Overabundant Information and Learning Traps*

Annie Liang† Xiaosheng Mu‡

October 24, 2017

Abstract

We study a model of sequential learning, where agents choose what *kind* of information to acquire from a large, fixed set of Gaussian signals with arbitrary correlation. In each period, a short-lived agent acquires a signal from this set of sources to maximize an individual objective. All signal realizations are public. We study the community's asymptotic speed of learning, and characterize the set of sources observed in the long run. A simple property of the correlation structure guarantees that the community learns as fast as possible, and moreover that a "best" set of sources is eventually observed. When the property fails, the community may get stuck in an inefficient set of sources and learn (arbitrarily) slowly. There is a specific, diverse set of possible final outcomes, which we characterize.

1 Introduction

Individuals have access to more sources of information than they can devote attention to. Thus, they must decide *which* sources to listen to. When individuals choose sources of information, what are the externalities on future agents, especially on their signal acquisitions? Will a community of short-lived agents eventually choose to acquire the "best" sources of information, or can its eventual demand for information focus on self-reinforcing sources that yield inefficiently slow learning?

*We thank Vasilis Syrgkanis for insightful comments in early conversations about this project. We are also grateful to Aislinn Bohren, Ben Golub, and Yuichi Yamamoto for suggestions that improved the paper.

†University of Pennsylvania

‡Harvard University

We study these questions within a model of sequential learning, with the new feature that individuals choose what kind of signal to observe out of a large set of information sources. There is a unidimensional payoff-relevant state, and additionally $K-1$ possible *biases* or *confounding terms*. Sources are modeled as different linear combinations of these K unknown states, plus an independent Gaussian error. We focus mainly on "informational overabundance," a situation in which not all sources must be observed to learn the payoff-relevant state.

Agents, indexed by $t \in \mathbb{N}$, move sequentially. Each agent chooses a source from which to acquire information (modeled as observation of an independent realization of that signal) and then chooses an action to maximize an individual objective. We assume that objectives may differ across agents. Each agent's choices are based on all signal choices and realizations thus far. Thus, all information is *public* in our setting, in contrast to the classic sequential learning model (Banerjee, 1992; Bikhchandani, Hirshleifer and Welch, 1992; Smith and Sorenson, 2000). Our main results characterize two (exhaustive) long-run outcomes, and the forces that determine which obtains. In certain informational environments, communities with different priors eventually observe different—potentially sub-optimal—sets of sources (in this case, we characterize the potential long-run observation sets). In others, *all* communities, irrespective of prior belief, eventually observe the same "best" set of sources. Which of these outcomes obtains turns out to depend on a simple property of the correlation structure across sources.

To fix ideas, consider agents who want to learn about a political leader's principles. There are many kinds of information that can shed light on this question: for example information about the leader's breaches of executive power, about his mis-management of domestic crises, and about his private indiscretions. Importantly, there are externalities to information acquisition: current information acquisitions affect what information future agents would like to learn. For example, news that the political leader requested a government official to drop a sensitive investigation may inspire interest in the fine details of "obstruction of justice." Alternatively, evidence of private indiscretions can spur enduring interest in the personal lives of the leader and his family. These topics differ in how ultimately revealing they are about the political leader, and they also differ in how easy they are to understand. Indeed, it may be that information that is most potentially revealing—that the political leader did in fact overreach executive power, for example—is poorly understood without prior acquisition of information explaining the limits of executive power.

Two long run outcomes may occur. First, the community may eventually converge to acquisition of information about the "most revealing" topics. Alternatively, acquisition of easily understood information may distract from issues of greater consequence and informativeness; indeed, the public may become "trapped" acquiring information about different angles of a relatively uninformative issue. Our results characterize the role of the correlation across topics in determining which of these outcomes obtains.

Our main analysis characterizes the evolution of society's acquisitions. Formally, for each period t, define a count vector $m(t) = (m_1(t), \ldots, m_N(t))$, where each $m_i(t)$ is the number of times that source i has been observed prior to time t. Our focus is on the *asymptotic frequency vector* $\lim_{t\to\infty} m(t)/t$ describing how often each source is observed in the long run, and also the implied *asymptotic speed of learning*, meaning the speed at which the community learns the payoff-relevant unknown.

A key feature of "overabundant" sources is that there is a multiplicity of ways to learn the payoff-relevant state. We refer to each (minimal) set that reveals the state as a *minimal spanning set*. Although asymptotic learning of the payoff-relevant state occurs if agents exclusively observe any minimal spanning set, the speed of learning can differ substantially across such sets.

We evaluate welfare by comparing society's acquisitions against an "optimal" benchmark, which we construct as follows. First, we show that for every number of observations t, there is an optimal division $n(t) = (n_1(t), \ldots, n_N(t))$ of t observations across signals, where $n_i(t)$ is the number of counts of signal i. This allocation is more informative (in the Blackwell sense) than any other allocation of t observations. Taking the limit $\lim_{t\to\infty} n(t)/t$ yields *optimal asymptotic frequencies*. We show that if there is a uniquely "best" minimal spanning set (a generic case that we define), then this limit is well-defined, and admits a simple closed-form expression. This optimal asymptotic frequency vector has several properties of independent interest. First, only sources that belong to the "best" minimal spanning set are observed with positive frequency in the long run. Second, a comparative static result shows that conditional on being viewed with strictly positive frequency, each signal's asymptotic frequency is (locally) decreasing in that its precision. Loosely, this means that sources are most frequently observed if they are *least informative* within the *most informative set*.

We turn next to our main analysis regarding whether long-run information acquisitions converge to the optimal asymptotic frequencies described above. We show that this outcome depends critically on whether there exists a minimally spanning

set that is of lower-dimension than the state space. The key intuition refers back to Sethi and Yildiz (2016); recall that an agent who observes a biased source learns *both* about the payoff-relevant state and also about the source's own bias. When biases across sources are correlated, there is a further spillover effect: learning from a biased source helps agents to understand the biases of all sources that are correlated with it.[1] Suppose now that agents repeatedly observe a set of sources that is of "full rank," meaning that the sources collectively reveal all K unknown states. Then, every time an agent observes a source from this set, he also improves society's understanding about all sources that are outside of the set. It can be shown that eventually agents come to evaluate sources by objective asymptotic values (which are independent of the community's prior). We thus present the following positive result: if every minimal spanning set is of full rank, then long-run acquisitions are optimal, independently of the prior belief.

In contrast, if there is some minimal spanning set of lower dimension, then inefficient long-run learning may obtain. Intuitively, continued observation of sources from a set of dimension $k < K$ provides limited positive spillovers for sources outside of the set. This is because agents can at most learn k unknown states from these sources, while the other sources may depend on the remaining $K-k$ states. Thus, the community's understanding of sources outside of the set need not improve. Formally, we show that for every minimal spanning set that is "best" in a lower-dimensional subspace, there is an open set of priors such that this set is observed in the long run. The implied inefficiency—measured as the ratio of the optimal speed of learning and the achieved speed of learning—can be an arbitrarily large constant.

Our work combines ideas from two literatures. First, recent work (Sethi and Yildiz, 2016; Che and Mierendorff, 2017; Fudenberg, Strack and Strzalecki, 2017; Liang, Mu and Syrgkanis, 2017; Mayskaya, 2017; Sethi and Yildiz, 2017) studies choice of information from a finite set of information sources. We build specifically upon our prior paper Liang, Mu and Syrgkanis (2017), which characterized optimal signal acquisitions from correlated Gaussian sources under the assumption of "exact-identification" (all sources must be observed to recover the state). Our work also builds on Sethi and Yildiz (2016, 2017), which study long-run acquisitions from a large number of Gaussian sources. There are a few key modeling differences: first, the related papers consider stochastic error variances, so that the "best" sources vary from

[1] This appears also in Sethi and Yildiz (2017), which studies a model in which biases are correlated across sources within a group, but not across groups.

period to period, while we fix noise variances, so that there is (generically) a unique "best" asymptotic set; second, Sethi and Yildiz (2016, 2017) focus on correlation structures that fall under our Theorem 2, for which long-run acquisitions do not necessarily achieve efficient learning, while we explore also those correlation structures that lead to optimal learning.[2] Thus, the welfare comparisons that we make here are particular to our framework.

Finally, our model contributes to the social learning literature. This literature has focused on the classic friction that decision-makers only observe coarse summary statistics of past information acquisitions. We assume instead that all information is perfectly passed on to future agents. It is immediate that asymptotic learning will occur, but we show that learning can be inefficiently slow when agents choose what kind of information to observe. Our paper relates in particular to Burguet and Vives (2000), Ali (2017), and Mueller-Frank and Pai (2016), who introduced costly information acquisition to the sequential learning model. We consider here choice from a set of information sources, and demonstrate the role of correlations across sources in determining speed of learning.

2 Model

There are K persistent states $\theta_1, \ldots, \theta_K \sim \mathcal{N}(\mu^0, V^0)$, where the prior covariance matrix V^0 is of full rank.[3] Agents have access to N different sources of information, and observation of source i corresponds to an independent realization of the signal

$$X_i^t = \langle c_i, \theta \rangle + \epsilon_i^t, \quad \epsilon_i^t \sim \mathcal{N}(0,1).$$

Each $c_i = (c_{i1}, \ldots, c_{ik})'$ is a constant $K \times 1$ vector and $\theta = (\theta_1, \ldots, \theta_K)'$ is the vector of unknown states. The noise terms ϵ_i^t are independent from each other and over time. Our assumption that noise terms have unit variance is without loss since the coefficients c_i are unrestricted. Allowing $N > K$ is key: this means that observation of all of the sources is not necessary for asymptotic learning of θ_1. We let C denote the $N \times K$ matrix whose i-th row is c_i'.

A countably infinite number of agents, indexed by $t \in \mathbb{N}$, moves sequentially. Each agent t acquires an independent realization of one of the N signals, and then

[2] Specifically, Sethi and Yildiz (2016) focuses on signals with independent biases, and Sethi and Yildiz (2017) focuses on signals that can be partitioned into groups (see Section 6).

[3] This rules out linear dependence across the states. If indeed some states are linearly dependent, we may work with a smaller set of states without changing the model.

chooses an action $a \in A$ to maximize an individual objective $u_t(a, \theta)$. He bases his action on his own signal acquisitions, as well as the history of signal acquisitions and realizations thus far. (Thus, all signal realizations are public.)

Payoff functions may differ across agents, but we impose the following restrictions. First, there is a single payoff-relevant state.

Assumption 1 (Single Payoff-Relevant State)**.** *For every t,*

$$u_t(a, \theta_1, \theta_{-1}) = u_t(a, \theta_1) \quad \textit{does not depend on } \theta_{-1}.$$

Thus, states $\theta_2, \ldots, \theta_K$ are not directly payoff-relevant. However, agents maintain beliefs over the complete K-dimensional state vector, since the payoff-irrelevant states can be important for interpreting signal realizations.

Additionally, we assume that the decision problems are non-trivial in the following way.

Assumption 2 (Payoff Sensitivity to Mean)**.** *For every t, any variance $\sigma^2 > 0$ and any action $a^* \in A$, there exists a positive measure of μ_1 for which a^* does not maximize $\mathbb{E}[u_t(a, \theta_1) \mid \theta_1 \sim \mathcal{N}(\mu_1, \sigma^2)]$.*

In words, holding the belief variance fixed, the expected value of θ_1 affects the optimal action to take.

A sufficient condition for Assumption 2 is that for every agent t and every action a^*, there exists some other action $\hat{a}$ such that $u_t(\hat{a}, \theta_1) > u_t(a^*, \theta_1)$ as $\theta_1 \to +\infty$ or as $\theta_1 \to -\infty$. That is, we require that the two limiting states $\theta_1 \to +\infty$ and $\theta_1 \to -\infty$ yield different optimal actions. This is true for all natural applications.

We use throughout the key concept of *spanning sets*. Let $[N] = \{1, \ldots, N\}$ index the set of signals and let $[K] = \{1, \ldots, K\}$ index the set of K states. A set of signals $\mathcal{S} \subset [N]$ is *spanning* or a *spanning set* if the vectors $\{c_i : i \in \mathcal{S}\}$ span e_1. Thus, it is possible to learn θ_1 by exclusively observing signals from $\mathcal{S}$. The set $\mathcal{S}$ is *minimally-spanning* or a *minimal spanning set* if it is spanning, and moreover has the property that no proper subset is spanning. Note that if a set $\mathcal{S}$ is minimally spanning, then it contains no more than K signals, and moreover we can write

$$e_1 = \sum_{i \in \mathcal{S}} \beta_i \cdot c_i$$

for (unique) nonzero coefficients β_i.

We assume throughout that θ_1 is revealed by the full set of signals.

Assumption 3. *The complete set of signals* $[N]$ *is spanning.*

This assumption allows for two interesting cases. Say that payoff-relevant state θ_1 is *exactly identified* if $[N]$ is minimally spanning. Additionally, say that payoff-relevant state θ_1 is *overidentified* if $[N]$ is spanning but not minimally spanning. Except for trivial cases, this latter setting corresponds to $N > K$.[4]

2.1 Interpretations and Examples

We mention below a few interpretations for this informational model.

Correlated biases. The most straightforward interpretation takes $\theta_2, \ldots, \theta_K$ (the states that are not payoff-relevant) to be different *biases*. In examples throughout this paper, we often invoke this interpretation more explicitly by relabeling the states $\theta_2, \ldots, \theta_K$ as $b_1, \ldots, b_{K-1}$. Then, each source provides a biased signal about θ_1, where the biases are potentially correlated across sources. For example, news sources CNN, NYTimes, and MSNBC share a left-leaning bias, but to different degrees. The coefficient matrix C determines the precise structure of this correlation.

Groups. A special kind of correlation is one in which signals can be partitioned into different groups, with group-specific unknowns.

Example 1. The unknown states are $\theta_1, \theta_2, \theta_3$, where only θ_1 is payoff-relevant. The sources are

$$\begin{aligned} X_1 &= \theta_1 + \theta_2 + \epsilon_1 \\ X_2 &= \theta_2 + \epsilon_1 \\ X_3 &= \theta_1 + \theta_3 + \epsilon_3 \\ X_4 &= \theta_3 + \epsilon_4 \end{aligned}$$

Then, there are two "groups" of sources, each of which is associated with a group-specific unknown. For example, the states θ_2 and θ_3 may represent comprehension of language or culture corresponding to the respective group. An agent who does not understand the language of the first group perceives X_1 to be a noisy signal about θ_1

[4]It is possible for θ_1 to be overidentified from a set of $N \leq K$ signals, e.g. $X_1 = \theta_1 + \epsilon_1$, $X_2 = \theta_1 + \theta_2 + \theta_3 + \epsilon_2$, and $X_3 = \theta_2 + \theta_3 + \epsilon_3$. In this case, the set $\{X_1, X_2, X_3\}$ is spanning, but not minimally spanning since both of its subsets $\{X_1\}$ and $\{X_2, X_3\}$ are also spanning. Although $N = K = 3$ in this example, it is equivalent to a model in which there are two states θ_1 and $\tilde{\theta}_2 = \theta_2 + \theta_3$, and the three signals are rewritten $X_1 = \theta_1 + \epsilon_1$, $X_2 = \theta_1 + \tilde{\theta}_2 + \epsilon_2$ and $X_3 = \tilde{\theta}_2 + \epsilon_3$. Then, we do have $N > K$ in this alternative model.

(because of the large variance on θ_2). As observations of signal X_2 accumulate (which improve understanding of θ_2), the informativeness of signal X_1 increases. Observe that the state θ_1 is revealed by signals from either group.

Composite of unknowns. A third interpretation takes the payoff-relevant state θ_1 to be a linear combination of unknowns $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ about which the agents can learn independently. In this case, we can interpret sources as experts with different specializations.

Example 2. The unknown states are $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$, where $\tilde{\theta}_1 + \tilde{\theta}_2 + \tilde{\theta}_3$ is payoff-relevant. The sources are

$$X_1 = \tilde{\theta}_1 + \epsilon_1$$
$$X_2 = \tilde{\theta}_2 + \epsilon_2$$
$$X_3 = \tilde{\theta}_3 + \epsilon_3$$

Then, source 1 is an expert regarding $\tilde{\theta}_1$, source 2 is an expert regarding $\tilde{\theta}_2$, and source 3 is an expert regarding $\tilde{\theta}_3$.

Linear best responses. A final interpretation micro-founds the signals as actions taken by another set of agents. Specifically, suppose that there are N "types" of agents (corresponding to the N sources). Each period, a new agent of each type i is born and receives a (Gaussian) signal

$$\theta + \eta_i,$$

where η_i is a K-dimensional vector of independent standard normal noise terms. He then takes an action a_i to match the private objective

$$-(a_i - \langle c_i, \theta \rangle)^2)$$

Then, each agent i's best response follows the distribution $\langle c_i, \theta \rangle + \epsilon_i$ where $\epsilon_i = \langle c_i, \eta_i \rangle$. It is important in this interpretation that agents are not long-lived, so that the distribution of best responses does not change.

3 Information Acquisition

Let $[N] = \{1, 2, \ldots, N\}$ denote the set of all signals. Each agent faces a *history* $h^{t-1} \in ([N] \times \mathbb{R})^{t-1} = H^{t-1}$ consisting of all past signal choices and realized signal

values. A *strategy* for the agent moving at time t is a measurable map from all $(t-1)$-length histories to signals—that is, $S : H^{t-1} \to [N]$, where $S(h^{t-1})$ represents the signal choice in period t following history h^{t-1}.[5]

Each agent's marginal belief about θ_1, updated to his own signal acquisition and all information revealed by past agents, is Gaussian. Write $\theta_1 \sim \mathcal{N}(\mu_1, V_{11})$ for this belief. His maximum payoff is

$$\max_{a \in A} \mathbb{E}[u_t(a, \theta_1) \mid \theta_1 \sim \mathcal{N}(\mu_1, V_{11})] \tag{1}$$

Each agent chooses the signal that maximizes (1) in expectation. We use a result from Liang, Mu and Syrgkanis (2017) to characterize this signal choice. First observe that since beliefs are Gaussian, the agent's posterior variance V_{11} about θ_1 following q_i observations of each signal i can be written as a deterministic function

$$V_{11} = f(q_1, \ldots, q_N).$$

In particular, the posterior variance does not depend on signal realizations; see Appendix B for the complete (closed-form) expression. It was shown in Liang, Mu and Syrgkanis (2017) that the signal that yields the greatest reduction in posterior variance *Blackwell dominates* the remaining signals. Thus, the signal choice that achieves the greatest reduction in posterior variance also maximizes expected payoffs.

Lemma 1 (Liang, Mu and Syrgkanis (2017))**.** *The optimal signal acquisition for every agent, at every history, is the signal that minimizes current posterior variance about θ_1.*

Using this lemma, we can track "society's acquisitions" in the following way. Write $m(t) = (m_1(t), \ldots, m_N(t))$ for the *division over signals at time t*, where $m_i(t)$ is the number of times signal i has been observed up to and including time t. Then, $m(t)$ evolves deterministically according to the following rule: $m(0) = \mathbf{0}$ and for $t \geq 0$,

$$m_i(t+1) = m_i(t) + 1 \quad \text{if} \quad f(m_i(t)+1, m_{-i}(t)) \leq f(m_j(t)+1, m_{-j}(t)) \ \forall j.$$

and $m_j(t+1) = m_j(t)$ for all other signals j.[6] We are primarily interested in the *long-run* acquisitions. Specifically, we will refer to the *asymptotic frequency* $\lim_{t\to\infty} m_i(t)/t$ with which source i is observed, and the *asymptotic observation set*, meaning the set

[5] Since information is public, agents do not need to additionally condition on past actions.

[6] We allow ties to be broken arbitrarily, so there may be multiple paths $m(t)$.

of signals that are observed with positive frequency in the long-run. Our subsequent results in Section 5 show that these limits are well-defined.

We will compare these long-run acquisitions against the following "optimal" benchmark. For each t, define

$$\operatorname*{argmin}_{(q_1,\ldots,q_K):q_i\in\mathbb{Z}^+,\sum_i q_i=t} f(q_1,\ldots,q_K).$$

to be the division(s) of t observations that minimizes posterior variance about θ_1. Below, we write $n(t) = (n_1(t),\ldots,n_N(t))$ for a typical optimal division of t observations, where generically $n(t)$ is unique. Applying Lemma 3 from Liang, Mu and Syrgkanis (2017), we have that $n(t)$ Blackwell dominates any other set of t observations and in fact maximizes the expected payoff of agent t. Thus, the sequence $(n(t))_{t\geq 1}$ pointwise (weakly) improves upon any other sequence $(m(t))_{t\geq 1}$, and for this reason we will use it as an optimal benchmark. We also define the *"optimal" asymptotic frequency* with which source i is observed to be $\lim_{t\to\infty} n_i(t)/t$, and the *optimal observation set* to be the signals that have positive optimal asymptotic frequency. These limits are well-defined under a simple condition, which we describe in the next section.

Note that a planner who has control over the agents' signal choices could dictate choosing signals according to their optimal frequencies (modulo adjustments to accommodate discrete time periods). Doing so would ensure that the signal counts at every time t are approximately given by $n(t)$. Thus, for a planner who is trying to maximize a discounted sum of agent's payoffs, the overall discounted payoff approximates the optimal benchmark associated with the sequence $(n(t))_{t\geq 1}$ in the infinitely patient limit. This justifies $n(t)$ as the right benchmark to study.

4 Optimal Benchmark

We begin by characterizing optimal asymptotic acquisitions, which we will subsequently use as a benchmark. We break up this characterization into two cases: in Section 4.1 we discuss the exact identification, where all sources are observed infinitely often; in Section 4.2 we turn to our primary case of interest, where asymptotic learning can occur from a strict subset of sources.

4.1 The Exactly-Identified Case

Suppose first that the number of signals and sources is the same ($N = K$); then, we return to the problem considered in our prior work Liang, Mu and Syrgkanis (2017), where the following was shown:

Proposition 1 (Liang, Mu and Syrgkanis (2017)). *Suppose $N = K$ and θ_1 is exactly identified. Then for $1 \leq i \leq K$, $n_i(t) = \lambda_i^* \cdot t + O(1)$, where*

$$\lambda_i^* = \frac{|[C^{-1}]_{1i}|}{\sum_{j=1}^K |[C^{-1}]_{1j}|}. \tag{2}$$

Note that in this case C is a $K \times K$ square matrix.

To interpret these frequencies, observe the vector identity

$$e_1 = \sum_{i=1}^K [C^{-1}]_{1i} \cdot c_i, \tag{3}$$

This represents the payoff-relevant state as a (unique) linear combination of the available signals. Thus $\lambda_i^* \propto |[C^{-1}]_{1i}|$ is a measure of signal i's contribution in this linear combination.

As a second and related intuition, observe that the random vector consisting of a single realization of each signal can be written

$$Y = (y_1, \ldots, y_K)' = C\theta + \varepsilon$$

where ε is the $K \times 1$ vector of error terms. The best linear unbiased estimate for the state vector is

$$(\hat{\theta}_1, \ldots, \hat{\theta}_K)' = C^{-1}Y. \tag{4}$$

Suppose now that we perturb each realization y_i by δ_i. Then, the estimate in (4) for the payoff-relevant state θ_1 changes by $[C^{-1}]_{1i} \cdot \delta_i$. This means that the larger $|[C^{-1}]_{1i}|$ is, the more $\hat{\theta}_1$ responds to changes in the realization of y_i. So (2) says that signals whose realizations more strongly influence the best linear estimate of θ_1 are observed more often in the long run.

In fact, the above result extends to $N < K$ when θ_1 is exactly identified, under an appropriate transformation of the problem. For example, suppose the signals are

$$X_1 = \theta_1 + \theta_2 + \theta_3$$
$$X_2 = \theta_1 - \theta_2 - \theta_3$$

so that $N = 2$ and $K = 3$. We can define a new state $\tilde{\theta}_2 = \theta_2 + \theta_3$ and rewrite

$$X_1 = \theta_1 + \tilde{\theta}_2$$
$$X_2 = \theta_1 - \tilde{\theta}_2$$

Then $N = K = 2$ in this equivalent model. This transformation applies in general: we can always choose N new states (including θ_1), each a linear combination of the original K states, and re-define the original N signals to be linear combinations of the new N states. This alternative model is equivalent to the original problem, but satisfies the conditions of Proposition 1. Thus, dropping the requirement that $N = K$, we obtain the following corollary:

Corollary 1. *Suppose θ_1 is exactly identified. Write*

$$e_1 = \sum_{i=1}^{N} \beta_i \cdot c_i$$

with nonzero coefficients β_i. Then for $1 \leq i \leq N$,

$$n_i(t) = \frac{|\beta_i|}{\sum_{j=1}^{N} |\beta_j|} \cdot t + O(1).$$

Moreover, the minimum posterior variance after t observations satisfies the following approximation:

$$f(n(t)) = \min_{\sum_{i=1}^{N} q_i = t} f(q_1, \ldots, q_N) \sim \left(\sum_{i=1}^{N} |\beta_i| \right)^2 / t.$$

Here and throughout the text, the notation "$F(t) \sim G(t)$" means $\lim_{t\to\infty} \frac{F(t)}{G(t)} = 1$.

Given this asymptotic formula for the posterior variance, we can interpret the sum $\sum_{i=1}^{N} |\beta_i|$ as representing the *speed of learning*: the *smaller* this sum is, the smaller the posterior variance at large t, and the *faster* society learns.

4.2 The Over-Identified Case

We turn now to our primary case of over-identification, where the number of signals exceeds the number of states ($N > K$).

For each minimal spanning set $\mathcal{S}$, we define the *asymptotic standard deviation*

$$Asd(\mathcal{S}) = \sum_{i \in \mathcal{S}} |\beta_i|.$$

By Corollary 1, agents who optimally choose from signals in $\mathcal{S}$ approximates a posterior variance of $(Asd(\mathcal{S}))^2/t$ at all large times t. Thus the smaller $Asd(\mathcal{S})$ is, the faster society learns. Notice that for any signal of the form

$$X = \alpha\theta_1 + \epsilon$$

the set $\{X\}$ is minimally spanning, and $Asd(\{X\}) = 1/|\alpha|$.

We can extend this definition to arbitrary set of signals $\mathcal{A} \subset [N]$ (not necessarily minimally-spanning) as follows. For any set that contains a minimal spanning set, define

$$Asd(\mathcal{A}) = \min_{\mathcal{S}\subset\mathcal{A}} Asd(\mathcal{S}),$$

where the minimum is taken over all minimal spanning sets $\mathcal{S}$ contained in $\mathcal{A}$. If such $\mathcal{S}$ does not exist (i.e., $\mathcal{A}$ is not itself spanning), we let $Asd(\mathcal{A}) = \infty$. In particular,

$$Asd([N]) = \min_{\mathcal{S}\subset[N]} Asd(\mathcal{S})$$

represents the minimum asymptotic standard deviation achieved by observing only those signals in some minimal spanning set. A priori, it is possible to do better by combining observations from multiple spanning sets. However, it will follow from Theorem 1 below that this is not the case when the following assumption on the coefficient matrix C is met:

Assumption 4 (Unique Minimizer)**.** *$Asd(\mathcal{S})$ has a unique minimizer $\mathcal{S}^*$ among minimal spanning sets $\mathcal{S} \subset [N]$.*

This assumption, which holds for generic coefficient matrices C, says that there is a unique minimal spanning set that maximizes speed of learning.

Under Assumption 4, let us write $e_1 = \sum_{i\in\mathcal{S}^*} \beta_i^* \cdot c_i$. Define the frequencies $\lambda^* \in \Delta^{N-1}$ by

$$\lambda_i^* = \frac{|\beta_i^*|}{\sum_{j\in\mathcal{S}^*}|\beta_j^*|}, \quad \forall\ i \in \mathcal{S}^* \tag{5}$$

and $\lambda_i^* = 0$ for $i \notin \mathcal{S}^*$. Our first theorem is now stated.

Theorem 1. *Suppose that the coefficient matrix C satisfies Unique Minimizer, with $\mathcal{S}^*$ uniquely minimizing $Asd(\mathcal{S})$. Let λ^* be given by (5). Then $n_i(t) \sim \lambda_i^* \cdot t$ for each signal $1 \leq i \leq N$.*[7]

[7]We conjecture that the stronger conclusion $n_i(t) = \lambda_i^* \cdot t + O(1)$ also holds. In Remark 2, we prove this result assuming $|\mathcal{S}^*| = K$.

The conclusion can be loosely interpreted as stating that λ^* is the "most efficient linear representation" of the payoff-relevant state in terms of the signal coefficients.[8]

We point out the following comparative static.

Corollary 2. *Suppose that the coefficient matrix C satisfies Unique Minimizer. Write each signal as $X_i = \alpha\langle c_i, \theta\rangle + \epsilon_i$, so that the precision of signal X_i is increasing in α. Then, either $\lambda_i^* = 0$ or λ_i^* is locally decreasing in α.*

That is, if signal i is viewed with positive frequency in the long run, then its asymptotic frequency is decreasing in its precision. This implies loosely that a signal is most frequently viewed when it is *least informative* subject to being in the *most informative* set.

The necessity of Assumption 4 for Theorem 1 is trivially seen by considering two duplicate sources, for example:

$$X_1 = \theta_1 + \epsilon_1$$
$$X_2 = \theta_1 + \epsilon_2$$

given which all divisions across signals are equally optimal. We show in the example below that it is possible for infinite observations of $N > K$ signals to be *strictly* optimal.

Example 3. There are $K = 3$ states $\theta_1, \theta_2, \theta_3$ independently drawn with prior variances $\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}$. $N = 4$ signals are available, and they are respectively

$$X_1 = \theta_1 + \theta_2 + \epsilon_1$$
$$X_2 = \theta_2 + \epsilon_2 + \epsilon_2$$
$$X_3 = \theta_1 + \theta_3 + \epsilon_3$$
$$X_4 = \theta_3 + \epsilon_4$$

with standard normal errors. Note that the former two signals and the latter two signals are both spanning, and these two sets generate the same asymptotic variance. Thus Assumption 4 is not satisfied.

[8] Specifically, consider the following constrained minimization problem:

$$\min \sum_{i=1}^{N} |\beta_i| \quad \text{subject to} \quad \sum_{i=1}^{N} \beta_i \cdot c_i = e_1.$$

It can be shown by linear programming that the minimum is attained exactly when $\beta_i = \beta_i^*$—that is, when focusing on a single minimal spanning set.

The posterior variance about θ_1 as a function of the number of observations q_1, q_2, q_3, q_4 of each signal type can be derived as follows. First, given q_2 observations of signal X_2 and q_4 observations of signal X_4, posterior variance about θ_2 and θ_3 are $1/(q_2+\beta)$ and $1/(q_4+\gamma)$ respectively. Consider now q_1 additional observations of X_1; this provides the same information about the payoff-relevant state θ_1 as the signal $\theta_1+\epsilon'$, where ϵ' is an independent noise term with variance $\frac{1}{q_1}+\frac{1}{q_2+\beta}$. Similarly, q_3 additional observations of X_3 are equivalent to the signal $\theta_1+\epsilon''$, where ϵ'' is an independent noise term with variance $\frac{1}{q_3}+\frac{1}{q_4+\gamma}$. From this we deduce that posterior variance about θ_1 is

$$f(q_1,q_2,q_3,q_4) = 1 \Bigg/ \left(\alpha + \frac{1}{\frac{1}{q_1}+\frac{1}{q_2+\beta}} + \frac{1}{\frac{1}{q_3}+\frac{1}{q_4+\gamma}}\right).$$

The optimal division vector thus seeks to *maximize*

$$\frac{1}{\frac{1}{q_1}+\frac{1}{q_2+\beta}} + \frac{1}{\frac{1}{q_3}+\frac{1}{q_4+\gamma}} \tag{6}$$

It is useful to rewrite (6) in the following way:

$$\frac{1}{4}\left(q_1+q_2+\beta+q_3+q_4+\gamma - \frac{(q_1-q_2-\beta)^2}{q_1+q_2+\beta} - \frac{(q_3-q_4-\gamma)^2}{q_3+q_4+\gamma}\right).$$

Then, since $q_1+q_2+\beta+q_3+q_4+\gamma = t+\beta+\gamma$ is fixed at any time t, it is equivalent to choose q_1, q_2, q_3, q_4 to minimize the sum of ratios

$$\frac{(q_1-q_2-\beta)^2}{q_1+q_2+\beta} + \frac{(q_3-q_4-\gamma)^2}{q_3+q_4+\gamma}.$$

Ideally, if signals were perfectly divisible, the optimum would be to choose $q_1 = q_2+\beta$ and $q_3 = q_4+\gamma$. But as each q_i is restricted to integer values, this continuous optimum is not feasible whenever β and γ are not integers.

The solution to this integer optimization problem is involved, and the details are relegated to Appendix A.1. To express the solution, we need some additional notation. Let r be the integer that minimizes $|r-\beta|$ (the distance to β) and let s be the integer that minimizes $|s-\gamma|$. Further, let $\langle\beta\rangle$ and $\langle\gamma\rangle$ be the value of these absolute differences. We show that when the parity of t and $r+s$ are the same, the optimal (q_1, q_2, q_3, q_4) satisfy

$$q_1, q_2 \approx \frac{\langle\beta\rangle}{2\langle\beta\rangle + 2\langle\gamma\rangle}\cdot t; \quad q_3, q_4 \approx \frac{\langle\gamma\rangle}{2\langle\beta\rangle + 2\langle\gamma\rangle}\cdot t.$$

and otherwise the optimal (q_1, q_2, q_3, q_4) satisfy

$$q_1, q_2 \approx \frac{\langle\beta\rangle}{2\langle\beta\rangle + 2 - 2\langle\gamma\rangle} \cdot t; \quad q_3, q_4 \approx \frac{1 - \langle\gamma\rangle}{2\langle\beta\rangle + 2 - 2\langle\gamma\rangle} \cdot t.$$

Thus, all four signals are observed with positive frequencies in the long run according to the optimal criterion.

Although the example is involved, its intuition is simple: we'd most like to set $q_1 = q_2 + \beta$ and $q_3 = q_4 + \gamma$, but this is not feasible when β and γ are not integers. Thus, there is inevitably some loss from the ideal case where signals are continuously divisible. This loss turns out to be convex in signal counts, so to minimize total loss, both groups of signals are observed infinitely often.

The conclusion of Theorem 1 fails to hold in a strong sense in the example above, since *all* signals are observed infinitely often. Appendix A provides another example that does not satisfy Assumption 4 where, in contrast, the conclusion of Theorem 1 holds qualitatively. Specifically, the conclusion of the theorem is shown to hold for λ^* defined with respect to *some* set $\mathcal{S}^*$ that minimizes Asd. The difference in these two examples, and in addition the complexity of derivation of the asymptotic frequencies above suggest that characterization of optimal acquisitions is in general difficult without Assumption 4.

5 Main Results

We move on now to our main analysis: characterization of long-run acquisitions, and when these converge to the optimal acquisitions discussed above. We show that whether society's acquisitions $m(t)$ eventually approximate the optimal acquisitions $n(t)$ depends critically on how many signals are required to identify θ_1.

To state our results, we need one more definition. For any spanning set of signals $\mathcal{A}$, let $\overline{\mathcal{A}} \subseteq [N]$ be the set of available signals whose coefficient vectors belong to the subspace spanned by signals in $\mathcal{A}$. Notice in particular that $\overline{\mathcal{A}}$ contains $\mathcal{A}$. We say a minimal spanning set $\mathcal{S}$ is *efficient in its subspace* if it uniquely minimizes Asd among subsets of $\overline{\mathcal{S}}$. For example, if the available signals are

$$X_1 = \theta_1/2 + \epsilon_1$$
$$X_2 = \theta_1 + \epsilon_2$$

then $\{X_1\}$ is a minimal spanning set, but it is not efficient in its subspace.[9]

[9] X_2 belongs to the subspace spanned by X_1, and $Asd(\{X_2\}) < Asd(\{X_1\})$.

Theorem 2. *Suppose minimal spanning set $\mathcal{S}$ is efficient in its subspace. Then, there exists an open set of prior beliefs under which long-run frequencies are strictly positive for signals in $\mathcal{S}$, and zero everywhere else.*[10]

A simpler to interpret special case of this result is the following.

Definition 1. *Say that the coefficient matrix C satisfies* Linear Independence *if $N \geq K$ and every $K \times K$ submatrix of C is of full rank.*

Corollary 3. *Suppose that the coefficient matrix satisfies Linear Independence. For any minimal-spanning set that contains less than K signals, there exists an open set of prior beliefs under which each agent observes a signal from this set.*

Thus, assuming Linear Independence, the possibility of inefficiency hinges on whether there exists a minimal spanning set with fewer than K signals.

The content of this theorem is illustrated in the example below, which shows how sequential information acquisition can become "stuck" in a sub-optimal spanning set.

Example 4. There are two states θ_1, θ_2 and three signals which are

$$X_1 = \theta_1/2 + \epsilon_1$$
$$X_2 = \theta_1 + \theta_2 + \epsilon_2$$
$$X_3 = \theta_1 - \theta_2 + \epsilon_3$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are independent standard Gaussian noise terms. Note that

$$Asd(\{X_1\}) = 2 > 1 = Asd(\{X_2, X_3\})$$

so the latter two signals maximize speed of learning.

However, consider a prior belief such that θ_1, θ_2 are independent, and the variance about θ_2 is larger than 3. Prior to any observations, the first signal $\frac{\theta_1}{2} + \epsilon_1$ has precision $\frac{1}{4}$ about θ_1, whereas the latter two signals $\theta_1 + \theta_2 + \epsilon_2$ and $\theta_1 - \theta_2 + \epsilon_3$ each has less precision. Thus the best choice in the first period is to observe X_1. Since this observation does not affect the variance of θ_2, the same argument shows that *every* agent observes signal 1.

[10] As shown in Liang, Mu and Syrgkanis (2017), these frequencies are the same as the optimal frequencies when only signals in $\mathcal{S}$ are available.

Theorem 2 generalizes this example to show that different priors can lead to different "absorbing sets." We note that the speed of learning from a sub-optimal set can be arbitrarily slower than the optimal speed, even though both have rate $O(1/t)$. Specifically, for any positive number L, there exists an environment in which

$$Asd(\mathcal{S})/Asd(\mathcal{S}^*) > L$$

where $\mathcal{S}$ is the asymptotic observation set and $\mathcal{S}^*$ is the optimal asymptotic observation set. This can be proved by direct construction: modify the example above so that

$$\begin{aligned} X_1 &= \theta_1/2 + \epsilon_1 \\ X_2 &= c\theta_1 + \theta_2 + \epsilon_2 \\ X_3 &= c\theta_1 - \theta_2 + \epsilon_3 \end{aligned}$$

with $c > \frac{L}{2}$. We note that the set of "inefficient" priors (which result in sub-optimal learning) does decrease in size as the level of inefficiency increases.

Converse to Theorem 2, our next result shows that starting from *any* prior, information acquisition eventually concentrates on a set of signals that is most efficient in its subspace. We use an assumption which strengthens Unique Minimizer.

Assumption 5 (Unique Minimizer in Every Subspace)**.** *For any spanning set $\mathcal{A} \subset [N]$, $\operatorname{argmin}_{\mathcal{S} \subset \overline{\mathcal{A}}} Asd(\mathcal{S})$ has a unique solution, where the minimum is taken over minimal spanning sets $\mathcal{S}$.*

This says that in every spanning subspace, there exists a unique minimal spanning subset $\mathcal{S}$ that minimizes asymptotic speed of learning. It is obviously guaranteed if different minimal spanning sets correspond to different values of Asd.

Theorem 3. *Suppose that the coefficient matrix C satisfies Assumption 5. Given any prior belief, long-run frequencies exist for every signal. Moreover, if $\mathcal{S}$ denotes the signals viewed with positive frequencies, then $\mathcal{S}$ is a minimal spanning set that is efficient in its subspace.*

As a special case of this theorem, notice that if every minimal spanning set is of size K, then all minimal spanning sets belong to the same subspace. Furthermore, if Unique Minimizer holds, there can only be one minimal spanning set that is efficient in its subspace, and moreover this is the "best" set (in the sense of Section 4). It is then a simple corollary of the above result that under these conditions, all communities (irrespective of their prior) converge to the optimal asymptotic acquisitions.

Corollary 4. *Suppose that the coefficient matrix C satisfies Unique Minimizer and that every minimal spanning set has size K. Then, starting from any prior belief, it holds that $m_i(t) \sim \lambda_i^* \cdot t, \forall i$.*

One may argue that if coefficient vectors are drawn at random from a full-support distribution over $\mathbb{R}^K$, then it holds with probability 1 that every minimal spanning set is of size K. While this consideration shows the efficiency result in Corollary 4 holds "generically," this notion of genericity ignores the fact that in many economic situations, information sources are not determined by a random process. Indeed, if we expect that sources are endogenous to design or strategic motivations, then important informational environments may be "non-generic." For example, the existence of any source that directly reveals θ_1 (that is, $X = c\theta_1 + \epsilon$) is non-generic by the above definition, but plausible in practice. Sets of signals that partition into different groups (as described in Section 2.1) are also economically interesting but non-generic.[11] Our earlier Theorem 2 shows that inefficiency is a likely outcome in these cases.

The intuition for the above results, and in particular the role of "low-dimensional" minimal spanning sets, is roughly as follows. If every minimal spanning set is of full rank, then as agents accumulate observations from any minimal spanning set, they learn not only about θ_1 but also about all other states. The aggregated information in the community must then eventually swamp the prior, so that agents' asymptotic evaluation of the value of different signals cannot be prior-dependent. In fact, this asymptotic evaluation returns the optimal comparisons in Section 4.

The argument above is no longer valid when there is a lower-dimensional set of signals that is minimally spanning. Intuitively, observation of $k < K$ signals can be self-reinforcing, since at most k unknown states are revealed from these sources. Thus, any uncertainty in the prior about the other $K-k$ states can persist, despite infinitely many observations of the k signals. Suppose that the remaining sources depend on these $K-k$ "poorly-understood" states. Agents may never acquire information from these sources, and thus never come to learn about these states.

Returning to our example in the introduction, in which we considered a community's acquisition of news about a political leader, recall that there were two possible

[11] Note that the set of coefficient matrices that satisfy Assumption 5 is "generic" in the following stronger sense: fixing the *directions* of coefficient vectors (as in Corollary 2), and suppose that the *precisions* are drawn at random, then different minimal spanning sets achieve different Asd values. In contrast, whether every minimal spanning set has size K is a condition on the directions themselves, so this stronger notion of genericity does not apply.

outcomes. One possible long run outcome is that the community acquires information about the "most revealing" topics—for example, whether the political leader broke laws limiting executive power. Alternatively, the community may become stuck reading about "decoy" topics that are easily understood, but which allow for inefficiently slow learning—for example, information about his personal indiscretions.

Our results suggest that the key property that separates these two long-run outcomes is whether the decoy topics are *revealing*—so that given sufficiently many articles about this topic, the community will learn if the leader is unprincipled—and moreover *self-contained*—so that in the intermediate term, they provide no information that would help the reader to better understand other relevant topics. In contrast, optimal long-run learning would occur if for example, coverage of personal indiscretions led also to better understanding of the limits of executive power, in which case readers would eventually divert attention to reporting about breaches of these limits.

Although not a focus of this paper, the described mechanism above suggests a distortion in information demand, and new considerations for the welfare analysis of news production. Insofar as the media has incentives to provide information that is of greatest immediate interest, provision of information to satisfy immediate demand may not be socially optimal. In contrast, limitation of news sources to investigative pieces—whose "intermediate steps," if released, would not be of public interest, but which are illuminating about the payoff-relevant unknown at the end—may allow for efficient long-run learning.

We sketch below the proof for Theorem 3, relegating the complete proof to the appendix.

5.1 Proof Sketch for Theorem 3

Instead of working directly with the posterior variance function $f(q_1, \ldots, q_N)$, we work with the function

$$f^*(\lambda_1, \ldots, \lambda_N) = \lim_{t \to \infty} t \cdot f(\lambda_1 t, \ldots, \lambda_N t)$$

which is defined on frequency vectors. Lemma 8 relates this function to the posterior variance function, stating that as each q_i grows large,

$$f(q_1, \ldots, q_N) \sim \frac{1}{t} \cdot f^* \left(\frac{q_1}{t}, \ldots, \frac{q_N}{t} \right)$$

where $t = \sum_{i=1}^{N} q_i$ is the total number of observations. Thus, division vectors $(q_1, \ldots, q_N)$ which minimize f^* will asymptotically also minimize the posterior variance function f.

Several properties of the function f^* are relevant. First, the function f^* is convex in λ. Additionally, under the assumption of a "best" minimal spanning set $\mathcal{S}^*$, the function f^* is uniquely minimized at the vector λ^* which assigns to each signal in $\mathcal{S}^*$ its asymptotic frequency, and puts 0 everywhere else (Lemma 6). The question of whether optimal long-run acquisitions are achieved is equivalent to the question of whether signal acquisition frequencies converge to λ^*.

Society's acquisitions follow "pseudo"-gradient descent, where the vector $\lambda(t) = m(t)/t$ evolves according to

$$\lambda(t+1) = \frac{t}{t+1}\lambda(t) + \frac{1}{t+1}e_i.$$

The vector e_i is the coordinate vector that yields the greatest (immediate) reduction in f (and roughly the greatest reduction in f^*). Unlike standard gradient descent, the descent here can occur only along a finite set of feasible directions. This limitation is without loss if f^* is continuously differentiable, which implies that the partial derivative in any direction is a convex combination of partial derivatives along basis vectors.

However, the function f^* can fail to be continuously differentiable at vectors with fewer than K nonzero coordinates. In particular, even when the derivative in the direction of $\lambda^* - \lambda$ is strictly positive, the directional derivative can be 0 along every coordinate vector. This results in agents becoming "stuck" at a sub-optimal point under the pseudo-gradient descent, as reflected in Theorem 2.

We show, however, that f^* is continuously differentiable at all vectors that have at least K nonzero coordinates, and moreover that agents will always eventually observe *some* minimal spanning set. Thus, if every minimal spanning set has size K, descent is well-behaved and ends at the global minimum f^*. This proves Corollary 4, and Theorem 3 follows from a similar argument.

6 Special Correlation Structures

Below, we apply the above results to characterize long-run acquisitions in special informational environments. We first consider an *island model*.

Definition 2. *Information sources with coefficient matrix C constitute an* island model *if the different minimal spanning sets $\mathcal{S}_1, \ldots, \mathcal{S}_M$ satisfy $M > 1$ and moreover, $\overline{\mathcal{S}_m} = \mathcal{S}_m$ for every m.*

An island model was introduced previously in Example 1 in Section 2.1, in which there were two disjoint minimal spanning sets (groups) with two signals each. The different groups may also be unbalanced in the following way:

Example 5. The unknown states are θ_1, b_1, b_2, b_3, where only θ_1 is payoff-relevant. The sources are

$$\begin{aligned} X_1 &= \theta_1 + b_1 + \epsilon_1 \\ X_2 &= b_1 + b_2 + \epsilon_1 \\ X_3 &= b_2 + \epsilon_3 \\ X_4 &= \theta_1 + b_3 + \epsilon_4 \\ X_5 &= b_3 + \epsilon_5 \end{aligned}$$

Then, $\{X_1, X_2, X_3\}$ and $\{X_4, X_5\}$ constitute the only minimal spanning sets.

While this example and Example 1 both have the property that minimal spanning sets are disjoint (and partition $[N]$), this need not be the case, as shown in the following example:

Example 6. The unknown states are θ_1, b_1, b_2, b_3, where only θ_1 is payoff-relevant. The sources are

$$\begin{aligned} X_1 &= \theta_1 + b_1 + \epsilon_1 \\ X_2 &= b_1 + b_2 + \epsilon_1 \\ X_3 &= b_2 + \epsilon_3 \\ X_4 &= b_1 + b_3 + \epsilon_4 \\ X_5 &= b_3 + \epsilon_5 \end{aligned}$$

Note that only X_4 differs from the previous example. Here, the only minimal spanning sets are $\{X_1, X_2, X_3\}$ and $\{X_1, X_4, X_5\}$, which span different subspaces.

A direct application of Theorem 2 yields:

Corollary 5. *For every island model, there is an open set of priors given which the agents eventually (sub-optimally) observes signals in $\mathcal{S} \neq \mathcal{S}^*$.*

Proof. By assumption, $\overline{\mathcal{S}} = \mathcal{S}$ for any minimally spanning set $\mathcal{S}$. Thus for any other minimally spanning set $\mathcal{S}'$, it holds that $\mathcal{S}' \subsetneq \mathcal{S} = \overline{\mathcal{S}}$. This shows $\mathcal{S}$ is the only minimal spanning set in its subspace, and it must be efficient in that subspace. Now choose any $\mathcal{S} \neq \mathcal{S}^*$. Theorem 2 yields the desired conclusion. □

Thus, inefficiency is possible in every island model.

Another special case is a *symmetric* model, in which there are M groups of K signals that are rotationally symmetric around the vector e_1.

Definition 3. *Information sources with coefficient matrix C constitute a* symmetric model *if the rows $[N]$ can be partitioned into M groups of size K, where each group can be ordered $c_1, \ldots, c_K$ such that*

$$c_k = R^{k-1} c_1, \quad \forall k$$

with R a rotation matrix around e_1 satisfying $R^K = I$ and $R^k \neq I$ for every $k \in \{1, \ldots, K-1\}$.[12]

An example of a symmetric model is given below:

Example 7. The unknown states are θ, b, where only θ is payoff-relevant. The sources are

$$\begin{aligned} X_1 &= \theta_1 + b + \epsilon_1 \\ X_2 &= 2\theta_1 + b + \epsilon_2 \\ X_3 &= \theta_1 - b + \epsilon_3 \\ X_4 &= 2\theta_1 - b + \epsilon_4 \end{aligned}$$

Then, the signals $\{X_3, X_4\}$ correspond to a rotation of $\{X_1, X_2\}$ around e_1 using the rotation matrix $R = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

The following is a corollary of Theorem 2:[13]

[12] A rotation matrix around e_1 is any matrix of the form

$$R = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & O_{N-1} \end{bmatrix}$$

with O_{N-1} an arbitrary orthonormal matrix.

[13] One can show that in symmetric models, the *optimal* observation set must be itself symmetric, and Unique Minimizer is satisfied whenever different symmetric minimal spanning set yield different speeds of learning.

Corollary 6. *Fix a symmetric model. If there is a final observation set $\mathcal{S} \neq \mathcal{S}^*$ consisting of signals from distinct groups, then at least $K+1$ different sets could be eventually observed depending on the prior.*

Proof. The rotated versions of $\mathcal{S}$ could all be eventually observed. They are pairwise distinct by the assumption that $\mathcal{S}$ contains signals from distinct groups, and they are also distinct from $\mathcal{S}^*$ since $\mathcal{S}$ necessarily has less than K signals. □

Thus, there can be a sharp discontinuity in the possible long-run outcomes in symmetric models. Either all communities converge to the optimal set $\mathcal{S}^*$, or communities with different priors end up in a range of final observation sets.[14]

7 Information Interventions

Section 5 demonstrated the possibility for sequential information acquisition to result in inefficient learning. We ask now whether it is possible for a benevolent outside party to help society achieve efficient learning by providing a one-time injection of free information. Naturally, this question applies only when agents (on their own) could eventually achieve a sub-optimal speed of learning. The conditions under which this occurs are given in Theorem 2.

Formally, suppose a policy-maker chooses M signals

$$\langle p_j, \theta \rangle + \mathcal{N}(0,1)$$

where each $\|p_j\|_2 \leq \gamma$, so that signal precisions are bounded by γ^2. At time 0, this information is made public. All subsequent agents update their prior beliefs based on this free information, and also on the history of signal acquisitions thus far. The goal of the policy-maker is to maximize the community's asymptotic speed of learning. Below, we use *efficient learning* to mean the case in which the asymptotic speed of learning achieves the optimum—that is, the final observation set is $\mathcal{S}^*$ and long-run frequencies are λ^*.

Is there a sufficient number of (kinds of) signals, such that efficient learning can be guaranteed? We answer in the affirmative below: $K-1$ precise signals are sufficient to produce efficient learning:

[14] When K is prime, the conclusion of Corollary 6 holds unconditionally. That is, the number of possible final observation sets is either 1, or at least K.

Proposition 2. *For any prior, there exists γ and $K-1$ signals with $\|p_j\|_2 \leq \gamma$ such that with these free signals at $t=0$, society achieves efficient learning.*

Intuitively, as long as the free signals make agents' beliefs about states $\theta_2, \ldots, \theta_K$ sufficiently precise, they can preclude the situation in which agents get stuck in a sub-optimal set as in Example 4. Notice that optimal information intervention does not need to teach directly about θ_1 (the parameter of interest), which the agents will learn on their own. Rather, the planner should provide auxiliary information that helps agents to better interpret the sources.

8 Related Literature

In addition to the references mentioned in the introduction, our results build on prior work regarding speed of learning (Vives, 1992; Golub and Jackson, 2012; Harel et al., 2017; Hann-Caruthers, Martynov and Tamuz, 2017), and is related also to the experimental design literature in statistics (see Chernoff (1972) for a survey). Specifically, our results in Section 4 are related to **c**-optimality, in which t experiments are chosen to minimize the posterior variance of a linear combination of the unknown states (in our case, simply the posterior variance of the first unknown state). Theorem 1 can be seen as an integer design version of the problem considered in Chaloner (1984). Chaloner (1984) showed that a c-optimal Bayesian continuous design exists on at most K points, but does not provide a construction of this design. Extending this, we supply a characterization of the optimal design itself; this improves on the prior result by showing uniqueness of the optimal design, and demonstrating that for certain correlational structures, the Bayesian continuous design exists on strictly fewer than K points.

9 Extensions

Non-Persistent i.i.d. States. So far, we have considered persistent states $\theta_1, \ldots, \theta_K$. All of our results extend if new states $\theta_1^t, \ldots, \theta_K^t$ are independently drawn each period according to $\theta_k^t = \theta_k + \gamma_k^t$, and the signals are $X_i^t = \sum_{i=1}^K c_{ik}\theta_k + \epsilon_i^t$ as before. The noise terms γ_k^t and ϵ_i^t are independent from one another. We assume that agent t has payoff function $u_t(a, \theta_1^t)$, which depends on the payoff-relevant state at that time. To see that our results extend, simply notice that the agent's posterior

variance about θ_1^t is the sum of his posterior variance about θ_1 and the variance of γ_1^t. Because the latter cannot be controlled by the agent, his optimal information acquisition strategy is unchanged.

Multiple Payoff-Relevant States. So far we have considered a single payoff relevant state θ_1. In Liang, Mu and Syrgkanis (2017), we pointed out that the reduction argument used in Section 3, which allows for general decision problems, relies on unidimensional payoff-relevant uncertainty. Nevertheless, in Liang, Mu and Syrgkanis (2017) we extended the main results to multiple states for the specific problem of prediction of the unknown states. When there are more sources than states, even the latter extension is quite challenging. We offer limited comments on this case in Appendix G, primarily characterizing bounds on the speed of asymptotic learning.

10 Conclusion

We study a model of sequential learning, where agents choose what kind of information to acquire from a large set of information sources. The key force of interest is the externality that current informational choices generate on future agents.

Our main results characterize two starkly different possibilities and the conditions under which either obtains: (1) the externality is *beneficial*: past information acquisitions help future agents to discern which sources are most informative, and in the long run, agents converge to acquiring information only from the most informative sources; (2) the externality is *harmful*: past information acquisitions increase the value of "low-quality" sources relative to "high-quality" sources, pushing future agents to acquire information from a set of sources that yields inefficiently slow learning. A simple property of the correlation structure across sources determines when such "learning traps" emerge, and which sources are a part of them.

When a community is stuck observing inefficient sources, what kind of information interventions might push the community towards efficient learning? One possibility is to limit the number of sources, and especially to remove "decoy" sources that are low-quality but self-reinforcing. Another possibility is to provide agents with free information. We show that a policy-maker can guarantee efficient long-run learning if he provides a sufficient number of sufficiently precise signals. The optimal information intervention does not inform directly about the payoff-relevant state, but rather provides auxiliary information that helps agents to interpret the best sources (so that

these are subsequently observed). This intervention may require educating agents along *many* different dimensions: we conjecture that provision of a single kind of information (no matter how precise) can be ineffective in a large number of environments. This points to the potential long-run ineffectiveness of information campaigns that are very informative but limited in scope.

Finally, although in this paper we focus on informational demand given a *fixed* set of information sources, one may also consider the reverse question of what kinds of information will be *endogenously* provided by strategic sources. Our results suggest that the answer to this question can be subtle: information sources most frequently viewed in the long run are those that are "least informative in a most informative set." Thus, a source that wants to maximize frequency of viewership has two competing incentives: first, to be viewed at all within the competitive market, it must provide sufficiently useful information; second, conditional on being viewed, it wants to reveal information slowly (so as to increase the number of observations). We leave characterization of the supply of information in an "informationally overabundant" environment for future work.

References

Ali, Nageeb. 2017. "Herding with Costly Information." Working Paper.

Banerjee, Abhijit. 1992. "A Simple Model of Herd Behavior." *Quaterly Journal of Economics*, 107(3): 797–817.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Information Cascades." *Journal of Political Economy*, 100(5): 992–1026.

Burguet, Roberto, and Xavier Vives. 2000. "Social Learning and Costly Information." *Economic Theory*.

Chaloner, Kathryn. 1984. "Optimal Bayesian Experimental Design for Linear Models." *The Annals of Statistics*, 12(1): 283–300.

Chernoff, Herman. 1972. *Sequential Analysis and Optimal Design*. Society for Industrial and Applied Mathematics.

Che, Yeon-Koo, and Konrad Mierendorff. 2017. "Optimal Sequential Decision with Limited Attention." Working Paper.

Fudenberg, Drew, Philip Strack, and Tomasz Strzalecki. 2017. "Stochastic Choice and Optimal Sequential Sampling." Working Paper.

Golub, Benjamin, and Matthew Jackson. 2012. "How Homophily Affects the Speed of Learning and Best-Response Dynamics." *The Quarterly Journal of Economics*.

Hann-Caruthers, Wade, Vadim Martynov, and Omer Tamuz. 2017. "The Speed of Sequential Asymptotic Learning." Working Paper.

Harel, Matan, Elchanan Mossel, Philipp Strack, and Omer Tamuz. 2017. "The Speed of Social Learning." Working Paper.

Liang, Annie, Xiaosheng Mu, and Vasilis Syrgkanis. 2017. "Dynamic Information Acquisition from Multiple Sources." Working Paper.

Mayskaya, Tatiana. 2017. "Dynamic Choice of Information Sources." Working Paper.

Mueller-Frank, Manuel, and Mallesh Pai. 2016. "Social Learning with Costly Search." *American Economic Journal: Microeconomics*.

Sethi, Rajiv, and Muhamet Yildiz. 2016. "Communication with Unknown Perspectives." *Econometrica*, 84(6): 2029–2069.

Sethi, Rajiv, and Muhamet Yildiz. 2017. "Culture and Communication." Working Paper.

Smith, Lones, and Peter Sorenson. 2000. "Pathological Outcomes of Observational Learning." *Econometrica*.

Vives, Xavier. 1992. "How Fast do Rational Agents Learn?" *Review of Economic Studies*.

A Examples Failing Assumption 4

A.1 Details for Example 3

To solve this integer optimization problem, let r be the integer that minimizes $|r-\beta|$ (the distance to β) and let s be the integer that minimizes $|s-\gamma|$. Further, let $\langle\beta\rangle$ and $\langle\gamma\rangle$ be the value of these absolute differences. We assume $2\beta, 2\gamma$ are not integers, so that $0 < \langle\beta\rangle, \langle\gamma\rangle < \frac{1}{2}$. We also assume $\langle\beta\rangle \neq \langle\gamma\rangle$, and by symmetry focus on the case of $\langle\beta\rangle < \langle\gamma\rangle$.

With these assumptions, it is clear that when q_1, q_2 are integers, the minimum value of $|q_1 - q_2 - \beta|$ is $\langle\beta\rangle$, achieved if and only if $q_1 = q_2 + r$. Similarly the minimum value of $|q_3 - q_4 - \gamma|$ is $\langle\gamma\rangle$, achieved when $q_3 = q_4 + s$. Now if the total number of observations t has the *same parity* as $r+s$, it is possible to choose q_1, q_2, q_3, q_4 such that their sum is t and $q_1 = q_2 + r$, $q_3 = q_4 + s$—any pair q_2, q_4 with sum $\frac{t-r-s}{2}$ leads to such a solution. Given these constraints, then, the optimum is to choose q_2, q_4 to minimize $\frac{\langle\beta\rangle^2}{2q_2+r+\beta} + \frac{\langle\gamma\rangle^2}{2q_4+s+\gamma}$. The optimal q_2 and q_4 satisfy $q_2/q_4 \approx \langle\beta\rangle/\langle\gamma\rangle$, which together with $q_2 + q_4 = \frac{t-r-s}{2}$ implies

$$q_1, q_2 \approx \frac{\langle\beta\rangle}{2\langle\beta\rangle + 2\langle\gamma\rangle} \cdot t; \quad q_3, q_4 \approx \frac{\langle\gamma\rangle}{2\langle\beta\rangle + 2\langle\gamma\rangle} \cdot t.$$

On the other hand, suppose t has the *opposite parity* to $r+s$. In this case $q_1 = q_2 + r$ and $q_3 = q_4 + s$ cannot both hold, thus $|q_1 - q_2 - \beta|$ and $|q_3 - q_4 - \gamma|$ cannot both take their minimum values $\langle\beta\rangle$ and $\langle\gamma\rangle$. It turns out that the best one can do is choose $q_1 = q_2 + r$ and $q_3 = q_4 + s \pm 1$ so that $|q_1 - q_2 - \beta| = \langle\beta\rangle$ and $|q_3 - q_4 - \gamma| = 1 - \langle\gamma\rangle$. Then, the optimal choice of q_2, q_4 with sum $\frac{t-r-s\mp1}{2}$ to minimize $\frac{\langle\beta\rangle^2}{2q_2+r+\beta} + \frac{(1-\langle\gamma\rangle)^2}{2q_4+s+\gamma\pm1}$. This yields

$$q_1, q_2 \approx \frac{\langle\beta\rangle}{2\langle\beta\rangle + 2 - 2\langle\gamma\rangle} \cdot t; \quad q_3, q_4 \approx \frac{1-\langle\gamma\rangle}{2\langle\beta\rangle + 2 - 2\langle\gamma\rangle} \cdot t.$$

Hence, in this example, all four signals are observed with positive frequencies in the long run according to the optimal criterion.

A.2 A Second Example: Qualitative Conclusion of Theorem 1 Holds

We give another example in which Assumption 4 (Unique Minimizer) is violated. However, the qualitative conclusion of Theorem 1 still holds. Namely, the t-optimal strategy eventually observes no more than K signals.

Consider two states θ_1, θ_2 independently drawn with variance $\frac{1}{a}$ and $\frac{1}{b}$ respectively. There are three signals $\theta_1 + \epsilon_1, \theta_2 + \epsilon_2$ and $\frac{\theta_1+\theta_2}{2} + \epsilon_3$, where each noise term is standard

normal. We assume that the payoff relevant state is $\frac{\theta_1+\theta_2}{2}$.[15] Observe that the first two signals are sufficient to identify the payoff-relevant state, and $Asd(\{1,2\}) = 1$. Meanwhile, the third signal itself is also spanning, with $Asd(\{3\})$ also equal to 1. Thus Assumption 4 fails.

We claim that for generic values of a and b, t-optimality at large t requires society to focus on the first two signals. Intuitively, this is because one observation of $\theta_1 + \epsilon_1$ and one observation of $\theta_2 + \epsilon_2$ contain at least as much information as their sum $\theta_1 + \theta_2 + \epsilon_1 + \epsilon_2$, which is the same as two observations of $\frac{\theta_1+\theta_2}{2} + \epsilon_3$. Thus, devoting any level of attention to the third signal is weakly worse than splitting that attention evenly between the first two signals. Furthermore, the combination of the first two signals also informs about the difference $\theta_1 - \theta_2$, which is correlated with the payoff-relevant state $\frac{\theta_1+\theta_2}{2}$ whenever the prior variances about θ_1 and θ_2 differ. Thus, society optimally "ignores" the third signal if its (prior and posterior) beliefs about θ_1 and θ_2 are *asymmetric*. As we show below, this occurs precisely when $a - b$ is not an integer.

To formalize the above intuition, we observe that given q_1 observations of signal 1 and q_2 observations of signal 2, society's posterior variance about $\frac{\theta_1+\theta_2}{2}$ is $\left(\frac{1}{q_1+a} + \frac{1}{q_2+b}\right)/4$. Thus, with q_3 additional observations of $\frac{\theta_1+\theta_2}{2} + \epsilon_3$, society's posterior variance becomes

$$f(q_1, q_2, q_3) = 1 \Bigg/ \left(\frac{4}{\frac{1}{q_1+a} + \frac{1}{q_2+b}} + q_3 \right).$$

The optimal problem at time t reduces to the following *maximization*:

$$\max_{q_1,q_2,q_3 \in \mathbb{Z}^+, q_1+q_2+q_3=t} \frac{4}{\frac{1}{q_1+a} + \frac{1}{q_2+b}} + q_3.$$

The maximand can be rewritten as

$$\frac{4}{\frac{1}{q_1+a} + \frac{1}{q_2+b}} + q_3 = q_1 + a + q_2 + b + q_3 - \frac{(q_1 + a - q_2 - b)^2}{q_1 + a + q_2 + b}.$$

Note that $q_1 + a + q_2 + b + q_3 = t + a + b$ is fixed, so society chooses q_1, q_2 to *minimize* the ratio $\frac{(q_1+a-q_2-b)^2}{q_1+a+q_2+b}$.

Suppose $a - b$ is not an integer, let $\langle a - b \rangle$ denote its distance to the nearest integer. Then, as q_1, q_2 are restricted to integers, the difference $|q_1 + a - q_2 - b|$ takes minimum value $\langle a - b \rangle > 0$. It follows that $\frac{(q_1+a-q_2-b)^2}{q_1+a+q_2+b}$ is uniquely minimized by choosing q_1, q_2 such that $|q_1 + a - q_2 - b| = \langle a - b \rangle$ and $q_1 + q_2$ is as large as possible. Hence, both q_1 and q_2 are close to $\frac{t}{2}$, and our earlier claim is verified.

[15] It is straightforward to linearly transform this environment into one that fits our model exactly, but we will not do that.

B Preliminaries

First, we review and extend a basic result from Liang, Mu and Syrgkanis (2017). Specifically, we show that the posterior variance weakly decreases over time, and the marginal value of any signal decreases in its signal count.

Lemma 2. *Given prior covariance matrix V^0 and q_i observations of each signal i, society's posterior variance about θ_1 is given by*

$$f(q_1, \ldots, q_N) = \left[((V^0)^{-1} + C'QC)^{-1}\right]_{11} \tag{7}$$

where $Q = \operatorname{diag}(q_1, \ldots, q_N)$. The function f is decreasing and convex in each q_i whenever these arguments take non-negative real values.

Proof. Note that $(V^0)^{-1}$ is the prior precision matrix, and $C'QC = \sum_{i=1}^{N} q_i \cdot [c_i c_i']$ is the total precision from the signals. Thus (7) simply represents the fact that for Gaussian prior and signals, the posterior precision matrix is the sum of prior and signal precision matrices. To prove the monotonicity of f, consider the partial order $\succeq$ on positive semi-definite matrices where $A \succeq B$ if and only if $A - B$ is positive semi-definite. As q_i increases, the matrix Q and $C'QC$ increase in this order. Thus the posterior covariance matrix $((V^0)^{-1} + C'QC)^{-1}$ decreases in this order, which implies that the posterior variance about θ_1 decreases. Intuitively, more information always improves the decision-maker's estimates.

To prove f is convex, it suffices to prove f is *midpoint-convex* since the function is clearly continuous. Take $q_1, \ldots, q_N$, $r_1, \ldots, r_N \in \mathbb{R}_+$ and let $s_i = \frac{q_i + r_i}{2}$. Define the corresponding diagonal matrices to be Q, R, S. Observe that $Q + R = 2S$. Thus by the AM-HM inequality for positive-definite matrices, we have in matrix order

$$((V^0)^{-1} + C'QC)^{-1} + ((V^0)^{-1} + C'RC)^{-1} \succeq 2((V^0)^{-1} + C'SC)^{-1}.$$

Using (7), we conclude

$$f(q_1, \ldots, q_N) + f(r_1, \ldots, r_N) \geq 2f(s_1, \ldots, s_N).$$

This proves the convexity of f. □

Second, we provide a definition of $[X^{-1}]_{11}$ for positive *semi-definite* matrices X. When X is positive definite, its eigenvalues are strictly positive, and its inverse matrix is defined as usual. In general, we can apply the spectrum theorem to write

$$X = UDU'$$

with U being a $K \times K$ orthogonal matrix whose columns are eigenvectors of X, and D being a $K \times K$ diagonal matrix consisting of non-negative eigenvalues. Even if some of these eigenvalues are zero, we can think of X^{-1} as

$$X^{-1} = (UDU')^{-1} = UD^{-1}U' = \sum_{j=1}^{K} \frac{1}{d_j} \cdot [u_j u_j']$$

with u_j being the j-th column vector of U. We thus define

$$[X^{-1}]_{11} = \sum_{j=1}^{K} \frac{(\langle u_j, e_1 \rangle)^2}{d_j}, \tag{8}$$

with the convention that $\frac{0}{0} = 0$. Note that by this definition,

$$[X^{-1}]_{11} = \lim_{\epsilon \to 0_+} \left(\sum_{j=1}^{K} \frac{(\langle u_j, e_1 \rangle)^2}{d_j + \epsilon} \right) = [(X + \epsilon I_K)^{-1}]_{11}$$

since the matrix $X + \epsilon I_K$ has the same set of eigenvectors as X, with eigenvalues increased by ϵ. Hence our definition of $[X^{-1}]_{11}$ is a continuous extension of the usual definition to positive semi-definite matrices. Note that we allow $[X^{-1}]_{11}$ to be infinite.

C Proof of Theorem 1

C.1 Characterization of Asymptotic Variance

We first approximate the posterior variance as a function of the frequencies with which each signal is observed. Specifically,

Lemma 3. *For any $\lambda_1, \ldots, \lambda_N \geq 0$, let $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$. Then*

$$\begin{aligned} f^*(\lambda_1, \ldots, \lambda_N) :=& \lim_{t \to \infty} t \cdot f(\lambda_1 t, \ldots, \lambda_N t) \\ =& [(C' \Lambda C)^{-1}]_{11} \end{aligned} \tag{9}$$

Note that the matrix $C'\Lambda C$ is positive semi-definite. So the value of $[(C'\Lambda C)^{-1}]_{11}$ is well defined, see (8).

Proof. Recall that $f(q_1, \ldots, q_N) = \left[((V^0)^{-1} + C'QC)^{-1}\right]_{11}$ with $Q = \operatorname{diag}(q_1, \ldots, q_N)$. Thus

$$t f(\lambda_1 t, \ldots, \lambda_N t) = \left[\left(\frac{1}{t} (V^0)^{-1} + C' \Lambda C \right)^{-1} \right]_{11}.$$

Hence by the continuity of $[X^{-1}]_{11}$ in the matrix X, we obtain the lemma. □

We note that $C'\Lambda C$ is the Fisher Information Matrix when the signals are observed according to frequencies λ. Thus the above lemma can also be seen as an application of the Bayesian Central Limit Theorem.

C.2 Reduction to the Study of f^*

The development of the function f^* is useful for the following reason:

Lemma 4. *Suppose $\hat{\lambda}$ uniquely minimizes $f^*(\lambda)$ subject to $\lambda \in \Delta^{N-1}$ (the $N-1$-dimensional simplex), then the t-optimal divisions satisfy $n_i(t) \sim \hat{\lambda}_i \cdot t$ for each i.*

Proof. Fix any increasing sequence of times $t_1, t_2, \ldots$. It suffices to show that whenever the limit $\lambda_i := \lim_{m\to\infty} \frac{n_i(t_m)}{t_m}$ exists for each i, this limit λ must be $\hat{\lambda}$. Suppose not, then by assumption $f^*(\lambda) > f^*(\hat{\lambda})$. For $\epsilon > 0$, define another vector $\tilde{\lambda} \in \mathbb{R}^N_+$ with $\tilde{\lambda}_i = \lambda_i + \epsilon, \forall i$. By the continuity of f^*, it holds that $f^*(\tilde{\lambda}) > f^*(\hat{\lambda})$ for sufficiently small ϵ.

Since $\lambda_i = \lim_{m\to\infty} \frac{n_i(t_m)}{t_m}$, there exists M sufficiently large such that $n_i(t_m) \leq \tilde{\lambda}_i \cdot t_m$ for each i and $m \geq M$. Hence, for $m \geq M$,

$$t_m \cdot f(n_1(t_m), \ldots, n_N(t_m)) \geq t_m \cdot f(\tilde{\lambda}_1 \cdot t_m, \ldots, \tilde{\lambda}_N \cdot t_m) \to f^*(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)$$

The first inequality uses the monotonicity of f. On the other hand,

$$t_m \cdot f(\hat{\lambda}_1 \cdot t_m, \ldots, \hat{\lambda}_N \cdot t_m) \to f^*(\hat{\lambda}_1, \ldots, \hat{\lambda}_N).$$

Comparing the above two displays, we see that for sufficiently large m, $f(n_1(t_m), \ldots, n_K(t_m)) > f(\hat{\lambda}_1 \cdot t_m, \ldots, \hat{\lambda}_N \cdot t_m)$. But this contradicts the t-optimality of the division $n(t_m)$, as society could do better by following frequencies $\hat{\lambda}$. The lemma is thus proved. □

C.3 Crucial Lemma about the Structure of Signal Vectors

We pause to demonstrate the following lemma:

Lemma 5. *Suppose $\mathcal{S}^* = \{1, \ldots, K\}$ uniquely minimizes $Asd(\mathcal{S})$ and suppose $[(C^*)^{-1}]_{1j}$ is positive for $1 \leq j \leq K$. Consider any $i > K$ and write $c_i = \sum_{j=1}^K \alpha_j \cdot c_j$. Then $|\sum_{j=1}^K \alpha_j| < 1$.*

Proof. By assumption, we have the vector identity

$$e_1 = \sum_{j=1}^K x_j \cdot c_j \quad \text{with } x_j = [(C^*)^{-1}]_{1j} > 0.$$

Suppose for contradiction that $\sum_{j=1}^{K} \alpha_j \geq 1$ (the opposite case where the sum is ≤ -1 can be similarly treated). In particular, some α_j is positive. Without loss of generality, we assume $\frac{\alpha_1}{x_1}$ is the largest among such ratios. Then $\alpha_1 > 0$ and

$$e_1 = \sum_{j=1}^{K} x_j \cdot c_j = \left(\sum_{j=2}^{K} (x_j - \frac{x_1}{\alpha_1} \cdot \alpha_j) \cdot c_j \right) + \frac{x_1}{\alpha_1} \cdot \left(\sum_{j=1}^{K} \alpha_j \cdot c_j \right)$$

This represents e_1 as a linear combination of the vectors $c_2, \ldots, c_K$ and c_i, with coefficients $x_2 - \frac{x_1}{\alpha_1} \cdot \alpha_2, \ldots, x_K - \frac{x_1}{\alpha_1} \cdot \alpha_K$ and $\frac{x_1}{\alpha_1}$. Observe that these coefficients are non-negative: for each $2 \leq j \leq K$, $x_j - \frac{x_1}{\alpha_1} \cdot \alpha_j$ is clearly positive if $\alpha_j \leq 0$ (since $x_j > 0$). And if $\alpha_j > 0$, then by assumption $\frac{\alpha_j}{x_j} \leq \frac{\alpha_1}{x_1}$ and $x_j - \frac{x_1}{\alpha_1} \cdot \alpha_j$ is again non-negative.

By definition, $Asd(\{2, \ldots, K, i\})$ is the sum of the absolute value of these coefficients. This sum is

$$\sum_{j=2}^{K} (x_j - \frac{x_1}{\alpha_1} \cdot \alpha_j) + \frac{x_1}{\alpha_1} = \sum_{j=1}^{K} x_j + \frac{x_1}{\alpha_1} \cdot (1 - \sum_{j=1}^{K} \alpha_j) \leq \sum_{j=1}^{K} x_j.$$

But then $Asd(\{2, \ldots, K, i\}) \leq Asd(\{1, 2, \ldots, K\})$, leading to a contradiction. Hence the lemma must be true. □

C.4 Proof of Theorem 1 when $|\mathcal{S}^*| = K$

Given Lemma 4, Theorem 1 will follow once we show that λ^* uniquely minimizes $f^*(\lambda)$ over the simplex—recall that λ^* denotes the optimal asymptotic frequencies for the minimal spanning set $\mathcal{S}^*$ that minimizes *Asd*. In this section, we prove λ^* is indeed the unique minimizer whenever this "best" subset $\mathcal{S}^*$ contains exactly K signals. Later on we will prove the same result even when $|\mathcal{S}^*| < K$, but that proof will require additional techniques.

Lemma 6. *Suppose $\mathcal{S}^* = \{1, \ldots, K\}$ is the unique minimizer of $Asd(\mathcal{S})$ over minimal spanning sets. Define $\lambda^* \in \Delta^{N-1}$ by*

$$\lambda_i^* = \frac{|[(C^*)^{-1}]_{1i}|}{\sum_{j=1}^{K} |[(C^*)^{-1}]_{1j}|}, 1 \leq i \leq K$$

with $C^* = C_{[K][K]}$,[16] and $\lambda_i^* = 0, \forall i > K$. Then $f^*(\lambda^*) < f^*(\lambda)$ for any $\lambda \in \Delta^{N-1}, \lambda \neq \lambda^*$.

Proof. First, we will assume that $[(C^*)^{-1}]_{1i}$ is positive for $1 \leq i \leq K$. This is without loss because we can always work with the "negative" of any signal (replace c_i with $-c_i$), which does not affect agents' behavior.

[16] For any subset $\mathcal{I} \subset [N]$ and $\mathcal{J} \subset [K]$, write $C_{\mathcal{I}\mathcal{J}}$ for the sub-matrix of C with row indices in $\mathcal{I}$ and column indices in $\mathcal{J}$. Likewise, let $C_{-\mathcal{I}\mathcal{J}}$ be the sub-matrix of C after deleting rows in $\mathcal{I}$ and columns in $\mathcal{J}$.

Since $f(q_1, \ldots, q_N)$ is convex in its arguments, $f^*(\lambda) = \lim_{t\to\infty} t \cdot f(\lambda_1 t, \ldots, \lambda_N t)$ is also convex in λ. To show $f^*(\lambda^*) < f^*(\lambda)$, we only need to show $f^*(\lambda^*) < f^*((1-\epsilon)\lambda^* + \epsilon\lambda)$ for some $\epsilon > 0$. In other words, it suffices to show $f^*(\lambda^*) < f^*(\lambda)$ for λ in an ϵ-neighborhood of λ^*. By assumption, $\mathcal{S}^*$ is minimally-spanning and so its signals are linearly independent. Thus its signals must span all of the K states. From this it follows that the $K \times K$ matrix $C'\Lambda^* C$ is positive definite, and by (9) the function f^* is continuously differentiable near λ^* (not just continuous, see Remark 1 below).

We claim that the partial derivatives of f^* satisfy the following inequality:

$$\partial_K f^*(\lambda^*) < \partial_i f^*(\lambda^*) \leq 0, \forall i > K. \tag{**}$$

Once this is proved, we will have, for λ close to λ^*,

$$f^*(\lambda_1, \ldots, \lambda_K, \lambda_{K+1}, \ldots, \lambda_N) \geq f^*(\lambda_1, \ldots, \lambda_{K-1}, \lambda_K + \lambda_{K+1} + \cdots + \lambda_N, 0, \ldots, 0) \geq f^*(\lambda^*). \tag{10}$$

The first inequality is based on (**) and continuous differentiability of f^*, while the second inequality is because λ^* uniquely minimizes f^* if society only observes the first K signals. Moreover, when $\lambda \neq \lambda^*$, one of these inequalities is strict so that $f^*(\lambda) > f^*(\lambda^*)$ strictly.

To prove (**), we recall that

$$f^*(\lambda_1, \ldots, \lambda_N) = e_1'(C'\Lambda C)^{-1} e_1.$$

Since $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$, its derivative is $\partial_i \Lambda = \Delta_{ii}$, which is an $N \times N$ matrix whose (i,i)-th entry is 1 and all other entries are zero. Using properties of matrix derivatives, we obtain

$$\partial_i f^*(\lambda) = -e_1'(C'\Lambda C)^{-1} C'\Delta_{ii} C (C'\Lambda C)^{-1} e_1.$$

As the i-th row vector of C is c_i', $C'\Delta_{ii}C$ is the $K \times K$ matrix $c_i c_i'$. The above simplifies to

$$\partial_i f^*(\lambda) = -[e_1'(C'\Lambda C)^{-1} c_i]^2.$$

At $\lambda = \lambda^*$, the matrix $C'\Lambda C$ further simplifies to $(C^*)' \cdot \text{diag}(\lambda_1^*, \ldots, \lambda_K^*) \cdot (C^*)$, which is a product of $K \times K$ invertible matrices. We thus deduce that

$$\partial_i f^*(\lambda^*) = -\left[e_1' \cdot (C^*)^{-1} \cdot \text{diag}\left(\frac{1}{\lambda_1^*}, \ldots, \frac{1}{\lambda_K^*}\right) \cdot ((C^*)')^{-1} \cdot c_i\right]^2.$$

It is crucial for our analysis that the term in the brackets is a linear function of c_i. To ease notation, we write $v' = e_1' \cdot (C^*)^{-1} \cdot \text{diag}\left(\frac{1}{\lambda_1^*}, \ldots, \frac{1}{\lambda_K^*}\right) \cdot ((C^*)')^{-1}$ and $\gamma_i = \langle v, c_i \rangle$. Then

$$\partial_i f = -\gamma_i^2, \ 1 \leq i \leq N. \tag{11}$$

For $1 \leq i \leq K$, $((C^*)')^{-1} \cdot c_i$ is just e_i. Thus, using the assumption $[(C^*)^{-1}]_{1j} > 0, \forall j$, we have

$$\gamma_i = e_1' \cdot (C^*)^{-1} \cdot \text{diag}\left(\frac{1}{\lambda_1^*}, \ldots, \frac{1}{\lambda_K^*}\right) \cdot e_i = \frac{[(C^*)^{-1}]_{1i}}{\lambda_i^*} = \sum_{j=1}^{K} [(C^*)^{-1}]_{1j} = Asd(\mathcal{S}^*),\ 1 \leq i \leq K. \tag{12}$$

On the other hand, choosing any $i > K$, we can uniquely write the vector c_i as a linear combination of $c_1, \ldots, c_K$. By Lemma 5, for any $i > K$ we have

$$\gamma_i = \langle v, c_i \rangle = \sum_{j=1}^{K} \alpha_j \cdot \langle v, c_j \rangle = \sum_{j=1}^{K} \alpha_j \cdot \gamma_j = Asd(\mathcal{S}^*) \cdot \sum_{j=1}^{K} \alpha_j. \tag{13}$$

The last equality uses (12). Since $|\sum_{j=1}^{K} \alpha_j| < 1$, the absolute value of γ_i for any $i > K$ is strictly smaller than the absolute value of γ_K. This together with (11) proves the desired inequality (**), and the lemma follows. □

Remark 1. The essence of this proof is the following non-obvious fact: the subset $\{1, \ldots, K\}$ uniquely minimizes Asd among all subsets of size K if and only if

$$Asd(\{1, \ldots, K\}) < Asd(\{1, \ldots, K\} \cup \{i\} \backslash \{j\}),\ \forall 1 \leq j \leq K < i \leq N.$$

That is, if a set of K signals does not minimize Asd, then we can improve the speed of learning simply by adding *one* signal to replace *one* existing signal. This property enables us to reduce the general problem with N signals to the much simpler problem with $K+1$ signals, and we are able to use calculus to resolve the latter problem, see (**).

However, the above fact relies on the original set containing exactly K signals. To see this, consider two states and three signals with coefficient vectors $c_1 = (0.5, 0), c_2 = (1, 1), c_3 = (1, -1)$. If we start with the first signal alone, adding *either* of the latter two signals does not decrease Asd. However, the latter two signals *combined* yield a faster speed of learning, as $Asd(\{2, 3\}) = 1 < 2 = Asd(\{1\})$. On the technical level, this occurs because f^* is not continuously differentiable at $(1, 0, 0)$. Thus, even though the partial derivatives satisfy (**), we cannot deduce that any *directional* derivative similarly satisfies (**). It is for this reason that we need a different proof of Lemma 6 when $|\mathcal{S}^*| < K$, which we present later.

Remark 2. Still assuming that the "best" subset $\mathcal{S}^*$ contains exactly K signals, we now show $n_i(t) = \lambda_i^* \cdot t + O(1), \forall i$, thus improving upon the conclusion of Theorem 1. First, we can apply Lemma 5 to find a positive constant $\eta < 1$ such that for *each* $i > K$, if $c_i = \sum_{j=1}^{K} \alpha_j c_j$ then $|\sum_{j=1}^{K} \alpha_j| \leq 1 - \eta$. By (11), (12) and (13), we have

$$\partial_1 f(\lambda^*) = \cdots = \partial_K f(\lambda^*) = -Asd(B^*)^2; \quad \partial_i f(\lambda^*) \geq -(1-\eta)^2 \cdot Asd(B^*)^2, \forall i > K. \tag{14}$$

For any $\lambda \in \Delta^{N-1}$, the convexity of f^* implies[17]

$$\begin{aligned} f^*(\lambda) &\geq f^*(\lambda^*) + \sum_{i=1}^{N} (\lambda_i - \lambda_i^*) \cdot \partial_i f^*(\lambda^*) \\ &= f^*(\lambda^*) + \sum_{i=1}^{N} (\lambda_i - \lambda_i^*) \cdot (\partial_i f^*(\lambda^*) + Asd(B^*)^2) \\ &\geq f^*(\lambda^*) + (2\eta - \eta^2) \cdot Asd(B^*)^2 \cdot \sum_{i=K+1}^{N} \lambda_i. \end{aligned} \tag{15}$$

The second line uses $\sum_{i=1}^{N} (\lambda_i - \lambda_i^*) = 0$ and the last inequality is due to (14).

Consider any division $(q_1, \ldots, q_N)$ at time t. A straightforward refinement of Lemma 3 gives that whenever $f^*(\lambda)$ is finite, $t \cdot f(\lambda t)$ approaches $f^*(\lambda)$ at the rate of $\frac{1}{t}$. In particular $f(\lambda^* \cdot t) = \frac{1}{t} \cdot f^*(\lambda^*) + O(\frac{1}{t^2})$. For $(q_1, \ldots, q_N)$ to be a t-optimal division, it is necessary that $f(q_1, \ldots, q_N) \leq f(\lambda^* \cdot t)$. Thus

$$f^*\left(\frac{q_1}{t}, \ldots, \frac{q_N}{t}\right) \leq f^*(\lambda^*) + O\left(\frac{1}{t}\right). \tag{16}$$

By (15) and (16), any t-optimal division $n(t)$ must satisfy $n_i(t) = O(1)$ for each signal $i > K$. Conditional on these signal counts, society's optimal choice over signals 1 through K must satisfy $n_i(t) = \lambda_i^* \cdot t + O(1), \forall 1 \leq i \leq K$, as shown in Proposition 1. This is what we desire to prove here.

C.5 A Perturbation Argument

We have shown that whenever $Asd(\mathcal{S})$ is uniquely minimized by a set $\mathcal{S}$ containing K signals,

$$\min_{\lambda \in \Delta^{N-1}} f^*(\lambda) = f^*(\lambda^*) = \min_{\mathcal{S} \subset [N]} Asd(\mathcal{S})^2 = Asd([N])^2$$

We now show this equality holds more generally.

Lemma 7. *For any coefficient matrix C,*

$$\min_{\lambda \in \Delta^{N-1}} f^*(\lambda) = Asd([N])^2. \tag{17}$$

Proof. We assume that θ_1 is identified from the available signals; otherwise $f^*(\lambda)$ and $Asd([N])$ are both infinite and equality holds trivially. Because society can choose to focus

[17] As mentioned in Remark 1, it is crucial that f^* is continuously differentiable at λ^*. The argument here relies on the directional derivative in the direction $\lambda - \lambda^*$ being well-defined and equal to a linear sum of partial derivatives.

on any a minimal spanning set, it is clear that $\min_\lambda f^*(\lambda) \leq Asd([N])^2 = \min_{\mathcal{S}}(Asd(\mathcal{S}))^2$. It remains to prove $f^*(\lambda) \geq Asd([N])^2$ for any fixed $\lambda \in \Delta^{N-1}$. By Lemma 3, we need to show $[(C'\Lambda C)^{-1}]_{11} \geq Asd([N])^2$.

This was already proved for *generic* coefficient matrices C; specifically, those for which $Asd(\mathcal{S})$ is minimized by a set of K signals. But even if C is "non-generic", we can approximate it by a sequence of "generic" matrices C_m.[18] Along this sequence, we have

$$[(C_m'\Lambda C_m)^{-1}]_{11} \geq Asd_m([N])^2$$

where Asd_m is the speed of learning from the N signals given by C_m. As $m \to \infty$, the LHS above approaches $[(C'\Lambda C)^{-1}]_{11}$. Thus the lemma will follow once we show that $\limsup_{m\to\infty} Asd_m([N]) \geq Asd([N])$.

For this we invoke the following characterization

$$Asd([N]) = \min_{\beta\in\mathbb{R}^N} \sum_{i=1}^N |\beta_i| \quad \text{s.t.} \quad e_1 = \sum_{i=1}^N \beta_i \cdot c_i.$$

If $e_1 = \sum_i \beta_i^{(m)} \cdot c_i^{(m)}$ along the sequence, then $e_1 = \sum_i \beta_i \cdot c_i$ for any limit point β of $\beta^{(m)}$. This enables us to conclude $\liminf_{m\to\infty} Asd_m([N]) \geq Asd([N])$, which is more than we needed. □

C.6 Proof of Theorem 1 when $|\mathcal{S}^*| < K$

Here we prove Theorem 1 for the case where the "best" subset $\mathcal{S}^*$ contains less than K signals. To be precise, let $\mathcal{S}^* = \{1, \ldots, k\}$ and define $\lambda^* \in \Delta^{N-1}$ to be the optimal frequencies when only the first k signals are observed. We will show $n_i(t) \sim \lambda_i^* \cdot t, \forall i$. By Lemma 4, we only need to show that λ^* uniquely minimizes $f^*(\lambda)$ over the simplex. Since $f^*(\lambda^*) = Asd(\mathcal{S}^*)^2 = Asd([N])^2$ by definition, we know from Lemma 7 that λ^* does minimize $f^*(\lambda)$.

It remains to show that λ^* is the unique minimizer. Suppose for contradiction that $f^*(\lambda^*) = f^*(\tilde{\lambda})$ for some $\tilde{\lambda} \in \Delta^{N-1}$ distinct from λ^*. For $\eta \in \mathbb{R}$, define $\lambda^\eta = \lambda^* + \eta \cdot (\tilde{\lambda} - \lambda^*)$, so that $\lambda^0 = \lambda^*, \lambda^1 = \tilde{\lambda}$. Observe that when $\eta \in (0, 1)$, λ^η is a convex combination between λ^* and $\tilde{\lambda}$. Thus the convexity of f^* implies

$$f^*(\lambda^\eta) \leq (1-\eta) f^*(\lambda^*) + \eta f^*(\tilde{\lambda}) = f^*(\lambda^*)$$

[18] First, we may add repetitive signals to ensure $N \geq K$. This does not affect the value of $\min f^*(\lambda)$ or $Asd([N])$. Whenever $N \geq K$, it is generically true that every minimal spanning set contains exactly K signals. Moreover, the equality $Asd(\mathcal{S}) = Asd(\tilde{\mathcal{S}})$ for $\mathcal{S} \neq \tilde{\mathcal{S}}$ induces a non-trivial polynomial equation over the entries in C. This means we can always find $C^{(m)}$ close to C such that for the coefficient matrix $C^{(m)}$, different subsets $\mathcal{S}$ (of size K) attain different values of $Asd(\mathcal{S})$.

Since $f^*(\lambda^*)$ is minimal, we must then have $f^*(\lambda^\eta) = f^*(\lambda^*)$ for $\eta \in (0,1)$. But for fixed λ^* and λ, (9) shows that the value of $f^*(\lambda^\eta)$ is a rational function (quotient of two polynomials) of η. Thus this rational function is itself a constant. Consequently, $f^*(\lambda^\eta) = f^*(\lambda^*)$ for all η (not just those in the unit interval) such that $\lambda^\eta \in \Delta^{N-1}$.

Because $\tilde{\lambda} \neq \lambda^*$, there exists some $j \in \{1, \ldots, k\}$ such that $\tilde{\lambda}_j < \lambda^*_j$. Without loss, we assume $\frac{\tilde{\lambda}_1}{\lambda^*_1}$ is the smallest among such ratios. Let $\eta = \frac{\lambda^*_1}{\lambda^*_1 - \tilde{\lambda}_1}$, then the vector λ^η has first-coordinate 0 and all other coordinates non-negative. By our preceding analysis, $f^*(\lambda^\eta) = f^*(\lambda^*)$ for this η. However, since λ^η "ignores" signal 1, Lemma 7 implies that

$$f^*(\lambda^\eta) \geq \min_{\lambda \in \Delta^{N-1},\ \lambda_1 = 0} f^*(\lambda) = Asd([N]\backslash\{1\})^2.$$

By assumption, $\mathcal{S}^* = \{1, \ldots, k\}$ is the *unique* minimal spanning set that minimizes Asd. Thus the RHS above is strictly larger than $Asd(\mathcal{S}^*)^2 = f^*(\lambda^*)$, leading to the contradictory result $f^*(\lambda^\eta) > f^*(\lambda^*)$.

This contradiction shows λ^* must uniquely minimize $f^*(\lambda)$, and the proof of Theorem 1 is complete.

D Proof of Theorem 2

Let signals $1, \ldots, k$ (with $k \leq K$) be a minimally spanning set that is efficient in its subspace. We will demonstrate an open set of prior beliefs given which *all agents* observe these k signals. Since these signals are minimally spanning, they must be linearly independent. Thus we can consider linearly transformed states $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ such that these k signals are simply $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ plus standard Gaussian noise. This linear transformation is invertible, so any prior over the original states is bijectively mapped to a prior over the transformed states. Thus it is without loss to work with the transformed model and look for prior beliefs over the transformed states.

By identifiability, the payoff-relevant state θ_1 becomes a linear combination $w_1\tilde{\theta}_1 + \cdots + w_k\tilde{\theta}_k$. We may without loss assume the weights w_i are all positive; otherwise simply replace $\tilde{\theta}_i$ with $-\tilde{\theta}_i$. Now, by assumption, the first k signals are efficient in its subspace. Thus Lemma 5 implies that any signal $j > k$ that belongs to the subspace of $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ can be written as

$$\sum_{i=1}^{k} \alpha_i \tilde{\theta}_i \ + \ \mathcal{N}(0,1)$$

with $|\sum_{i=1}^{k} \alpha_i| < 1$. On the other hand, if a signal $j > k$ does not belong to this subspace, it must take the form of

$$\sum_{i=1}^{K} \beta_i \tilde{\theta}_i \ + \ \mathcal{N}(0,1)$$

with $\beta_{k+1}, \ldots, \beta_K$ not all equal to zero.

Now consider a prior belief such that $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ are independent from each other. Given prior variances $v_1, \ldots, v_K$, the reduction in the variance of $w_1\tilde{\theta}_1 + \cdots + w_k\tilde{\theta}_k$ by any signal $\sum_{i=1}^k \alpha_i\tilde{\theta}_i + \mathcal{N}(0,1)$ is

$$\frac{(\sum_{i=1}^k \alpha_i w_i v_i)^2}{1 + \sum_{i=1}^k \alpha_i^2 v_i}$$

If $v_1, \ldots, v_k$ are small positive numbers and if the product $w_i v_i$ is approximately constant across $1 \leq i \leq k$, then the above is approximately $(\sum_{i=1}^k \alpha_i)^2 w_1^2 v_1^2$. Since $|\sum_{i=1}^k \alpha_i| < 1$, we deduce that any signal $j > k$ that belongs to the subspace of the first k signals is worse than signal 1 (in the first period), whose variance reduction is $\frac{w_1^2 v_1^2}{v_1+1}$.

Meanwhile, take any signal $j > k$ that does not belong to the subspace. The variance reduction by such a signal $\sum_{i=1}^K \beta_i\tilde{\theta}_i + \mathcal{N}(0,1)$ is

$$\frac{(\sum_{i=1}^k \beta_i w_i v_i)^2}{1 + \sum_{i=1}^K \beta_i^2 v_i}$$

As $\beta_{k+1}, \ldots, \beta_K$ are not all zero, the denominator above can be arbitrarily large if $v_{k+1}, \ldots, v_K$ are chosen to be large. Then again this signal is worse than signal 1 for the first agent, just as we showed in Example 4.

To summarize, we have shown that whenever the prior variances $v_1, \ldots, v_K$ satisfy the following three conditions, the first agent chooses among the first k signals:

1. $v_1, \ldots, v_k$ are close to 0;
2. $w_1 v_1, \ldots, w_k v_k$ have pairwise ratios close to 1;
3. $v_{k+1}, \ldots, v_K$ are large.[19]

To show that the signal choice stays among the first k signals *in every period*, it suffices to check that starting from any prior satisfying the above conditions, the posterior after observing a signal continues to satisfy these conditions. Since variances decrease over time, the first condition is obviously satisfied. By independence, learning about $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ does not affect the variances of the remaining states. So $v_{k+1}, \ldots, v_K$ are unchanged, and the third condition is verified. Finally, the second condition holds for the posterior because the signal i that is chosen has the greatest value of $\frac{w_i^2 v_i^2}{v_i+1}$. This choice ensures that $v_i \propto \frac{1}{w_i}$, as shown also in Liang, Mu and Syrgkanis (2017). Theorem 2 is proved.[20]

[19] Formally, we require that for some $\xi > 0$, it holds that $v_1, \ldots, v_k < \xi$; $\max_{1 \leq i \leq k} w_i v_i \leq (1+\xi) \cdot \min_{1 \leq i \leq k} w_i v_i$; and $v_{k+1}, \ldots, v_K > \frac{1}{\xi}$.

[20] Strictly speaking, the above construction does not provide an *open set* of prior beliefs given which agents always observe the first k signals. This is because we restricted attention to priors

E Proof of Theorem 3

E.1 Preliminaries

Given any prior, let $\mathcal{A} \subset [N]$ be the set of signals that are observed by infinitely many agents. We first show that $\mathcal{A}$ is a spanning set.

Indeed, by definition we can find some period t after which agents only observe signals in $\mathcal{A}$. Also note that the variance reduction of any signal approaches zero as its signal count gets large. Thus, along society's signal path, the variance reduction is close to zero at sufficiently late periods.

If $\mathcal{A}$ is not spanning, society's posterior variance remains bounded away from zero. Thus in the limit where each signal in $\mathcal{A}$ has infinite signal counts, there still exists some signal j outside of $\mathcal{A}$ whose variance reduction is strictly positive.[21] By continuity, at sufficiently late periods, observing signal j would reduce the variance by a positive amount. This is a profitable deviation from observing some signal in $\mathcal{A}$, leading to a contradiction!

Now that $\mathcal{A}$ is spanning, we can take $\mathcal{S}$ to be the efficient minimal spanning set in the subspace spanned by $\mathcal{A}$. To prove Theorem 3, we will show the long-run frequencies are positive precisely for the signals in $\mathcal{S}$. Ignoring the initial periods, it is without loss to assume that only signals in $\overline{\mathcal{A}}$ are available. It suffices to show that whenever the signals observed infinitely often *span the entire subspace*, agents eventually observe the efficient subset $\mathcal{S}$. To ease notation, we assume this subspace is the entire $\mathbb{R}^K$, and prove the following result:

Theorem 3 Restated. Suppose that the signals observed infinitely often span $\mathbb{R}^K$. Then society eventually observes signals in $\mathcal{S}^*$ with frequencies λ^*.

The next sections are devoted to the proof of this restatement.

---

that are independent over $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$. But it could be shown that the argument extends to mild correlation across states. We omit the somewhat cumbersome details, which do not add any further intuition.

[21] Formally, let $s_1, \ldots, s_N$ denote the limit signal counts, where $s_i = \infty$ if and only if $i \in \mathcal{A}$. Then there exists j such that $f(s_j + 1, s_{-j}) < f(s_j, s_{-j})$. This is because if $f(s_j + 1, s_{-j}) = f(s_j, s_{-j})$ for each j, then the partial derivatives of f at s are all zero. Since f is continuously differentiable, this would imply all directional derivatives of f are also zero. By the convexity of f, $f(s)$ must achieve minimum value. But by assumption there exists a spanning set, so $f(q) = 0$ if $q_1, \ldots, q_N$ are all infinite. This contradicts $f(s) > 0$.

E.2 Controlling the Derivatives

To study the posterior variance function f, it will be convenient to instead work with the homogenous function f^* we introduced in Lemma 3. We formalize this connection as follows:

Lemma 8. *Suppose that signals in $\mathcal{A}$ span $\mathbb{R}^K$. Then, as $q_i \to \infty$ for each $i \in \mathcal{A}$,*

$$f(q_1, \dots, q_N) \sim \frac{1}{t} \cdot f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right) \quad \text{with} \quad t = \sum_{i=1}^{N} q_i$$

The partial derivatives and second partial derivatives also satisfy the approximations

$$\partial_j f(q_1, \dots, q_N) \sim \frac{1}{t^2} \cdot \partial_j f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right)$$

$$\partial_{jj} f(q_1, \dots, q_N) \sim \frac{1}{t^3} \cdot \partial_{jj} f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right)$$

Proof. Recall that

$$f(q_1, \dots, q_N) = \left[((V^0)^{-1} + C'QC)^{-1}\right]_{11}.$$

Since $q_i \to \infty$ for $i \in \mathcal{A}$, the least eigenvalue of the matrix $C'QC$ approaches infinity. That is, for any $\epsilon > 0$, it holds eventually that $(V^0)^{-1} \preceq \epsilon \cdot C'QC$ in matrix order. Then

$$\frac{1}{1+\epsilon} \cdot [(C'QC)^{-1}]_{11} \leq f(q_1, \dots, q_N) \leq [(C'QC)^{-1}]_{11}.$$

Equivalently, this shows

$$\frac{1}{(1+\epsilon)t} \cdot f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right) \leq f(q_1, \dots, q_N) \leq \frac{1}{t} \cdot f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right).$$

Similar approximation holds for the derivatives, proving the lemma. □

Lemma 9. *Under the same assumptions as in Lemma 8, it holds that*

$$\frac{\partial_{jj} f(q_1, \dots, q_N)}{\partial_j f(q_1, \dots, q_N)} \to 0$$

and similarly

$$\frac{\partial_{jj} f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right)}{t \cdot \partial_j f^*\left(\frac{q_1}{t}, \dots, \frac{q_N}{t}\right)} \to 0$$

Proof. It suffices to prove the first result. From $f(q_1, \dots, q_N) = e_1' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot e_1$ we compute that

$$\partial_j f = -e_1' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot c_j \cdot c_j' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot e_1$$

and

$$\partial_{jj} f = 2e_1' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot c_j \cdot c_j' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot c_j \cdot c_j' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot e_1.$$

Let $\gamma_j = e_1' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot c_j$, which is a number. Then the above shows

$$\partial_j f = -\gamma_j^2; \qquad \partial_{jj} f = 2\gamma_j^2 \cdot c_j' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot c_j.$$

Again, all eigenvalues of the matrix $(V^0)^{-1} + C'QC$ become large as $q_i \to \infty$ for $i \in \mathcal{A}$. Thus for arbitrarily large constant L, eventually $(V^0)^{-1} + C'QC \succeq L \cdot c_j c_j'$ in matrix norm. Then the number $c_j' \cdot [(V^0)^{-1} + C'QC]^{-1} \cdot c_j$ is arbitrarily small, and the above display shows $\partial_{jj} f$ is small compared to $\partial_j f$. □

The above lemmata imply that at sufficiently late periods along society's signal path, the variance reduction of any *discrete* signal can be approximated by the continuous partial derivative of f (or f^*). A direct corollary is the following:

Lemma 10. *For any $\epsilon > 0$, there exists sufficiently large $t(\epsilon)$ such that if signal j is observed in any period $t+1$ later than $t(\epsilon)$, then*

$$\partial_j f^* \left(\frac{m(t)}{t} \right) \leq (1 - \epsilon) \min_{1 \leq l \leq N} \partial_l f^* \left(\frac{m(t)}{t} \right).$$

That is, the signal choice in any sufficiently late period *almost* minimizes the directional derivative of f^*.

E.3 (Pseudo) Gradient Descent of f^*

We define $\lambda(t) = \frac{m(t)}{t} \in \Delta^{N-1}$. If j is the signal choice in period $t+1$, then it is easily checked that

$$\lambda(t+1) = \frac{t}{t+1} \lambda(t) + \frac{1}{t+1} e_j.$$

The frequencies $\lambda(t)$ move in the direction of e_j, which is the direction where f^* decreases almost the fastest (by Lemma 10). Thus, the evolution of $\lambda(t)$ over time resembles the gradient descent dynamics—the value of $f^*(\lambda(t))$ roughly decreases over time, and we can expect that eventually $\lambda(t)$ approaches the unique minimizer λ^* of f^*.

To formalize this intuition, we consider (for fixed $\epsilon > 0$ and sufficiently large t)

$$\begin{aligned}
f^*(\lambda(t+1)) &= f^* \left(\frac{t}{t+1} \lambda(t) + \frac{1}{t+1} e_j \right) \\
&= f^* \left(\frac{t}{t+1} \lambda(t) \right) + \frac{1}{t+1} \cdot \partial_j f^* \left(\frac{t}{t+1} \lambda(t) \right) + O\left(\frac{1}{(t+1)^2} \cdot \partial_{jj} f^* \left(\frac{t}{t+1} \lambda(t) \right) \right) \\
&\leq f^* \left(\frac{t}{t+1} \lambda(t) \right) + \frac{1-\epsilon}{t+1} \cdot \partial_j f^* \left(\frac{t}{t+1} \lambda(t) \right) \\
&= \frac{t+1}{t} \cdot f^*(\lambda(t)) + \frac{(1-\epsilon)(t+1)}{t^2} \cdot \partial_j f^*(\lambda(t)) \\
&\leq f^*(\lambda(t)) + \frac{1}{t} \cdot f^*(\lambda(t)) + \frac{1-2\epsilon}{t} \cdot \min_{1 \leq l \leq N} \partial_l f^*(\lambda(t)).
\end{aligned} \tag{18}$$

The first inequality uses Lemma 9, the next equality uses the homogeneity of f^*, and the last inequality uses Lemma 10.

Write $\lambda = \lambda(t)$ for short. Observe that f^* is continuously differentiable at λ, since $\lambda_i(t) > 0$ for $i \in \mathcal{A}$, which spans the entire space. Thus the convexity of f^* yields

$$f^*(\lambda^*) \geq f^*(\lambda) + \sum_{j=1}^{N} (\lambda_j^* - \lambda_j) \cdot \partial_j f^*(\lambda).$$

The homogeneity of f^* implies $\sum_{j=1}^{N} \lambda_j \cdot \partial_j f^*(\lambda) = -f^*(\lambda)$. This enables us to rewrite the above display as

$$\sum_{j=1}^{N} \lambda_j^* \cdot \partial_j f^*(\lambda) \leq f^*(\lambda^*) - 2f^*(\lambda).$$

Thus, in particular,

$$\min_{1 \leq l \leq N} \partial_l f^*(\lambda(t)) \leq f^*(\lambda^*) - 2f^*(\lambda). \tag{19}$$

Combining (18) and (19), we have for all large t:

$$f^*(\lambda(t+1)) \leq f^*(\lambda(t)) + \frac{1}{t} \cdot [(1-2\epsilon) \cdot f^*(\lambda^*) - (1-4\epsilon) \cdot f^*(\lambda(t))]. \tag{20}$$

We claim this implies $f^*(\lambda(t)) \leq (1+4\epsilon) \cdot f^*(\lambda^*)$ holds for all large t. Indeed, if this holds for *some* t, then (20) implies the same is true at future periods. It thus suffices to show the opposite inequality $f^*(\lambda(t)) > (1+4\epsilon) \cdot f^*(\lambda^*)$ cannot hold at every large t. By (20), that would give $f^*(\lambda(t+1)) \leq f^*(\lambda(t)) - \frac{\epsilon \cdot f^*(\lambda^*)}{t}$. But since the harmonic series diverges, $f^*(\lambda(t))$ would then decrease without bound, leading to a contradiction!

Hence we have shown that for any fixed ϵ, $f^*(\lambda(t)) \leq (1+4\epsilon) \cdot f^*(\lambda^*)$ eventually. As λ^* is the unique minimizer of f^*, this implies $\lambda(t) \to \lambda^*$, which proves Theorem 3.

E.4 A Stronger Result

In the above, we showed that if the signals observed infinitely often span $\mathbb{R}^K$, then the signals observed with positive frequencies are exactly those in the best minimally spanning set $\mathcal{S}^*$. However, this leaves open the possibility that some signals outside of $\mathcal{S}^*$ are observed infinitely often, yet with zero long-run frequency. Below we show this is not possible when $|\mathcal{S}^*| = K$. More specifically, suppose $|\mathcal{S}^*| = K$ and $m_i(t) \sim \lambda_i^* \cdot t, \forall i$, then in fact the stronger conclusion $m_i(t) = \lambda_i^* \cdot t + O(1)$ also holds.[22] Together with Remark 2, this suggests that the difference between $m_i(t)$ and the optimal $n_i(t)$ remains bounded.[23]

[22] Thus, the conclusion of Corollary 4 can be strengthened.

[23] We believe but cannot prove that $m_i(t) = \lambda_i^* \cdot t + O(1)$ holds more generally, even if $|\mathcal{S}^*| < K$. Equivalently, we conjecture that any signal with zero long-run frequency is in fact only observed finitely many times.

So let us assume $|\mathcal{S}^*| = K$. Without loss, $\mathcal{S}^* = \{1, \ldots, K\}$ is the first K signals. By the previously established (**), the first K partial derivatives of f^* are equal at λ^* and they are strictly smaller (i.e., more negative) than the other partial derivatives. Since these partial derivatives are continuous, we can find $\epsilon > 0$ such that whenever λ is within ϵ distance from λ^*, it holds that

$$\partial_i f^*(\lambda) < (1+\epsilon) \cdot \partial_j f^*(\lambda), \quad \forall 1 \leq i \leq K < j$$

By assumption we have $\lambda(t) = \frac{m(t)}{t} \to \lambda^*$. Thus at sufficiently late periods, Lemma 10 implies that the signal choice must be within the first K signals. This shows signals outside of $\mathcal{S}^*$ are observed finitely often, as desired. And for any signal i in $\mathcal{S}^*$, its signal count satisfies $m_i(t) = \lambda_i^* \cdot t + O(1)$ by Proposition 1. This completes the proof of the stronger result here.

F Proof of Proposition 2

We will prove that given any prior belief, the policy-maker can provide $K-1$ sufficiently precise signals so that once they are processed, society eventually observes the best set $\mathcal{S}^*$. In fact, the following argument shows that the planner can provide these free signals at any time t, not necessarily before endogenous information acquisition takes place.

The proof of the proposition closely resembles the proof of the restated Theorem 3, see Appendix E. Indeed, with sufficiently high precision on the free signals, it is as if each free signal has unit precision but is observed many times. Thus, as long as the $K-1$ free signals span $\theta_2, \ldots, \theta_K$, the restated Theorem 3 applies since society eventually learns θ_1 anyways. Of course, the assumption of that theorem is not exactly satisfied, and one may wonder whether *observing a signal many times has the same consequence as observing it infinitely often*. In what follows we show how to resolve this concern.

Consider for simplicity that the planner provides L i.i.d. free signals in $\mathcal{A} \subset [N]$ (which spans $\theta_2, \ldots, \theta_K$), where we are free to choose L by making γ sufficiently large. This corresponds to restricting $m_i(t) \geq L$ for each $i \in \mathcal{A}$. Fix any $\epsilon > 0$, there exists such an L that the approximations in Lemma 8 and 9 hold up to a margin of error no more than ϵ. That is, for Lemma 8, we now have

$$(1-\epsilon) \cdot f(q_1, \ldots, q_N) \leq \frac{1}{t} \cdot f^*\left(\frac{q_1}{t}, \ldots, \frac{q_N}{t}\right) \leq (1+\epsilon) \cdot f(q_1, \ldots, q_N)$$

etc., and we similarly modify Lemma 9 to

$$\frac{\partial_{jj} f(q_1, \ldots, q_N)}{\partial_j f(q_1, \ldots, q_N)} \leq \epsilon.$$

These hold because the signal precision matrix $C'QC$ eventually dominates the prior precision matrix $(V^0)^{-1}$.

Then, Lemma 10 still holds, with fixed ϵ and sufficiently large L. We could then derive (18), (19) and (20) in the same way as before. This enables us to conclude $f^*(\lambda(t)) \leq (1+4\epsilon)\cdot f^*(\lambda^*)$ at every late period t. Note that ϵ has been fixed. Thus, the inequality

$$f^*(\lambda(t)) \leq (1+4\epsilon)\cdot f^*(\lambda^*)$$

does not by itself imply that $\lambda(t) \to \lambda^*$. However, if we had chosen ϵ to be sufficiently small, then $\lambda(t)$ eventually belongs to a small neighborhood of λ^*. In particular, we could have chosen ϵ so that the above inequality implies $\lambda_i(t) \geq \frac{\lambda_i^*}{2} > 0$ for each $i \in \mathcal{S}^*$.

For such a choice of ϵ and corresponding L, we know that society observes each signal in $\mathcal{S}^*$ with positive frequencies. But Theorem 3 shows that the set of signals with positive frequencies is a minimal spanning set. So this set must be $\mathcal{S}^*$ itself, and the long-run frequencies must be λ^*. This proves that any intervention with a sufficiently large L achieves efficient learning. Proposition 2 follows.

G Multiple Payoff-Relevant States

In this appendix, we consider optimal long-run acquisitions for the problem of predicting multiple states. We assume that society seeks to minimize the sum of his belief variances about $\theta_1, \ldots, \theta_K$. His objective function is to minimize

$$F(q_1, \ldots, q_N) = Tr\left[((V^0)^{-1} + C'QC)^{-1}\right].$$

subject to the signal counts q_i being integers and summing up to t. We use "Tr" to denote the trace of a matrix.

The solution to this minimization problem turns out to be very complex when $N > K$. To make the problem more tractable, we impose a further assumption that the signal coefficient vectors c_i have the same norm. This allows us to focus the analysis on the directions of the signals, rather than their precisions.

Assumption 6 (Unit Norm)**.** *Each vector $c_i \in \mathbb{R}^K$ has norm 1.*

Given this assumption, a basic question is to understand how fast society can jointly learn about different states. If the signals are simply $\theta_1 + \epsilon_1, \ldots, \theta_K + \epsilon_K$, then society cannot do better (in the long run) than spending the same number ($\frac{t}{K}$) of observations on each signal. In so doing, its posterior variance at time t about each state θ_i is approximately $\frac{K}{t}$, and the sum of these variances is $\frac{K^2}{t}$. Our next result shows this is asymptotically best, even when additional signals are available.

Proposition 3. *Under Assumption 6, we have*

$$\liminf_{q_1+\cdots+q_N\to\infty} (q_1 + \cdots + q_N)\cdot F(q_1, \ldots, q_N) \geq K^2.$$

For the special case of $K = 2$, we are able to determine the exact asymptotic variance (the value of the LHS above) for any given set of signals, see later in this Appendix. Deriving the analogous result for general K is left for future work.

We highlight that unlike the case of a single payoff-relevant state, here the minimum asymptotic variance can in general be achieved by more than one vector of frequencies. Thus, the above results only describe agents' payoffs at large t, but they do not pin down agents' optimal behavior. When q_i is not restricted to integer values, Chaloner (1984) showed that the minimum posterior variance at any *fixed* time t is achieved by focusing on at most $\frac{K(K+1)}{2}$ signals. However, it is not known whether the same subset of $\frac{K(K+1)}{2}$ signals are observed for all large t, and her result also does not extend to our integer design problem.

G.1 Proof of Proposition 3

We first show that

$$F^*(\lambda) := \lim_{t\to\infty} t \cdot F(\lambda t) = Tr\left[(C'\Lambda C)^{-1}\right] \tag{21}$$

If at least K of $\lambda_1, \ldots, \lambda_N$ are positive, this follows from the previous formula for F. Suppose instead that only $\lambda_1, \ldots, \lambda_k$ are positive, with $k < K$. Consider the limit of $Tr\left[(C'\Lambda C)^{-1}\right]$ as $\lambda_{k+1}, \ldots, \lambda_N$ approaches zero. In this limit, the $K \times K$ matrix $C'\Lambda C$ approaches a rank k matrix, so an eigenvalue of $C'\Lambda C$ approaches zero. This means an eigenvalue of $(C'\Lambda C)^{-1}$ approaches infinity, and since all its eigenvalues are non-negative by positive-definiteness, we deduce $Tr\left[(C'\Lambda C)^{-1}\right] \to \infty$. Meanwhile, $F(\lambda t)$ is bounded away from zero since the first k signals cannot identify all of the states $\theta_1, \ldots, \theta_K$. Thus (21) always hold.

We need to show that if each signal coefficient vector c_i has norm 1, then $F^*(\lambda) \geq K^2$ for all $\lambda \in \Delta^{N-1}$. For this, consider the positive-definite $K \times K$ matrix $C'\Lambda C$. Let its K (positive) eigenvalues be $\beta_1, \ldots, \beta_K$, then we have

$$\beta_1 + \cdots + \beta_K = Tr(C'\Lambda C) = \sum_{i=1}^{N} \lambda_i \sum_{j=1}^{K} c_{ij}^2 = \sum_{i=1}^{N} \lambda_i = 1,$$

Observe that the eigenvalues of the inverse matrix $(C'\Lambda C)^{-1}$ are simply $\frac{1}{\beta_1}, \ldots, \frac{1}{\beta_K}$. Thus, by (21) and Cauchy-Schwartz inequality,

$$F^*(\lambda) = Tr\left[(C'\Lambda C)^{-1}\right] = \frac{1}{\beta_1} + \cdots + \frac{1}{\beta_K} \geq \frac{K^2}{\beta_1 + \cdots + \beta_K} = K^2.$$

This proves Proposition 3.

G.2 Characterization of Asymptotic Variance when $K = 2$

Suppose there are just two states and each signal has unit norm, we determine here the exact value of $\min_{\lambda \in \Delta^{N-1}} F^*(\lambda)$ for any given coefficient matrix C. By what we have shown, this value (divided by t) approximates the minimum of the objective function F that can be achieved given t observations.

Applying Lemma 3 and adding up the variances about θ_1 and θ_2, we have for $K = 2$,

$$F^*(\lambda) = \frac{\sum_{i=1}^N \lambda_i (x_i^2 + y_i^2)}{\sum_{1 \le i < j \le N} \lambda_i \lambda_j (x_i y_j - x_j y_i)^2},$$

where each signal coefficient vector $c_i = (x_i, y_i)'$. By Assumption 6, $x_i^2 + y_i^2 = 1$ for each i. Thus the numerator above is exactly 1, and we only need to *maximize* the denominator. It will be convenient to parametrize $(x_i, y_i) = (\cos \phi_i, \sin \phi_i)$, with $\phi_i \in [0, \pi)$ distinct from one another.[24] Then, the denominator becomes

$$\begin{aligned}
\sum_{1 \le i < j \le N} \lambda_i \lambda_j (x_i y_j - x_j y_i)^2 &= \sum_{1 \le i < j \le N} \lambda_i \lambda_j \sin^2(\phi_i - \phi_j) = \frac{1}{4} \cdot \sum_{1 \le i < j \le N} \lambda_i \lambda_j (2 - 2\cos(2\phi_i - 2\phi_j)) \\
&= \frac{1}{4}(\lambda_1 + \cdots + \lambda_N)^2 - \frac{1}{4} \sum_{i=1}^N \lambda_i^2 - \frac{1}{4} \sum_{1 \le i < j \le N} \lambda_i \lambda_j 2\cos(2\phi_i - 2\phi_j) \\
&= \frac{1}{4} - \sum_{i,j=1}^N \lambda_i \lambda_j \cos(2\phi_i - 2\phi_j) \\
&= \frac{1}{4} - \left(\sum_{i=1}^N \lambda_i \cos 2\phi_i \right)^2 - \left(\sum_{i=1}^N \lambda_i \sin 2\phi_i \right)^2.
\end{aligned}$$

This recovers the result of Proposition 3 that $F^*(\lambda) \ge 4$. More generally, given $\phi_1, \ldots, \phi_N$, let $u_i = (\cos 2\phi_i, \sin 2\phi_i)$ be a vector/point lying on the unit circle. Then society seeks to *minimize* $(\sum_{i=1}^N \lambda_i \cos 2\phi_i)^2 + (\sum_{i=1}^N \lambda_i \sin 2\phi_i)^2$, which is the squared norm of the vector $\sum_{i=1}^N \lambda_i u_i$. Taking a geometric perspective, this problem is to choose a point in the convex hull of points $u_1, \ldots, u_N$ that is closest to the origin. There are two possibilities:

1. Suppose the points $u_1, \ldots, u_N$ lie on a semi-circle. Without loss, we label u_1 as the point closest to one end of this semi-circle and u_2 being closest to the other end. Then the point in $Conv(u_1, \ldots, u_N)$ that is closest to the origin is the mid-point between u_1 and u_2. In this case F^* is *uniquely* minimized at $\lambda = (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$. The minimum value of F^* is strictly larger than 4 except when $u_1 = -u_2$ (equivalently, when the original signal coefficients c_1, c_2 are orthogonal).

[24] $\phi_i \in [\pi, 2\pi)$ can be replaced by $\phi_i - \pi$, corresponding to replacing the vector c_i by $-c_i$.

2. Suppose the points $u_1, \ldots, u_N$ do not lie on a semi-circle. Then their convex hull contains the origin in the interior and in particular $N > 2$. We can find three of these N points, say u_1, u_2, u_3, such that the triangle connecting these three points contains the origin. Then, F^* is minimized at $\lambda = (\lambda_1, \lambda_2, \lambda_3, 0, \ldots, 0)$, where $\lambda_1, \lambda_2, \lambda_3$ are unique weights such that $\lambda_1 u_1 + \lambda_2 u_2 + \lambda_3 u_3 = \mathbf{0}$. In this case the minimum value of F^* is exactly 4.

We note that in the latter case, whenever $N > 3$, there is not a unique set of three points whose convex hull contains the origin. Thus F^* is not uniquely minimized, and we cannot use the analogue of Lemma 4 to characterize society's optimal divisions.