$ for $a \neq 0$.

(L2) (Differentiability) l is three-times differentiable in both a and. θ

(L3) (Finite expectations) For any a, $l(a,\theta)$ is integrable with respect to the prior, μ.

(L4) (Locally bounded curvature)For θ restricted to any compact K, $l_{aa}(\theta,\theta)$ is bounded.

(L5) (Absolutely integrable derivatives) For any compact K,

a) $\sup_{a\in K} \int |l_a(a,\theta)|\mu(\theta)d\theta < \infty$

b) $\sup_{a\in K} \int |l_{aa}(a,\theta)|\mu(\theta)d\theta < \infty$

2.A.2 Proof of Lemma 2.1

We wish to show

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

under state θ_0, where $\hat{\theta}$ is the Bayes optimal action.

Rearranging terms we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\hat{\theta}_{MLE} - \theta_0) + \sqrt{n}(\hat{\theta} - \hat{\theta}_{MLE})$$

where $\hat{\theta}_{MLE}$ is the maximum likelihood estimate.

By standard asymptotic theory, $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$, so by application of Slutsky's theorem, it suffices to show $\sqrt{n}(\hat{\theta} - \hat{\theta}_{MLE}) \xrightarrow{d} 0$.

This convergence of the Bayes estimator to the maximum likelihood is well-known for squared-deviation loss (See Theorem 8.3 in Lehmann and Casella, 1998). For more general differentiable losses, we can apply the following result of Strasser (1975):

Lemma (Strasser, 1975). *Let the prior density μ be continuously differentiable, experiment $\mathcal{E}$ satisfy conditions (E), and the loss function, l, satisfy conditions (L). Then for any compact K, there exists*

c_K such that

$$\sup_{\theta \in K} \mathbb{P}_\theta^n \left\{ \mathbf{x}_n \middle| |\hat{\theta} - \hat{\theta}_{MLE}| \geq c_K \frac{\log(n)}{n} \right\} = o\left(\frac{1}{\sqrt{n}}\right)$$

where $\mathbb{P}_\theta^n$ denotes the probability under state θ with n samples from $\mathscr{E}$.

Clearly, the above lemma implies $\sqrt{n}(\hat{\theta} - \hat{\theta}_{MLE}) \xrightarrow{p} 0$ (in fact, very quickly).

The claim follows immediately then because convergence in probability implies convergence in distribution. □

2.A.3 Proof of Lemma 2.2

First, notice by Taylor's theorem we can write

$$l(\hat{\theta}, \theta) = l(\hat{\theta}, 0) - l(\theta, \theta) = \frac{1}{2}(\hat{\theta} - \theta)^2 l_{aa}(\theta, \theta) + \frac{1}{6}(\hat{\theta} - \theta)^3 l_{aaa}(\tilde{\theta}, \theta)$$

for some $\tilde{\theta}$ between $\hat{\theta}$ and θ. Now, let $y = nI(\theta)(\hat{\theta} - \theta)^2$. So we can then write the above equation

$$nI(\theta) l(\hat{\theta}, \theta) = \frac{1}{2} y l_{aa}(\theta, \theta) + \frac{1}{6} y(\hat{\theta} - \theta) l_{aaa}(\tilde{\theta}, \theta)$$

Because $\sqrt{nI(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$ by Lemma 2.1, we must have $y \xrightarrow{d} \chi_1^2$. Further, the second term converges in distribution to zero by Slutsky's theorem because $(\hat{\theta} - \theta) \xrightarrow{d} 0$. Applying Slutsky's theorem again, we then have

$$\frac{2nI(\theta) l(\hat{\theta}, \theta)}{l_{aa}(\theta, \theta)} \xrightarrow{d} \chi_1^2$$

Taking expectations—guaranteed to exist by regularity conditions (E) and (L)—we have

$$nR(\theta) = \frac{1}{2} I(\theta)^{-1} l_{aa}(\theta, \theta) + o(1)$$

Rearranging terms gives the result. □

2.A.4 Proof of Proposition 2.2

As with the proof of Proposition 1.1, I send budget $Y \to \infty$ because that is equivalent to sending all costs to zero at the same rate. First note that for a fixed $\mathbf{r}$, if the DM purchases the bundle on the budget frontier with proportions closest to $\mathbf{r}$, her "payoff" will be

$$U(\mathbf{r}, Y) = \frac{Y}{2} \frac{\left(\int \frac{1}{I_{\mathbf{r}}(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta\right)^{-1}}{(\mathbf{r} \cdot \mathbf{c})}(1 + o(1))$$

as $Y \to \infty$. So by Proposition 2.1, for fixed $\mathbf{r}$, spending the entire budget on a bundle with proportions as close as possible to $\bar{\mathbf{r}}$ maximizing payoff per dollar

$$\bar{\mathbf{r}} = \arg\max_{\mathbf{r}} \left\{ \left(\int \frac{1}{I_{\mathbf{r}}(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta \right)^{-1} \Big/ \mathbf{r} \cdot \mathbf{c} \right\}$$

will eventually yield lower expected than losses than $\mathbf{r}$.

Now let $\mathbf{r}(Y)$ be the expected-loss-minimizing proportions for budget Y. We wish to show $\mathbf{r}(Y) \to \bar{\mathbf{r}}$ as $Y \to \infty$. For contradiction, suppose not. Let Y_k be an increasing sequence such that $\mathbf{r}(Y_k)$ converges to $\mathbf{r}$.

By the above logic, for high enough Y, $\mathbf{r}$ performs strictly worse than $\bar{\mathbf{r}}$. In particular we must have for Y large enough we must have

$$U(\bar{\mathbf{r}}, Y) - U(\mathbf{r}, Y) \geq \epsilon > 0$$

But because $\mathbf{r}(Y_k) \to \mathbf{r}$, we must also have

$$\begin{aligned} U(\mathbf{r}(Y_k), Y_k) &= \frac{Y}{2} \left(\frac{\left(\int \frac{1}{I_{\mathbf{r}}(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta\right)^{-1}}{(\mathbf{r} \cdot \mathbf{c})} + o(1) \right)(1 + o(1)) \\ &= U(\mathbf{r}, Y)(1 + o(1)) \end{aligned}$$

as $k \to \infty$. Thus, for k high enough, we must have

$$U(\bar{\mathbf{r}}, Y_k) - U(\mathbf{r}(Y_k), Y_k) \geq \epsilon > 0$$

which contradicts the fact that $\mathbf{r}(Y_k)$ was the optimal relative proportions at budget Y_k.

Thus, as $Y \to \infty$, we must have $\mathbf{r}(Y) \to \infty$ as claimed. □

2.A.5 Proof of Proposition 2.4

We proceed similarly to the proof of Proposition 1.4A. To simplify notation, I work with the case of two experiments; however the proof logic easily applies with more than two experiments.

If k_2 is the minimum number of additional samples from $\mathscr{E}_2$ required to just compensate for a loss of k_1 samples from $\mathscr{E}_1$, we must have

$$\int \frac{1}{n_1 I_1(\theta) + n_2 I_2(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta \geq \left(\int \frac{1}{(n_1 - k_1) I_1(\theta) + (n_2 + k_2) I_2(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta \right)(1 + o(1))$$

where $o(1)$ is as $n_1 + n_2 \to \infty$.

Because the radius of convergence for the Taylor series of $1/x$ around $x = x_0$ is x_0, we can write

$$\begin{aligned}\int \frac{1}{(n_1 - k_1) I_1(\theta) + (n_2 + k_2) I_2(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta &= \int \frac{1}{n_1 I_1(\theta) + n_2 I_2(\theta)} l_{aa}(\theta, \theta)\mu(\theta)d\theta \\ + \int \frac{k_1 I_1(\theta)}{(n_1 I_1(\theta) + n_2 I_2(\theta))^2} l_{aa}(\theta, \theta)\mu(\theta)d\theta &- \int \frac{k_2 I_2(\theta)}{(n_1 I_1(\theta) + n_2 I_2(\theta))^2} l_{aa}(\theta, \theta)\mu(\theta)d\theta + o\left(\frac{1}{(n_1 + n_2)^2}\right)\end{aligned}$$

Plugging this in to the previous inequality and rearranging, we have

$$0 \geq \left(\int \frac{k_1 I_1(\theta)}{(n_1 I_1(\theta) + n_2 I_2(\theta))^2} l_{aa}(\theta, \theta)\mu(\theta)d\theta - \int \frac{k_2 I_2(\theta)}{(n_1 I_1(\theta) + n_2 I_2(\theta))^2} l_{aa}(\theta, \theta)\mu(\theta)d\theta \right)(1 + o(1))$$

Rearranging once again, we have

$$k_2 \geq k_1 \left(\int \frac{I_1(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \Big/ \int \frac{I_2(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \right) + o(1)$$

Repeating the above exercise using that fact that $L(k_1, k_2) \leq L(k_1, k_2 - 1)$ gives

$$k_2 \leq k_1 \left(\int \frac{I_1(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \Big/ \int \frac{I_2(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \right) + 1 - \frac{1}{k_1} + o(1)$$

Thus for $n_1 + n_2$ high enough, k_2 is squeezed to be

$$k_2 = \left\lceil k_1 \left(\int \frac{I_1(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \Big/ \int \frac{I_2(\theta)}{I_{\mathbf{r}}(\theta)^2} l_{aa}(\theta, \theta) \mu(\theta) d\theta \right) \right\rceil$$

as claimed. □

2.B MULTIVARIATE RESULTS

In order to generalize the results to multivariate settings, we need only use the multivariate version of Taylor's theorem for the loss approximation. Namely,

$$l(a, \theta) = \frac{1}{2}(a - \theta)^T H_a l(\theta, \theta)(a - \theta) + o(||a - \theta||)$$

where $H_a l$ denotes the Hessian of $l(\cdot, \theta)$.

We can then use the multivariate version of Lemma 2.1: $\sqrt{n}(a - \theta) \xrightarrow{D} \mathcal{N}(0, I(\theta)^{-1})$ where $I(\theta)$ is now the Fisher information *matrix*. We can then write the state-θ risk as

$$R_n(\theta) = \frac{1}{2n} \mathrm{tr}(H_a l(\theta, \theta) I(\theta)^{-1})(1 + o(1))$$

Taking expectations with respect to the prior we have

$$L(n) = \frac{1}{2n} \int \text{tr}(H_a l(\theta, \theta) I(\theta)^{-1}) \mu(\theta) d\theta (1 + o(1))$$

Note that this expression is essentially identical to the usual statement of the A-criterion for experiment design. In that setting, losses are assumed to be quadratic, with $l(a, \theta) = (a-\theta)^T A(a-\theta)$ for some positive-semidefinite matrix A. See Clyde (1993) for the derivation in this setting. Here we simply replace A by the Hessian of the loss function.

So at large samples we have that the optimal proportions $\mathbf{r}$ will approximately minimize

$$L(\mathbf{n}) = \frac{1}{2N} \int \text{tr} \left(H_a l(\theta, \theta) \left(\sum_{j=1}^{j} r_j I_j(\theta) \right)^{-1} \right) \mu(\theta) d\theta$$

BIBLIOGRAPHY

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

Chaloner, K. and I. Verdinelli (1995): "Bayesian Experimental Design: A Review," *Statistical Science*, 10.

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, 493–507.

Clyde, M. A. (1993): "Bayesian optimal design for approximate normality," Ph.D. thesis, University of MInnesota.

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

Daniels, H. E. (1954): "Saddlepoint Approximations in Statistics," *The Annals of Mathematical Statistics*, 25, 631–650.

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

Gil-Pelaez, J. (1951): "Note on the Inversion Theorem," *Biometrika*, 38, 481.

Jeffreys, H. and B. Jeffreys (1956): *Methods of Mathematical Physics*, Cambridge University Press, 3rd ed.

Keppo, J., G. Moscarini, and L. Smith (2008): "The demand for information: More heat than light," *Journal of Economic Theory*, 138, 21–50.

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

Lehmann, E. L. and G. Casella (1998): *Theory of Point Estimation*, Springer New York, 2nd ed.

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

Moscarini, G. and L. Smith (2002): "The law of large demand for information," *Econometrica*, 70, 2351–2366.

Pukelsheim, F. (1993): *Optimal design of experiments*, New York: John Wiley & Sons.

Radner, R. and J. E. Stiglitz (1984): "A nonconcavity in the value of information," .

Strasser, H. (1975): "The asymptotic equivalence of Bayes and maximum likelihood estimation," *Journal of Multivariate Analysis*, 5, 206–226.

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

van der Vaart, A. W. (1998): *Asymptotic statistics*, Cambridge University Press.