The value of conceptual knowledge

Benjamin Davies and Anirudh Sankar*

Draft version: April 2, 2026

Click here for latest version

Abstract

We define and quantify the instrumental value of conceptual knowledge. Such knowledge tells agents how unknown, payoff-relevant states relate. It is distinct from the statistical knowledge gained from data on those states. We formalize this distinction in a Bayesian decision framework with Gaussian priors and quadratic losses. Conceptual knowledge is valuable because it empowers agents to collect more valuable data. It is more valuable when states are more "reducible": when they can be explained with fewer common concepts. Its value is non-monotone in sample size and vanishes when samples have infinite size. Agents who know more concepts can attain the same payoffs with less data. This is especially true when states are highly reducible.

Keywords: concepts, dimension reduction, eigenvalues, mental models, statistical decisions, value of information
JEL classification: C44, D83

*Department of Economics, Stanford University; bldavies@stanford.edu and asankar@stanford.edu. We thank Steve Callander, Arun Chandrasekhar, Ben Golub, Matt Jackson, Annie Liang, Jann Spiess, and seminar participants at Motu and Stanford for helpful discussions and comments.

1 Introduction

Humans use mental models to make sense of the world (Johnson-Laird, 1983). The building blocks of these models are "concepts": mental representations we use to describe objects and how they relate (Murphy, 2002). Understanding these relationships allows us to use information about one object to draw inferences about another (Mitchell, 2021).

This paper studies the interaction between the conceptual knowledge embedded in mental models and the statistical knowledge gained from data. It is well-known that data are instrumentally valuable insofar as they contain payoff-relevant information (Howard, 1966; Raiffa and Schlaifer, 1961). We ask: when and why are *concepts* valuable?

An illustrative example. Suppose a farmer wants to learn which fertilizers to apply to his crops. He views fertilizers as "black boxes," knowing *that* they help crops grow but not *why*. He does not know how fertilizers' effects relate and cannot extrapolate one from another's. So when he tries fertilizers to learn their effects, he has to try them separately.

Now suppose the farmer knows fertilizers supply nitrogen, a nutrient that helps crops grow. This tells him how fertilizers' effects relate: they share a common "nitrogen component." He can use this component to extrapolate one fertilizer's effect from another's. Moreover, when he tries different fertilizers to learn their effects, he can combine them to isolate the nitrogen component. This is better than trying each fertilizer separately because it allows him to learn their effects from one trial rather than many.

Nitrogen is a mental construct—the farmer cannot see it. All he can see are the effects of using different fertilizers. Yet his knowledge of the concept "nitrogen" allows him to learn more efficiently. It empowers him to run trials that are more informative and instrumentally valuable.[1] *How much more valuable* is a quantity we define and characterize in this paper. We call this quantity the "value of conceptual knowledge."

We also quantify the value of having "deeper" knowledge. For example, beyond knowing that fertilizers supply nitrogen, the farmer may know nitrogen is present in different forms (e.g., ammonium and nitrate) that become available to plants at different rates. Then he can reason not only in terms of a "nitrogen component" but also a deeper "release rate component." This allows the farmer to extract more value from each trial and obtain the same payoffs with fewer trials. We study *how many fewer* in this paper.

[1] Knowing about nitrogen also enables out-of-sample prediction: once the farmer learns the nitrogen component, he can predict the effect of any nitrogen-containing fertilizer he has not tried.

Contributions. We make three contributions. First, we describe what it means to "know concepts" and define their instrumental value. Our definition builds on that of the value of information (Howard, 1966; Raiffa and Schlaifer, 1961): whereas information is valuable because it leads to better decisions, conceptual knowledge is valuable because it leads to better information.

Second, we characterize the value of conceptual knowledge in a Bayesian decision framework with Gaussian priors and quadratic losses. This framework admits a closed-form expression for the value of optimally acquired information, allowing us to analyze and clarify how and why this value changes. Moreover, our focus on a specific decision problem allows us to generate insights and intuitions that are not immediate from studying abstract problems (as in, e.g., Blackwell (1951, 1953) and Whitmeyer (2026)).[2]

Third, we use our framework to formalize what it means to have "deeper" conceptual knowledge. This allows us to compare the marginal values of having deeper knowledge and having more data. It also recognizes conceptual knowledge as an economic good one may acquire in the same way information is a good one acquires. In this way, we advance the literature on learning and information acquisition that treats conceptual knowledge as fixed and minimally restrictive (e.g., Bardhi, 2024; Callander, 2011; Schwartzstein, 2014).

Model and results. We consider a Bayesian agent who collects a sample of observations before taking actions. Each observation is a noisy signal of a linear combination of unknown, payoff-relevant states. The agent chooses these combinations based on his prior beliefs about the state vector. These beliefs encode his conceptual knowledge: the eigenvectors of the prior variance matrix represent concepts, and the corresponding eigenvalues measure concepts' explanatory power. An agent with conceptual knowledge knows which concepts explain more of the states' variance—intuitively, he can do principal component analysis (hereafter "PCA") *before* seeing any data. This "pre-data PCA" leads him to collect more valuable data because he focuses on the dimensions that matter most. In contrast, a naïve agent with no conceptual knowledge cannot focus on these dimensions. The value of conceptual knowledge equals the gain in maximal sample value from knowing which dimensions to focus on.

We establish three main results. First, conceptual knowledge is more valuable when states are more "reducible": when their prior variances are explained by fewer common

[2]For example, our non-monotonicity result (Theorem 2) could not be derived from Blackwell (1951, 1953) or Whitmeyer's (2026) frameworks without imposing more structure on them.

concepts (Theorem 1). Formally, this happens when the eigenvalues of the prior variance matrix are more spread out. If the states are explained by a single concept, then the agent gains a lot from knowing that concept and focusing on it when he collects data (i.e., "asking the right question"). In contrast, if every concept has the same explanatory power, then the agent gains nothing from knowing those concepts because he designs the same sample as he would if he was naïve.

Second, the value of conceptual knowledge is non-monotone in sample size and vanishes as this size grows without bound (Theorem 2). If the agent has more data, then he can learn more from them, raising the value gain from knowing which concepts to focus on. However, having more data also prompts him to broaden his focus, lowering the gain from knowing which concepts to focus on. The first effect dominates the second when the sample size is small. As it becomes large, the agent's posterior becomes independent of his prior, and so the conceptual knowledge embedded in his prior becomes irrelevant and loses its instrumental value. Intuitively, if he has infinite data, then he does not benefit from doing "pre-data PCA" because he can do traditional (post-data) PCA.

Third, an agent with "deeper" conceptual knowledge (i.e., who knows more principal components of the prior variance matrix) can attain the same payoffs with less data (Theorem 3). This is especially true when states are highly reducible. Then the agent can extract more value from each observation, lowering the number they need to attain a given payoff.

These theoretical results are consistent with empirical evidence. In Sankar et al. (2025), we experimentally study the effect of sharing relevant concepts with farmers who observe data on fertilizer outcomes. We compare farmers given concepts and data to farmers given data only, and find that farmers given concepts make more profitable decisions. This validates our theoretical prediction that conceptual knowledge makes data more valuable.

Our results also bear on comparisons between human and machine intelligence. Humans devise concepts that unite seemingly unrelated phenomena: Newton devised a concept (gravity) that unites apples falling on Earth with the orbits of other planets; Watson and Crick devised a concept (DNA) that unites crime scene investigation with the origins of domesticated rice; Bernoulli devised a concept (risk aversion) that unites choices in poker with choices between insurance policies. These and other concepts allow humans to generalize (Mitchell, 2021) and learn from limited data (Tenenbaum et al., 2011). In contrast, machines rely on recognizing patterns in large sets of data (Goodfellow et al., 2016; Halevy et al., 2009). Our results capture this asymmetry: conceptual knowledge compensates for data scarcity (Theorem 3), but loses its value when data are abundant (Theorem 2).

Roadmap. Section 2 elaborates on our example of a farmer learning about fertilizers. Section 3 presents our theoretical framework. Section 4 shows how information is optimally acquired in our framework. Section 5 defines and characterizes the value of conceptual knowledge. Section 6 formalizes our notion of "deeper" knowledge, and compares having deeper knowledge to having more data. Section 7 discusses our modeling assumptions and related literatures. Section 8 concludes. Appendix A contains proofs of our mathematical claims.

A prior version of this paper (Davies and Sankar, 2026) contains additional discussions and results.

2 An illustrative example

We begin with an example inspired by our empirical work in Uganda (Sankar et al., 2025), which shows that conceptual knowledge helps farmers learn about fertilizers.

Environment. A Bayesian farmer wants to learn the effect $\theta_k \in \mathbb{R}$ of applying fertilizer $k \in \{1, 2\}$ to his crops. His prior on $\theta \equiv (\theta_1, \theta_2)$ is a normal distribution with variance $\mathbb{V}(\theta)$. He observes the outcome

$$y = \theta_1 w_1 + \theta_2 w_2 + u$$

of using $w_1 \in \mathbb{R}$ more units of fertilizer 1 and $w_2 \in \mathbb{R}$ more units of fertilizer 2.[3] The vector $w = (w_1, w_2)$ has Euclidean length $\|w\| = 1$ and the error $u \in \mathbb{R}$ is independently normally distributed with variance $\sigma_u^2 > 0$.[4] It captures the randomness in y due to variation in unobserved factors.

The farmer's data $\mathcal{S} \equiv \{(w, y)\}$ comprise the vector w and outcome y. These data are valuable insofar as they make the farmer's beliefs about θ more precise. We measure the value of $\mathcal{S}$ via the mean difference

$$\pi(\mathcal{S}) \equiv \frac{1}{2} \sum_{k=1}^{2} (\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k \mid \mathcal{S}))$$

between the prior and posterior variances of θ_1 and θ_2.[5]

[3] We interpret negative values of w_k as using less of fertilizer k than the farmer uses currently.

[4] We fix $\|w\| = 1$ so that only the direction of w (and not its magnitude) affects the informativeness of y.

[5] In Section 3.1, we derive $\pi(\mathcal{S})$ as the value of information in a more general decision problem.

The farmer chooses the weight vector w that maximizes $\pi(\mathcal{S})$. Intuitively, he chooses the combination of fertilizers that teaches him as much as possible about their effects. That he can only choose *one* combination reflects the scarcity and cost of relevant data: our Ugandan setting is one of many where humans must learn from limited data.

Conceptual knowledge. The farmer knows the two fertilizers supply equal amounts of nitrogen, a nutrient that helps crops grow. He cannot see or touch nitrogen; it is a mental construct. But he can use his conceptual knowledge of nitrogen to express each fertilizer's effect θ_k as the sum of a common "nitrogen effect" and an orthogonal "other effect." He encodes these effects by the scalars

$$\gamma_1 \equiv \frac{\theta_1 + \theta_2}{\sqrt{2}} \quad \text{and} \quad \gamma_2 \equiv \frac{\theta_1 - \theta_2}{\sqrt{2}},$$

allowing him to express the effect vector

$$\theta = \gamma_1 v_1 + \gamma_2 v_2$$

as a linear combination of two unit vectors

$$v_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} \quad \text{and} \quad v_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}.$$

These vectors form an orthonormal basis for the Euclidean space $\mathbb{R}^2$ containing θ. The common and other effects γ_1 and γ_2 are the coordinates of θ over this basis.

The farmer knows v_1 and v_2, but does not know γ_1 or γ_2. Knowing v_1 and v_2 makes learning θ equivalent to learning $\gamma \equiv (\gamma_1, \gamma_2)$. So he does not have to learn each fertilizer's effect separately; instead, he can learn the common nitrogen effect and extrapolate the overall effects. This helps because he only has one observation (w, y) from which to infer two unknowns θ_1 and θ_2. Knowing how these unknowns relate (via v_1 and v_2) allows him to learn about both at the same time by choosing w appropriately.

The farmer's choice of w depends on the relative contributions of γ_1 and γ_2 to the prior variances of θ_1 and θ_2. He knows γ_1 contributes more: the fertilizers' effects are mostly determined by how much nitrogen they supply. So he assumes γ_1 and γ_2 are independently distributed with variances $\lambda_1 = \sigma^2(1+\rho)$ and $\lambda_2 = \sigma^2(1-\rho)$. The sum

$$\begin{aligned}\lambda_1 + \lambda_2 &= \mathbb{V}\left(\frac{\theta_1 + \theta_2}{\sqrt{2}}\right) + \mathbb{V}\left(\frac{\theta_1 - \theta_2}{\sqrt{2}}\right)\\ &= \mathbb{V}(\theta_1) + \mathbb{V}(\theta_2)\end{aligned}$$

of these variances equals the sum of the prior variances of θ_1 and θ_2, and $\rho \in [0, 1)$ determines the share

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1+\rho}{2}$$

of this sum contributed by γ_1. This share equals $1/2$ when $\rho = 0$, in which case γ_1 and γ_2 contribute equally. It equals one in the limit as $\rho \to 1$, in which case only γ_1 contributes.

Given the specifications of (v_1, v_2) and (λ_1, λ_2), the effect vector θ has prior variance

$$\mathbb{V}(\theta) = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{1}$$

Thus θ_1 and θ_2 have equal prior variances σ^2 and correlation ρ. Intuitively, the more θ_1 and θ_2 are determined by the common effect γ_1 of supplying nitrogen, the more likely they are to have similar values.

The prior variance matrix (1) has eigendecomposition

$$\mathbb{V}(\theta) = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T.$$

Each eigenvalue λ_k equals the prior variance of θ in the direction of the corresponding eigenvector v_k. So θ has the most prior variance in the direction of v_1 and the least in the direction of v_2.

Value of information. The value $\pi(\mathcal{S})$ of $\mathcal{S} = \{(w, y)\}$ is largest when $w = \pm v_1$ and smallest when $w = \pm v_2$ (see Corollary 1). For example, choosing $w = v_1$ makes $y = \gamma_1 + u$ a "pure signal" of the common nitrogen effect γ_1. This makes $\mathcal{S}$ maximally valuable because it provides information about the component of θ with the most prior variance, leading to the largest difference between prior and posterior variances. In contrast, choosing $w = v_2$ makes $y = \gamma_2 + u$ a pure signal of the other effect γ_2. This makes $\mathcal{S}$ *minimally* valuable because it provides information about the component of θ with the *least* prior variance, leading to the *smallest* difference between prior and posterior variances.

Value of conceptual knowledge. The data $\mathcal{S}$ have maximal value

$$\pi^* \equiv \max_{\|w\|=1} \pi(\mathcal{S}).$$

The farmer attains π^* by choosing $w = \pm v_1$; that is, by combining fertilizers so as to isolate their common nitrogen effect. This choice relies on his conceptual knowledge that

the nitrogen effect (i) exists, (ii) corresponds to the eigenvector v_1, and (iii) explains most of the fertilizer effects' prior variances. Without this knowledge, the farmer would have no way to represent each fertilizer's effect θ_k as the sum of components with differential contributions to its prior variance. So he would assume equal contributions (i.e., $\rho = 0$) and his data would have maximal value

$$\pi^{(0)} \equiv \max_{\|w\|=1} \Big[\pi(\mathcal{S})\big|_{\rho=0}\Big].$$

The difference

$$\Pi \equiv \pi^* - \pi^{(0)}$$

between π^* and $\pi^{(0)}$ captures the value of the farmer's conceptual knowledge: the value of knowing about nitrogen and using this knowledge to make his data more valuable.

The value Π of the farmer's conceptual knowledge is larger when ρ is larger.[6,7] Intuitively, if most of fertilizers' effects come from supplying nitrogen, then the farmer can refine his prior a lot by isolating the nitrogen effect when he tries fertilizers. We formalize this intuition in Section 5, and generalize it to a setting in which $\mathcal{S}$ has arbitrary size and θ has arbitrary length. In this setting, conceptual knowledge is more valuable when the eigenvalues of the prior variance matrix $\mathbb{V}(\theta)$ are more spread out (see Theorem 1). This is why raising ρ raises Π: it raises $\lambda_1 = (1+\rho)\sigma^2$ and lowers $\lambda_2 = (1-\rho)\sigma^2$ without changing their mean $(\lambda_1 + \lambda_2)/2 = \sigma^2$.

3 Framework

We consider a Bayesian agent who collects data before making a statistical decision. This section describes the agent's environment and his conceptual knowledge of it.

3.1 Environment

Prior. There is a true but unknown vector $\theta \equiv (\theta_1, \ldots, \theta_K)$ of real-valued states. The agent's prior on θ is a probability distribution $\mathbb{P}$ over the K-dimensional Euclidean space $\mathbb{R}^K$.

[6] We have

$$\pi^* = \frac{(1+\rho)^2\sigma^4}{2((1+\rho)\sigma^2 + \sigma_u^2)} \quad \text{and} \quad \pi^{(0)} = \frac{\sigma^4}{2(\sigma^2 + \sigma_u^2)}$$

by Corollary 1 and the definition of $\pi^{(0)}$. So $\partial\pi^*/\partial\rho > 0$ and $\partial\pi^{(0)}/\partial\rho = 0$, implying $\partial\Pi/\partial\rho > 0$.

[7] For example, the correlation ρ will be close to one when the fertilizers supply nitrogen only, and close to zero when their nutrient profiles are very different.

This distribution is normal with known mean $\mu \in \mathbb{R}^K$ and variance matrix $\Sigma \in \mathbb{R}^{K \times K}$:

$$\mathbb{P} = \mathcal{N}(\mu, \Sigma).$$

We assume $K \geq 2$ is finite and Σ is invertible.

The agent derives $\mathbb{P}$ from his conceptual knowledge about θ. We describe this knowledge in Section 3.2.

Sample. The agent observes a sample $\mathcal{S} \equiv \{(w^{(i)}, y^{(i)})\}_{i=1}^n$ of finite size n. Each observation comprises a "covariate" $w^{(i)} \in \mathbb{R}^K$ with Euclidean length $\|w^{(i)}\| = 1$,[8] and an "outcome"

$$y^{(i)} = \theta^T w^{(i)} + u^{(i)} \tag{2}$$

equal to the sum of

$$\theta^T w^{(i)} = \sum_{k=1}^K \theta_k w_k^{(i)}$$

and an independently normally distributed error $u^{(i)}$ with mean zero and variance $\sigma_u^2 > 0$. Thus, each outcome $y^{(i)}$ provides a noisy signal of a weighted combination of states, where the weights are determined by the covariate $w^{(i)} \equiv (w_1^{(i)}, \ldots, w_K^{(i)})$.

Actions and losses. The agent uses his prior $\mathbb{P}$, the sample $\mathcal{S}$, and Bayes' rule to form posterior beliefs about θ. Then he chooses a vector $a \equiv (a_1, \ldots, a_K)$ of real-valued actions. These actions induce a loss

$$L(\theta, a) \equiv \frac{1}{K} \sum_{k=1}^K (a_k - \theta_k)^2$$

equal to the mean squared difference between them and the corresponding states.

Let $\mathbb{E}$ and $\mathbb{V}$ take expectations and variances with respect to the prior distribution $\mathbb{P}$. The agent chooses the action vector that minimizes his posterior expected loss:[9]

$$a \in \operatorname*{arg\,min}_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}]. \tag{3}$$

[8] Assuming the covariates have unit length normalizes the scales of the signals $y^{(1)}, \ldots, y^{(n)}$ so that only the directions of $w^{(1)}, \ldots, w^{(n)}$ (and not their magnitudes) affect signals' informativeness. It also ensures the Gram matrix (12) always has trace n (see Footnote 13). This allows us to associate optimal samples of size n with a unique Gram matrix (21).

[9] In Davies and Sankar (2026, Section A1), we show that the choice problem (3) is equivalent to a prediction problem that arises in the machine and statistical learning literatures.

Intuitively, the agent wants to estimate the states $\theta_1, \ldots, \theta_K$ accurately, and the accuracy of his estimates $a_1, \ldots, a_K$ is determined by their squared errors $(a_k - \theta_k)^2$.

Lemma 1 characterizes the optimal action vector (3) and the posterior expected loss it induces. This vector equals the posterior mean of θ. It induces a posterior expected loss equal to the mean of the posterior variances of $\theta_1, \ldots, \theta_K$.

Lemma 1. *The optimal action vector $a = \mathbb{E}[\theta \mid \mathcal{S}]$ induces posterior expected loss*

$$\min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{V}(\theta_k \mid \mathcal{S}). \tag{4}$$

Value of $\mathcal{S}$. If the agent did not observe the sample $\mathcal{S}$, then his minimized prior and posterior expected losses would be equal. The information in $\mathcal{S}$ is instrumentally valuable because it helps the agent take actions with lower expected losses. Accordingly, we define the "value of $\mathcal{S}$" to be the difference between his minimized prior and posterior expected losses:

$$\pi(\mathcal{S}) \equiv \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a')] - \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}]. \tag{5}$$

Substituting (4) into (5) yields an expression for $\pi(\mathcal{S})$ in terms of the states' prior and posterior variances:

$$\pi(\mathcal{S}) = \frac{1}{K} \sum_{k=1}^{K} (\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k \mid \mathcal{S})). \tag{6}$$

Intuitively, the sample is valuable insofar as it lowers states' variances, allowing the agent to estimate them more accurately.

Since the states $\theta_1, \ldots, \theta_K$ and signals $y^{(1)}, \ldots, y^{(n)}$ are jointly normally distributed, the posterior variance $\mathbb{V}(\theta_k \mid \mathcal{S})$ depends on $\mathcal{S}$ only via the covariates $w^{(1)}, \ldots, w^{(n)}$ (see Section 4.1). We assume the agent chooses these covariates to maximize $\pi(\mathcal{S})$ subject to the length constraints $\|w^{(i)}\| = 1$.

3.2 Conceptual knowledge

Mental model and concepts. Our definition of conceptual knowledge draws upon psychologists' and cognitive scientists': concepts are the building blocks of mental models (Johnson-Laird, 1983), are used to describe objects and how they relate (Murphy, 2002), and allow humans to generalize across objects (Mitchell, 2021). Accordingly, our agent's

conceptual knowledge allows him to describe the states $\theta_1, \ldots, \theta_K$ and how they relate. It gives him a mental model of

$$\theta = \sum_{k=1}^{K} \gamma_k v_k \tag{7}$$

as an unknown combination of known vectors $v_1, \ldots, v_K \in \mathbb{R}^K$. We call these vectors "concepts." They are the building blocks of the agent's mental model. They capture his environment's generalizable structure: each state θ_j depends on the j^{th} component of v_k via an unknown coefficient $\gamma_k \in \mathbb{R}$ that is independent of j. This allows the agent to generalize across states: signals of θ_1 provide information about $\gamma_1, \ldots, \gamma_K$, from which he can extrapolate $\theta_2, \ldots, \theta_K$.

For example, suppose $\theta_1, \ldots, \theta_K$ represent the effects of applying different fertilizers. Then $v_1, \ldots, v_K$ could encode nutrient quantities and $\gamma_1, \ldots, \gamma_K$ the fertilizer-invariant effects of supplying different nutrients. If a farmer learns one fertilizer's overall effect, then he can extrapolate the others' via their joint dependence on $\gamma_1, \ldots, \gamma_K$.

For convenience and without loss of generality, we assume $v_1, \ldots, v_K$ are orthonormal for the remainder of the paper.

Eigendecomposition. The agent does not know the coefficients $\gamma_1, \ldots, \gamma_K$, but he knows some contribute more to states' prior variances than others. Specifically, he knows each coefficient γ_k is independently distributed with prior variance $\lambda_k > 0$ non-increasing in k.[10] Then θ has prior variance

$$\begin{aligned} \Sigma &= V \Lambda V^T \\ &= \sum_{k=1}^{K} \lambda_k v_k v_k^T, \end{aligned} \tag{8}$$

where

$$\Lambda \equiv \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{bmatrix}$$

[10] It is without loss of generality to assume $\gamma_1, \ldots, \gamma_K$ are independently distributed. This is because Λ is positive-semidefinite, and so, by the spectral theorem, there is an orthogonal matrix $A \in \mathbb{R}^{K \times K}$ and diagonal matrix $\Lambda' \in \mathbb{R}^{K \times K}$ such that $\Lambda = A \Lambda' A^T$. Then $V' \equiv VA$ is orthogonal and Σ has eigendecomposition $V' \Lambda' (V')^T$, so we can carry out our analysis by replacing V with V' and Λ with Λ'.

is the $K \times K$ diagonal matrix with entries $\lambda_1 \geq \cdots \geq \lambda_K \geq 0$ and

$$V \equiv \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix}$$

is the $K \times K$ orthogonal matrix with columns $v_1, \ldots, v_K$.

Equation (8) is an eigendecomposition of Σ. The k^{th} largest eigenvalue $\lambda_k = \mathbb{V}(\gamma_k)$ of Σ equals the prior variance of θ in the direction of the corresponding unit eigenvector v_k. The trace

$$\operatorname{tr}(\Sigma) = \sum_{k=1}^{K} \lambda_k$$

of Σ equals the sum of the eigenvalues $\lambda_1, \ldots, \lambda_K$. So these eigenvalues' mean

$$\begin{aligned} \overline{\lambda} &\equiv \frac{1}{K} \sum_{k=1}^{K} \lambda_k \\ &= \frac{1}{K} \sum_{k=1}^{K} \mathbb{V}(\theta_k) \end{aligned}$$

equals the mean of the states' prior variances. The ratio $\lambda_k / \operatorname{tr}(\Sigma)$ equals the share of these variances contributed by γ_k. If the shares contributed by $\gamma_1, \ldots, \gamma_K$ are equal, then $\lambda_k = \operatorname{tr}(\Sigma)/K = \overline{\lambda}$ is constant in k and so $\Sigma = V \Lambda V^T$ is proportional to $K \times K$ identity matrix I_K:

$$V\big(\overline{\lambda} I_K\big) V^T = \overline{\lambda} I_K.$$

In contrast, if $\lambda_1 / \operatorname{tr}(\Sigma) \approx 1$, then γ_1 contributes most of the states' prior variances.

Intuitively, the agent does "pre-data" principal component analysis. PCA is a dimension reduction technique that projects a distribution onto its highest variance dimensions. Traditional PCA estimates these dimensions from data. In contrast, our agent derives them from his conceptual knowledge: he knows which dimensions have the highest variance *before* observing any data.

Reducibility. The distribution of $\lambda_1, \ldots, \lambda_K$ around their mean $\overline{\lambda} = \operatorname{tr}(\Sigma)/K$ captures the states' "reducibility." They are more "reducible" when their prior variances are explained by fewer common concepts: when $\lambda_1, \ldots, \lambda_K$ are more spread out around $\overline{\lambda}$. The agent's conceptual knowledge allows him to "reduce" the state vector θ by representing it as a low-dimensional combination of higher-dimensional concepts.

We formalize what it means for $\lambda_1, \ldots, \lambda_K$ to be "more spread out" as follows. Define their (empirical) cumulative distribution function (hereafter "CDF") by

$$F(z) = \frac{|\{k \in \{1, \ldots, K\} : \lambda_k \leq z\}|}{K} \tag{9}$$

for all $z > 0$. A "mean-preserving spread" (hereafter "MPS") of F is a CDF F' such that

(i) The distributions described by F and F' have the same mean:

$$\int_0^\infty z \, \mathrm{d}F(z) = \int_0^\infty z \, \mathrm{d}F'(z).$$

(ii) For all $z > 0$, the area under F' from 0 to z is at least the area under F from 0 to z:

$$\int_0^z (F'(t) - F(t)) \, \mathrm{d}t \geq 0.$$

These are the "integral conditions" defined by Rothschild and Stiglitz (1970). Condition (ii) says that F' has more weight in its tails than F, capturing how F' is more spread out.

We say $\lambda_1, \ldots, \lambda_K$ "undergo a MPS" (and so the states become "more reducible") when their CDF (9) undergoes a MPS. This changes the trace of the posterior variance matrix without changing the trace of Σ. So if $\lambda_1, \ldots, \lambda_K$ undergo a MPS, then the agent's posterior expected loss changes but his prior expected loss $\overline{\lambda}$ does not. This makes MPSs useful for analyzing how the value $\pi(\mathcal{S})$ of the sample $\mathcal{S}$ depends on the distribution of $\lambda_1, \ldots, \lambda_K$. We discuss this dependence in Section 5.

Naïve baseline. If the agent had no conceptual knowledge—i.e., if he did not have a mental model of θ as a combination of concepts with different explanatory powers—then he would not be able to reduce states in the manner described above. His prior variance matrix

$$\Sigma^{(0)} \equiv \overline{\lambda} I_K$$

would equal the matrix that obtains when $\lambda_k = \overline{\lambda}$ for each $k \in \{1, \ldots, K\}$. We call

$$\mathbb{P}^{(0)} \equiv \mathcal{N}(\mu, \Sigma^{(0)})$$

the "naïve" prior because it ignores the covariances among states stemming from their dependence on common concepts.

3.3 Example with pairwise correlated states

Finally, we give an example of how the prior variance matrix Σ encodes the agent's conceptual knowledge of the states.[11]

Suppose the agent knows each state θ_k has two components: a common component that is proportional to the states' mean and an idiosyncratic component that is independent across states. He encodes the common component by the unit vector

$$v_1 = \frac{1}{\sqrt{K}}\mathbf{1}_K,$$

where $\mathbf{1}_K \equiv (1, \ldots, 1)$ is the K-vector of ones. He encodes the idiosyncratic components by unit vectors $v_2, \ldots, v_K$ that are orthogonal to v_1 and each other. The k^{th} coefficient γ_k in (7) has prior variance

$$\lambda_k = \sigma^2 \begin{cases} 1 + \rho(K-1) & \text{if } k = 1 \\ 1 - \rho & \text{if } k > 1, \end{cases} \tag{10}$$

where $\sigma^2 > 0$ is the mean of $\lambda_1, \ldots, \lambda_K$ and where $\rho \in [0, 1)$ determines the share

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_K} = \frac{1}{K} + \rho\left(1 - \frac{1}{K}\right)$$

of the prior variances of $\theta_1, \ldots, \theta_K$ contributed by the coefficient γ_1 on v_1. This share equals $1/K$ when $\rho = 0$, in which case λ_k is constant in k and so $\gamma_1, \ldots, \gamma_K$ contribute to the prior variances of $\theta_1, \ldots, \theta_K$ equally. It equals one in the limit as $\rho \to 1$, in which case only γ_1 contributes. As ρ rises, the prior distribution of θ becomes more concentrated along the subspace of $\mathbb{R}^K$ spanned by v_1, and so the states become "more reducible" in the sense described in Section 3.2.[12]

[11] This example generalizes the setting described in Section 2. It builds Σ from first principles, starting with the eigenvalues and eigenvectors. In Davies and Sankar (2026, Section 3.2), we provide a second example that builds Σ from knowledge of how the states are generated, implying specific eigenvalues and eigenvectors.

[12] Indeed, the eigenvalues $\lambda_1, \ldots, \lambda_K$ defined by (10) have k^{th} partial sum

$$\sum_{j=1}^{k} \lambda_j = (k + \rho(K-k))\sigma^2,$$

which is increasing in ρ when $k < K$ and constant in ρ when $k = K$. Thus, by Lemma A1, these eigenvalues undergo a MPS when ρ rises.

Since $v_1, \ldots, v_K$ are orthonormal, the sum of their outer products equals the $K \times K$ identity matrix I_K. So the prior variance matrix

$$\begin{aligned}\Sigma &= \lambda_1 v_1 v_1^T + \lambda_K \big(I_K - v_1 v_1^T\big) \\ &= \rho\sigma^2 \mathbf{1}_K \mathbf{1}_K^T + (1-\rho)\sigma^2 I_K \\ &= \sigma^2 \begin{bmatrix} 1 & \rho & \cdots \\ \rho & 1 & \\ \vdots & & \ddots \end{bmatrix}\end{aligned} \tag{11}$$

is the $K \times K$ matrix with diagonal entries equal to σ^2 and off-diagonal entries equal to $\rho\sigma^2$. Thus, under the agent's prior, the states have equal variances σ^2 and pairwise correlations ρ.

4 Designing the sample $\mathcal{S}$

This section describes the maximally valuable design of the sample $\mathcal{S} \equiv \{(w^{(i)}, y^{(i)})\}_{i=1}^n$. First, we derive sharp bounds on the value $\pi(\mathcal{S})$ of $\mathcal{S}$. Second, we describe how the agent can choose $w^{(1)}, \ldots, w^{(n)}$ to attain the upper bound on $\pi(\mathcal{S})$. Finally, we characterize the sample designed by a "naïve" agent who lacks conceptual knowledge.

4.1 Bounds on $\pi(\mathcal{S})$

We can express the value (6) of $\mathcal{S}$ in terms of the traces of the prior and posterior variance matrices:

$$\pi(\mathcal{S}) = \frac{1}{K}(\mathrm{tr}(\Sigma) - \mathrm{tr}(\mathbb{V}(\theta \mid \mathcal{S}))).$$

Lemma 2 characterizes $\mathbb{V}(\theta \mid \mathcal{S})$ in terms of the prior variance matrix Σ and "Gram matrix"

$$G \equiv \sum_{i=1}^n w^{(i)} (w^{(i)})^T. \tag{12}$$

Lemma 2. *The state vector has posterior variance*

$$\mathbb{V}(\theta \mid \mathcal{S}) = \left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right)^{-1}. \tag{13}$$

The Gram matrix (12) is symmetric and positive semi-definite. So, by the spectral theorem, there is a $K \times K$ diagonal matrix

$$\Delta \equiv \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_K \end{bmatrix}$$

with entries $\delta_1 \geq \cdots \geq \delta_K \geq 0$ and a $K \times K$ orthogonal matrix

$$\Omega = \begin{bmatrix} \omega_1 & \cdots & \omega_K \end{bmatrix}$$

such that

$$\begin{aligned} G &= \Omega \Delta \Omega^T \\ &= \sum_{k=1}^{K} \delta_k \omega_k \omega_k^T. \end{aligned} \tag{14}$$

Then $\delta_1, \ldots, \delta_K$ are the eigenvalues of G and $\omega_1, \ldots, \omega_K \in \mathbb{R}^K$ are the corresponding unit eigenvectors. Proposition 1 uses the eigendecompositions (8) and (14) of the prior variance and Gram matrices to provide sharp bounds on $\pi(\mathcal{S})$.

Proposition 1. *The value $\pi(\mathcal{S})$ of $\mathcal{S}$ satisfies*

$$\frac{1}{K}\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2}\right)^{-1}\right) \overset{\star}{\leq} \pi(\mathcal{S}) \overset{\star\star}{\leq} \frac{1}{K}\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1}\right), \tag{15}$$

where $\star$ holds with equality if $\omega_k = v_{K-k+1}$ for each $k \in \{1, \ldots, K\}$ and $\star\star$ holds with equality if $\omega_k = v_k$ for each $k \in \{1, \ldots, K\}$.

Proposition 1 says that the sample $\mathcal{S}$ is most valuable when the eigenvectors of Σ and G are maximally "aligned": when $v_k = \omega_k$ for each $k \in \{1, \ldots, K\}$ and hence $V = \Omega$. Then $\mathcal{S}$ contains more information about components of θ with larger prior variances. In contrast, the sample is *least* valuable when the eigenvectors of Σ and G are maximally "*mis*-aligned": when $v_k = \omega_{K-k+1}$ for each $k \in \{1, \ldots, K\}$. Then $\mathcal{S}$ contains *less* information about components of θ with larger prior variances.

For example, suppose $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ contains a single observation. Then the Gram matrix $G = w^{(1)}(w^{(1)})^T$ has eigenvalues $\delta_1 = 1$ and $\delta_2 = \cdots = \delta_K = 0$. Substituting them into (15) gives us bounds on the value of $\mathcal{S}$:

Corollary 1. *Suppose $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ contains one observation. Then its value $\pi(\mathcal{S})$ satisfies*

$$\frac{\lambda_K^2}{K(\lambda_K + \sigma_u^2)} \overset{\star}{\leq} \pi(\mathcal{S}) \overset{\star\star}{\leq} \frac{\lambda_1^2}{K(\lambda_1 + \sigma_u^2)}, \tag{16}$$

where $\star$ holds with equality if $\Sigma w^{(1)} = \lambda_K w^{(1)}$ and $\star\star$ holds with equality if $\Sigma w^{(1)} = \lambda_1 w^{(1)}$.

The value of a sample with size $n = 1$ is largest when $w^{(1)}$ is an eigenvector of Σ with corresponding eigenvalue $\lambda_1 = \max\{\lambda_1, \ldots, \lambda_K\}$. The value is smallest when $w^{(1)}$ is an eigenvector of Σ with corresponding eigenvalue $\lambda_K = \min\{\lambda_1, \ldots, \lambda_K\}$. Intuitively, the more "weight" $w^{(1)}$ puts on directions in which the prior variance of θ is large, the more valuable it is to observe $(w^{(1)}, y^{(1)})$ because the larger is the variance reduction it delivers.

4.2 Optimal samples

Suppose the eigenvectors of Σ and G are maximally aligned (and hence $V = \Omega$). Then, by Proposition 1, the value $\pi(\mathcal{S})$ of $\mathcal{S}$ rises when the trace

$$\operatorname{tr}(\mathbb{V}(\theta \mid \mathcal{S})) = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1}$$

of the posterior variance matrix falls. This trace depends on the eigenvalues $\delta_1, \ldots, \delta_K$ of the Gram matrix G, which are non-negative, non-increasing, and sum to n.[13] So $\pi(\mathcal{S})$ is maximized when $\delta_1, \ldots, \delta_K$ solve

$$\min_{\delta_1, \ldots, \delta_K \in \mathbb{R}} \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1} \text{ subject to } \delta_1 \geq \ldots \geq \delta_K \geq 0 \text{ and } \sum_{k=1}^{K} \delta_k = n. \tag{17}$$

Proposition 2 describes a solution to (17). It uses the integer

$$R^* \equiv \max\left\{ k \in \{1, \ldots, K\} : \sum_{j=1}^{k} \frac{1}{\lambda_j} + \frac{n}{\sigma_u^2} \geq \frac{k}{\lambda_k} \right\} \tag{18}$$

[13] Indeed

$$\sum_{k=1}^{K} \delta_k = \operatorname{tr}(G) = \operatorname{tr}\left(\sum_{i=1}^{n} w^{(i)} (w^{(1)})^T \right) \overset{\star}{=} \sum_{i=1}^{n} \operatorname{tr}\Big((w^{(i)})^T w^{(i)} \Big) \overset{\star\star}{=} n,$$

where $\star$ uses the linearity and cyclic property of matrix traces, and $\star\star$ uses the fact that $\|w^{(i)}\| = 1$ for each i.

to provide a sharp upper bound

$$\pi^* \equiv \frac{1}{K}\left(\sum_{k=1}^{R^*}\lambda_k - (R^*)^2\left(\sum_{k=1}^{R^*}\frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \tag{19}$$

on the value of $\mathcal{S}$.

Proposition 2. *Define*

$$\delta_k^* \equiv \begin{cases} \frac{n}{R^*} + \sigma_u^2\left(\frac{1}{R^*}\sum_{j=1}^{R^*}\frac{1}{\lambda_j} - \frac{1}{\lambda_k}\right) & \text{if } k \leq R^* \\ 0 & \text{if } k > R^* \end{cases} \tag{20}$$

for each $k \in \{1, \ldots, K\}$. Then $\delta_1^*, \ldots, \delta_K^*$ solve (17)*. Moreover, we have $\pi(\mathcal{S}) \leq \pi^*$ with equality if $\mathcal{S}$ induces the Gram matrix*

$$G = \sum_{k=1}^{K}\delta_k^* v_k v_k^T. \tag{21}$$

We call the sample $\mathcal{S}$ "optimal" if it induces the Gram matrix (21). The agent can construct such a sample as follows: for each $k \in \{1, \ldots, K\}$, collect δ_k^* observations with covariate v_k.[14] Then the outcomes $y^{(1)}, \ldots, y^{(n)}$ are pure signals of the coefficients $\gamma_1, \ldots, \gamma_K$. For example, suppose $w^{(1)} = v_1$. Then, since $v_1, \ldots, v_K$ are orthonormal, we have

$$\begin{aligned} y^{(1)} &= \theta^T w^{(1)} + u^{(1)} \\ &= \left(\sum_{k=1}^{K}\gamma_k v_k\right)^T v_1 + u^{(1)} \\ &= \gamma_1 + u^{(1)} \end{aligned}$$

An optimal sample contains pure signals of the coefficients $\gamma_1, \ldots, \gamma_{R^*}$ that contribute most to the states' prior variances. In contrast, it provides no information about the coefficients $\gamma_{R^*+1}, \ldots, \gamma_K$ that contribute least to the states' prior variances. Thus, the agent optimally focuses on the coefficients that "matter" and ignores those that do not. The number R^* that "matter" grows as the sample size n grows. We call R^* the "rank" of an optimal sample because it is the rank of the Gram matrix (21).

[14] This may be infeasible for two reasons: (i) the eigenvalues $\delta_1^*, \ldots, \delta_K^*$ may not be integers; (ii) the agent may not be able to choose $v_1, \ldots, v_K$ as covariates (since, e.g., it would require him to combine negative quantities of fertilizers). We abstract from these issues for convenience and expositional clarity.

If $\mathcal{S}$ is optimal, then the posterior variance matrix $\mathbb{V}(\theta \mid \mathcal{S})$ has k^{th} largest eigenvalue

$$\left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2}\right)^{-1} = \begin{cases} R^*\left(\sum_{j=1}^{R^*} \frac{1}{\lambda_j} + \frac{n}{\sigma_u^2}\right)^{-1} & \text{if } k \leq R^* \\ \lambda_k & \text{if } k > R^*. \end{cases}$$

This eigenvalue equals the posterior variance of the unknown coefficient γ_k. So if $\mathcal{S}$ is optimal, then it equates the posterior variances of $\gamma_1, \ldots, \gamma_{R^*}$ to each other and the posterior variances of $\gamma_{R^*+1}, \ldots, \gamma_K$ to their prior variances. Intuitively, the agent has a target variance and designs $\mathcal{S}$ so as to bring the posterior variances of $\gamma_1, \ldots, \gamma_K$ below that target.

4.3 Naïve baseline

Now suppose the agent lacks conceptual knowledge and, consequently, uses the "naïve" prior $\mathbb{P}^{(0)}$ defined in Section 3.2. Then the state vector θ has prior variance $\Sigma^{(0)} = \overline{\lambda} I_K$, a matrix with eigenvalues $\lambda_1^{(0)} = \cdots = \lambda_K^{(0)} = \overline{\lambda}$. So, by analogy to (18) and (19), the naïve agent's optimal sample has rank

$$\begin{aligned} R^{(0)} &\equiv \max\left\{k \in \{1, \ldots, K\} : \sum_{j=1}^{k} \frac{1}{\lambda_j^{(0)}} + \frac{n}{\sigma_u^2} \geq \frac{k}{\lambda_k^{(0)}}\right\} \\ &= K \end{aligned}$$

and value

$$\begin{aligned} \pi^{(0)} &\equiv \frac{1}{K}\left(\sum_{k=1}^{R^{(0)}} \lambda_k^{(0)} - (R^{(0)})^2 \left(\sum_{k=1}^{R^{(0)}} \frac{1}{\lambda_k^{(0)}} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \\ &= \overline{\lambda} - \left(\frac{1}{\overline{\lambda}} + \frac{n}{K\sigma_u^2}\right)^{-1}. \end{aligned}$$

This sample is equally informative about all components of the state vector θ: the optimal Gram matrix (21) has equal (to n/K) eigenvalues. Intuitively, if the agent does not know which concepts have more explanatory power, then he has no reason to prioritize some components of θ over others when collecting data, and so he collects the same amount on every component.

5 Value of conceptual knowledge

Whereas information is valuable insofar as it helps the agent make better decisions, conceptual knowledge is valuable insofar as it helps him obtain better information. We quantify

this idea by comparing the information acquired by two agents:

(i) one who has conceptual knowledge and uses the true prior $\mathbb{P}$;

(ii) one who lacks conceptual knowledge and uses the naïve prior $\mathbb{P}^{(0)}$ (see Section 3.2).

The optimal samples designed by these agents have values π^* and $\pi^{(0)}$. (See Sections 4.1–4.3 for derivations of π^* and $\pi^{(0)}$.) Their difference

$$\Pi \equiv \pi^* - \pi^{(0)}$$

measures how much knowing concepts empowers the agent to collect more valuable information. Accordingly, we call Π the "value of conceptual knowledge." It depends on the eigenvalues $\lambda_1, \ldots, \lambda_K$ of Σ and the sample size n (which jointly determine the rank R^*). We characterize this dependence in Theorems 1 and 2.

Theorem 1. *The value of conceptual knowledge*

(i) is non-negative,

(ii) equals zero when the eigenvalues $\lambda_1, \ldots, \lambda_K$ are equal, and

(iii) does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a mean-preserving spread.

Theorem 1 says that conceptual knowledge is more valuable when states are more reducible. If a few common concepts explain most of states' prior variances, then the agent gains a lot from identifying those concepts via "pre-data PCA" and learning about the corresponding coefficients $\gamma_1, \ldots, \gamma_{R^*}$ (i.e., "asking the right questions"). In contrast, if every concept has the same explanatory power, then the agent gains nothing from identifying those concepts because he designs the same sample as he would if he was naïve.

Theorem 2. *There is a finite threshold $n' \geq 0$ such that the value of conceptual knowledge Π is increasing in the sample size n if and only if $n < n'$. Moreover,*

$$\lim_{n \to \infty} \Pi = 0.$$

Theorem 2 says that the value of conceptual knowledge is non-monotone in the sample size n. This is because raising n has two effects:

(i) it gives the agent more information about the unknown coefficients $\gamma_1, \ldots, \gamma_{R^*}$, raising the gain from knowing which concepts to focus on;

(ii) it leads the agent to learn about more coefficients (i.e., it raises R^*), lowering the gain from knowing which concepts to focus on.

The first effect dominates the second precisely when $n < n'$.

Theorem 2 also says that the value of conceptual knowledge vanishes as n grows without bound. This is because the agent's posterior becomes less dependent on his prior as n grows and is independent in the limit as $n \to \infty$. Intuitively, if the agent has infinite data, then he does not benefit from doing "pre-data PCA" because he can do traditional (post-data) PCA. Having access to unlimited data washes out the benefit of knowing what data to collect.[15]

As an illustration of Theorems 1 and 2, consider the prior variance matrix (11) derived in Section 3.3. Its eigenvalues are equal when $\rho = 0$ and undergo a MPS when $\rho \in [0, 1)$ rises. So, by Theorem 1, the value Π of conceptual knowledge equals zero when $\rho = 0$ and is non-decreasing in ρ. Moreover, by Theorem 2, there is a threshold $n' \geq 0$ such that Π is increasing in the sample size n if and only if $n < n'$. We characterize this threshold below.

Proposition 3. *Suppose the states have equal prior variances $\sigma^2 > 0$ and pairwise correlation $\rho \in [0, 1)$. Then the value Π of conceptual knowledge*

(i) equals zero when $\rho = 0$,

(ii) is increasing in ρ, and

(iii) is increasing in the sample size n if and only if

$$n < \frac{\rho K \sigma_u^2}{(1 + \rho(K-1))\sigma^2}. \tag{22}$$

Whereas Theorem 1 implies Π is non-decreasing in ρ, Proposition 3 says Π is *increasing* in ρ. This is because Theorem 1 refers to an arbitrary MPS, which may not affect the largest R^* eigenvalues or, as a result, the value of an optimal sample. In contrast, the MPS induced by raising ρ always affects the largest eigenvalue $\lambda_1 = (1 + \rho(K-1))\sigma^2$ of (11).

[15] This washout relies on having *unrestricted* access: the agent must be able to choose covariates that span the K-dimensional space containing θ. If the covariates do not span $\mathbb{R}^K$, then $\mathcal{S}$ may contain no information about some high-variance components of θ, the agent's posterior expected loss may be arbitrarily large, and the value of $\mathcal{S}$ may be arbitrarily small. We illustrate this possibility in Davies and Sankar (2026, Section A2.3).

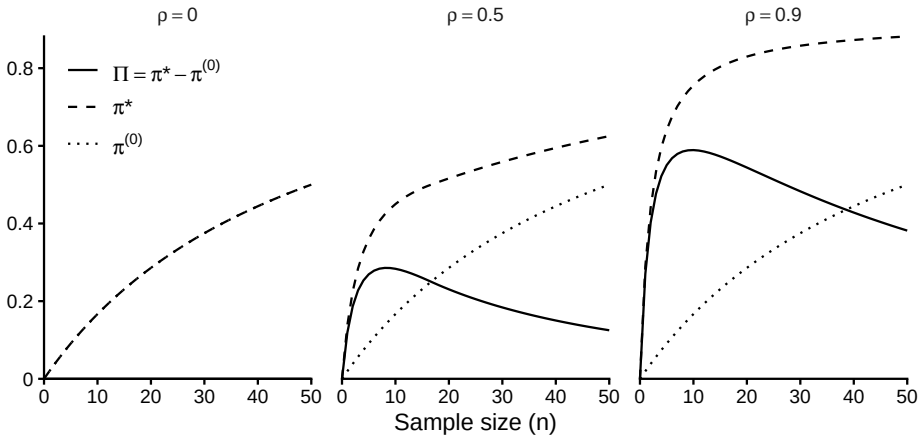

Figure 1: Values Π, π^*, and $\pi^{(0)}$ when θ has prior variance (11) and $(K, \sigma^2, \sigma_u^2) = (5, 1, 10)$

Figure 1 shows how Π depends on ρ and the sample size n. If $\rho = 0$, then the values π^* and $\pi^{(0)}$ of the optimal samples collected by agents with and without conceptual knowledge are equal, and so $\Pi = 0$ for all $n > 0$. In contrast, if $\rho > 0$, then $\pi^* > \pi^{(0)}$ for all $n > 0$ and hence $\Pi > 0$ for all $n > 0$. Both π^* and $\pi^{(0)}$ grow as n grows, and π^* grows faster if and only if (22) holds. Thus $\Pi = \pi^* - \pi^{(0)}$ is increasing in n if and only if (22) holds.

6 Deeper conceptual knowledge

This section extends our notion of conceptual knowledge to one of "deeper" knowledge. We suppose the agent knows some, but not all, of the relevant concepts, and refer to the "depth" of his knowledge as the number he knows. This gives us a language for comparing the values of conceptual and statistical knowledge: would the agent rather know more concepts or have more data?

6.1 J-deep conceptual knowledge

We formalize our notion of "deeper" conceptual knowledge as follows. Suppose the agent knows the trace

$$\operatorname{tr}(\Sigma) = \sum_{k=1}^{K} \lambda_k$$

of the true prior variance matrix Σ and its first $J \in \{0, 1, \ldots, K\}$ eigenpairs (λ_k, v_k), but does not know the last $(K - J)$ eigenpairs. Intuitively, he knows the J concepts with the

most explanatory power, but does not know the $(K - J)$ concepts with the least explanatory power. So he assumes the components of θ orthogonal to $v_1, \ldots, v_J$ have equal prior variances; specifically, he assumes θ has prior variance

$$\Sigma^{(J)} = \sum_{k \leq J} \lambda_k v_k v_k^T + \lambda_K^{(J)} \left(I_K - \sum_{k \leq J} v_k v_k^T \right), \tag{23}$$

where

$$\lambda_K^{(J)} \equiv \frac{1}{K - J} \sum_{k > J} \lambda_k$$

is the mean of the smallest $(K - J)$ eigenvalues of Σ.[16] The matrix (23) has the same trace as Σ but (possibly) different eigenvalues; its k^{th} largest eigenvalue

$$\lambda_k^{(J)} \equiv \begin{cases} \lambda_k & \text{if } k \leq J \\ \lambda_K^{(J)} & \text{if } k > J \end{cases}$$

equals that of Σ if and only if $k \leq J$. The eigenvalues of $\Sigma^{(J)}$ have mean

$$\frac{1}{K} \sum_{k=1}^{K} \lambda_k^{(J)} = \overline{\lambda}$$

independently of J. Likewise $\lambda_K^{(0)} = \overline{\lambda}$ by definition. So $\Sigma^{(0)} = \overline{\lambda} I_K$ is the naïve prior variance matrix discussed in Sections 3.2 and 4.3. The parameter J interpolates between $\Sigma^{(0)}$ and $\Sigma^{(K)} = \Sigma$. It captures the "depth" of the agent's conceptual knowledge: the larger is J, the more concepts he knows and the richer is his mental model of θ.[17]

We say the agent has "J-deep conceptual knowledge" if his prior on θ has variance $\Sigma^{(J)}$. Suppose he has such knowledge and designs an optimal sample. Then, by analogy to (18) and (19), this sample has rank

$$R^{(J)} \equiv \max \left\{ k \in \{1, \ldots, K\} : \sum_{j=1}^{k} \frac{1}{\lambda_j^{(J)}} + \frac{n}{\sigma_u^2} \geq \frac{k}{\lambda_k^{(J)}} \right\}$$

[16] If $J = K$, then the bracketed term in (23) equals zero and $\lambda_K^{(K)}$ can be defined arbitrarily. We set $\lambda_K^{(K)} \equiv \lambda_K$.

[17] Since $\lambda_1 \geq \cdots \geq \lambda_K$ (by assumption), there are non-increasing returns to knowing more concepts (i.e., increasing J): each additional concept contributes a non-increasing share of states' prior variances. Intuitively, the agent prioritizes discovering concepts with more explanatory power (e.g., he reads textbooks that provide "high-level summaries" before "digging into the details").

and value

$$\pi^{(J)} \equiv \frac{1}{K}\left(\sum_{k=1}^{R^{(J)}} \lambda_k^{(J)} - \left(R^{(J)}\right)^2 \left(\sum_{k=1}^{R^{(J)}} \frac{1}{\lambda_k^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-1}\right).$$

For example, letting $J = 0$ yields the rank $R^{(0)}$ and value $\pi^{(0)}$ of an optimal sample collected by a naïve agent (see Section 4.3). We refer to the difference

$$\Pi^{(J)} \equiv \pi^{(J)} - \pi^{(0)}$$

between $\pi^{(J)}$ and $\pi^{(0)}$ as the "value of J-deep conceptual knowledge."

Proposition 4 says that deeper knowledge is (weakly) more valuable. Intuitively, knowing more concepts allows the agent to design samples that provide more payoff-relevant information.

Proposition 4. *The value of J-deep conceptual knowledge*

(i) is non-negative,

(ii) equals zero when $J = 0$,

(iii) is non-decreasing in J, and

(iv) equals the value of full (i.e., K-deep) knowledge when $J \geq R^*$.

The value of J-deep conceptual knowledge is bounded above by the value $\Pi^{(K)} = \Pi$ of "full" knowledge, and attains this bound when $J \geq R^*$. Thus, the agent gains no additional value from knowing more than the R^* concepts with the most explanatory power. This is because he ignores the other $(K - R^*)$ concepts when he designs samples (see Lemma A7), so knowing those concepts does not change his optimal sample.

For example, suppose the true prior variance matrix Σ has k^{th} largest eigenvalue

$$\lambda_k = \frac{K\alpha(1-\alpha)^{k-1}}{1-(1-\alpha)^K}$$

with $0 < \alpha < 1$. Then $\lambda_1, \ldots, \lambda_K$ are strictly positive, have mean $\overline{\lambda} = 1$, are constant in the limit as $\alpha \to 0$, and get more spread out as α rises.[18] This parameter determines the

[18] For each $k \in \{1, \ldots, K\}$ we have $\lambda_k \to 1$ as $\alpha \to 0$ by L'Hôpital's rule. Moreover, the partial sum

$$\sum_{j=1}^{k} \lambda_k = \frac{K\left(1-(1-\alpha)^k\right)}{1-(1-\alpha)^K}$$

is non-decreasing in α and is constant in α when $k = K$. So, by Lemma A1, the eigenvalues $\lambda_1, \ldots, \lambda_K$ undergo a MPS when α rises.

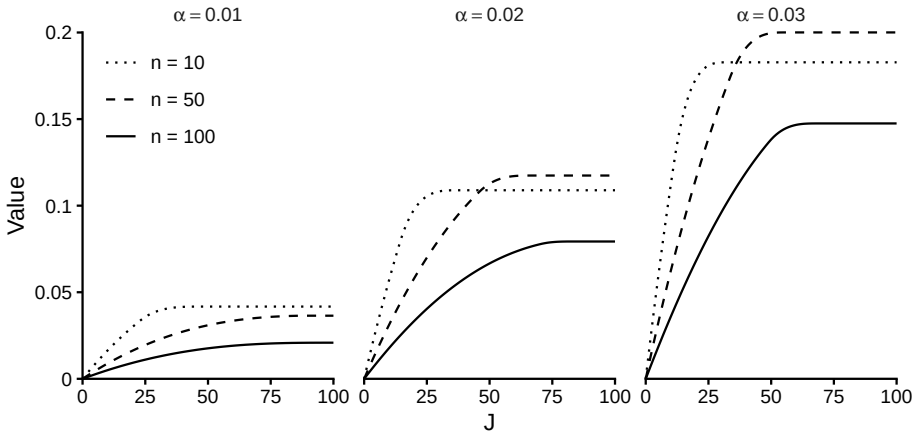

Figure 2: Value $\Pi^{(J)}$ when $\lambda_{k+1} = (1-\alpha)\lambda_k$ and $(K, \sigma_u^2, \overline{\lambda}) = (100, 1, 1)$

rate

$$\frac{\lambda_{k+1} - \lambda_k}{\lambda_k} = -\alpha$$

at which λ_k decays as k grows. Intuitively, the larger is α, the faster concepts' marginal explanatory power falls. Thus, if α is larger, then states are more reducible.

Figure 2 shows how $\Pi^{(J)}$ depends on J and α when $(K, \sigma_u^2) = (100, 1, 1)$. It is increasing in J when $J < R^*$ and constant in J when $J \geq R^*$. It is increasing in α, consistent with Theorem 1: conceptual knowledge is more valuable when states are more reducible. Likewise $\Pi^{(J)}$ is non-monotone in n, consistent with Theorem 2: raising n allows the agent to learn more about the "in-sample" coefficients $\gamma_1, \ldots, \gamma_{R^{(J)}}$ (raising $\Pi^{(J)}$), but also prompts him to expand his sample and learn about more coefficients (lowering $\Pi^{(J)}$).

6.2 More concepts or more data?

Finally, we compare the marginal values of knowing more concepts (i.e., increasing the depth J) and having more data (i.e., increasing the sample size n).

Suppose the agent has J-deep conceptual knowledge and designs an optimal sample of size n. The value $\pi^{(J)}$ of this sample is larger when his minimized posterior expected loss (4) is smaller. Suppose he has a target value $\pi_0 \geq 0$ and let

$$n_{\pi_0}^{(J)} \equiv \min\{n \geq 0 : \pi^{(J)} \geq \pi_0\}$$

be the minimum sample size necessary to attain this value. This size is smaller when the agent knows more concepts and when states are more reducible:

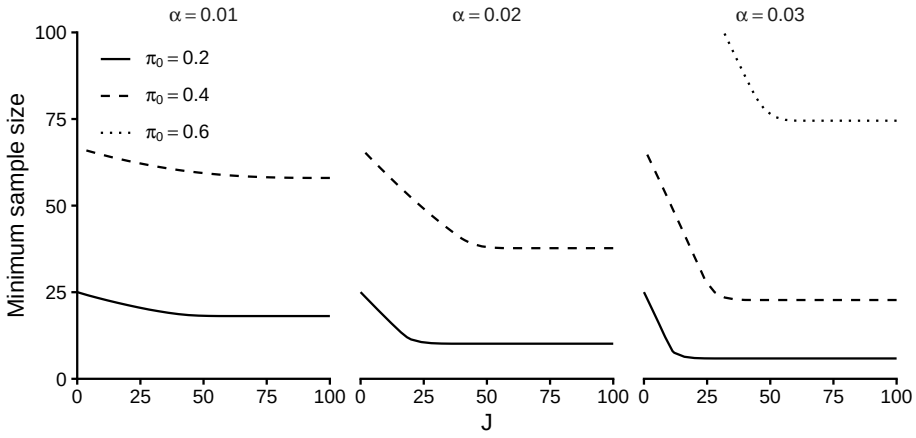

Figure 3: Minimum sample sizes $n_{\pi_0}^{(J)}$ when $\lambda_{k+1} = (1-\alpha)\lambda_k$ and $(K, \sigma_u^2, \overline{\lambda}) = (100, 1, 1)$

Theorem 3. *The minimum sample size $n_{\pi_0}^{(J)}$ necessary to design a sample with value π_0*

(i) is non-increasing in the depth J of the agent's conceptual knowledge and

(ii) does not rise when the eigenvalues $\lambda_1, \ldots, \lambda_K$ undergo a mean-preserving spread.

Theorem 3 says that if the agent knows more concepts, then he can attain the same payoffs with less data, especially when states are highly reducible. This is because he can design better samples and extract more value from each observation, lowering the number he needs to attain the target π_0.

For example, suppose the true eigenvalues $\lambda_1, \ldots, \lambda_K$ have mean $\overline{\lambda} = 1$ and decay at rate α as in Section 6.1. Figure 3 shows how the minimum sample size $n_{\pi_0}^{(J)}$ depends on the target value π_0 and depth J when $(K, \sigma_u^2) = (100, 1)$. Given π_0, the size $n_{\pi_0}^{(J)}$ is decreasing in J when $J < R^*$ and constant in J when $J \geq R^*$. Intuitively, if the agent knows too few concepts, then he cannot design samples that focus on all of the "right" concepts. Giving him more concepts empowers him to design better samples, extract more value from each observation, and require fewer observations to attain π_0. However, once he knows enough concepts, giving him more does not change how he designs samples or the marginal value of each observation. Then the only way to obtain *more valuable* data is to obtain *more* data, thus making $n_{\pi_0}^{(J)}$ constant in $J \geq R^*$.

The curves in Figure 3 are indifference curves: they trace of sets of depth-size pairs (J, n) that allow the agent to attain different value targets π_0. The slope of each curve equals the marginal rate of substitution (hereafter "MRS") between concepts and data. Intuitively, this

MRS captures the number of observations the agent would give up to know another concept. When J is large, the MRS equals zero because the learning another concept does not change the agent's optimal sample (see Section 6.1). However, when J is small, Figure 3 shows that the MRS rises in absolute value when α rises: the marginal concept is "worth more observations" when states are more reducible.

7 Discussion

7.1 Modeling assumptions

We assume states and outcomes are jointly normally distributed under the agent's prior, and that his actions are real-valued and induce quadratic losses. This setup is ubiquitous in the literature on statistical decisions (Hastie et al., 2009). It is also implicit in empirical economics papers that estimate linear models via Ordinary Least Squares. Our agent has a linear model (7) of the unknown state vector. If his prior is uninformative, then his optimal actions equal the estimates obtained via OLS.

We also assume the agent knows how states covary *a priori*. This separates the conceptual knowledge embedded in his prior from the statistical knowledge he infers from his sample. The assumption allows us to measure the agent's conceptual and statistical knowledge on independent scales, and to study their relative contributions to his payoffs.

Finally, we assume there is a correct model of the agent's environment (i.e., a true prior variance matrix) that he can know at different "depths." This separates our paper from the literatures on model uncertainty (Chatfield, 1995; Marinacci, 2015) and mis-specification (Esponda and Pouzo, 2016; Spiegler, 2016), which study agents who do not know the correct model or use an incorrect model. Our analysis complements those literatures: rather than asking "what if the agent does not know the correct model," we ask "what does he gain from knowing the correct model?"

7.2 Related literatures

Value of information. Seminal work by Blackwell (1951, 1953) establishes an order over information sources: one is more informative than another if it produces a sufficient statistic for the other's information. Howard (1966) and Raiffa and Schlaifer (1961) link informativeness to instrumental value, defining the "value of information" as the gain in

expected payoffs it delivers. More recently, Brooks et al. (2024) extend Blackwell's order to settings with multiple sources, Frankel and Kamenica (2019) characterize measures of information's value, and Whitmeyer (2026) studies value-increasing transformations in abstract decision problems.[19]

Our contribution is to illustrate *why* some information is more valuable than others, focusing on a specific decision problem relevant to economists and statisticians. We show that the value of a sample depends on its alignment with the conceptual structure encoded by the prior variance matrix. Proposition 1 provides sharp bounds: a sample is most (least) valuable when the eigenvectors of the induced Gram and prior variance matrices are maximally aligned (mis-aligned). Proposition 2 characterizes the design of maximally valuable samples, showing that they focus on dimensions with the most prior variance. Theorem 1 quantifies how much this focus raises samples' value. Theorem 2 shows that the gain is non-monotonic in sample size—a result made possible by our modeling assumptions (i.e., Gaussian priors and quadratic losses) and consequent tractability.

Model-based inference. Economists (e.g., Koopmans, 1947; Lucas, 1976; Wolpin, 2013), statisticians (e.g., Cox, 1990), and computer scientists (e.g., Wolpert, 1996) emphasize the importance of models for interpreting data. Manski (2003) shows formally that data alone are insufficient for inference: one must also make assumptions about the data-generating process (i.e., impose a model).

In our framework, a model specifies which components of the states vary together (the eigenvectors $v_1, \ldots, v_K$) and how much variance each component contributes (the eigenvalues $\lambda_1, \ldots, \lambda_K$). The depth parameter J interpolates between a diffuse model ($J = 0$), a partial model ($J < K$), and a complete model ($J = K$). Thus J is an alternative to Fudenberg et al.'s (2022) measure of model "completeness." Theorem 3 characterizes the trade-off between having a more complete model and having more data. This complements work by Dominitz and Manski (2017), who study the trade-off between having more data and "better" data. We also rationalize Mailath and Samuelson's (2020, p. 1463) claim that "people work with models that are deliberately incomplete, including the most salient variables and excluding others." Our agent optimally focuses on high-variance dimensions (Proposition 2) and may not benefit from knowing about more than a few (Proposition 4).

[19] Others study information's value in specific decision problems. For example, Lehmann (1988) and Athey and Levin (2018) study its value in monotone decision problems, Persico (2000) studies its value in auctions, and Cabrales et al. (2013) study its value in investment decisions.

Human cognition. Murphy (2002) explains how humans build mental models from concepts, while Tenenbaum et al. (2011) and Mitchell (2021) discuss how concepts help humans generalize. We embed these ideas in a Bayesian decision framework. Our agent has a mental model of states as combinations of concepts. This allows him to generalize: he can use signals of one state to make inferences about another.

Ilut and Valchev (2025) offer a different framework. They isolate two learning modes—"abstract reasoning" and "integrating experience"—analogous to our notions of conceptual and statistical knowledge. Ilut and Valchev study a dynamic setting, and focus on the "learning traps" that arise from reasoning too little or having the wrong data. In contrast, we study a static setting, and focus on the benefits of reasoning correctly and having the "right" data.

Whereas humans use concepts to learn from limited data (Tenenbaum et al., 2011), machines rely on recognizing patterns in large sets of data (Goodfellow et al., 2016; Halevy et al., 2009). Our framework formalizes this distinction: conceptual knowledge compensates for data scarcity (Theorem 3) but loses its value when data are abundant (Theorem 2). Iakovlev and Liang (2025) obtain a similar washout result in a different framework. Ours offers additional insight: the value of conceptual knowledge rises before it falls, changing direction at the threshold n' identified in Theorem 2.

8 Conclusion

This paper defines the "value of conceptual knowledge," and quantifies it in a Bayesian decision framework with Gaussian priors and quadratic losses. We show that conceptual knowledge aids information acquisition in a static, single-agent setting. One could extend our analysis to a dynamic and/or multi-agent setting. This would support a theory of how concepts are discovered and facilitate social learning. One could also use our framework to develop a consumer choice theory of concepts and data—we study their relative values, but not their relative costs.

A Proofs

Many of our proofs invoke the following lemma, which connects mean-preserving spreads to majorization.

Lemma A1. *Let $\lambda_k > 0$ and $\lambda'_k > 0$ be non-increasing in $k \in \{1, \ldots, K\}$, and let F and F' be their CDFs defined as in* (9)*. The following are equivalent:*

(i) *F' is a mean-preserving spread of F.*

(ii) *$\sum_{k=1}^{K} g(\lambda'_k) \geq \sum_{k=1}^{K} g(\lambda_k)$ for all convex functions $g : (0, \infty) \to \mathbb{R}$.*

(iii) *$\sum_{j=1}^{k} \lambda'_j \geq \sum_{j=1}^{k} \lambda_j$ for each $k \in \{1, \ldots, K\}$, with equality when $k = K$.*

(iv) *$\sum_{j=k}^{K} \lambda'_j \leq \sum_{j=k}^{K} \lambda_j$ for each $k \in \{1, \ldots, K\}$, with equality when $k = 1$.*

Proof. The result follows from establishing three equivalences:

1. *(i) $\iff$ (ii).* Rothschild and Stiglitz (1970, Theorem 2) show that (i) is equivalent to

   (ii') $\int_0^\infty g(z)\,\mathrm{d}F'(z) \geq \int_0^\infty g(z)\,\mathrm{d}F(z)$ for all convex functions $g : (0, \infty) \to \mathbb{R}$,

   which is equivalent to (ii) by the definitions of F and F'.

2. *(ii) $\iff$ (iii).* Consider the K-vectors $\lambda' \equiv (\lambda'_1, \ldots, \lambda'_K)$ and $\lambda \equiv (\lambda_1, \ldots, \lambda_K)$. Arnold (1987, Theorem 2.9) shows that (ii) holds precisely when λ' majorizes λ. But the components of λ' and λ are non-increasing, and so λ' majorizes λ if and only if (iii) holds.

3. *(iii) $\iff$ (iv).* For each $k \in \{1, \ldots, K\}$ we have

$$\begin{aligned}\sum_{j=1}^{k} \lambda'_j - \sum_{j=1}^{k} \lambda_j &= \left(\sum_{j=1}^{K} \lambda'_j - \sum_{j>k} \lambda'_j\right) - \left(\sum_{j=1}^{K} \lambda_j - \sum_{j>k} \lambda_j\right) \\ &= \left(\sum_{j=1}^{K} \lambda'_j - \sum_{j=1}^{K} \lambda_j\right) - \left(\sum_{j>k} \lambda'_j - \sum_{j>k} \lambda_j\right),\end{aligned}$$

   from which it follows that (iii) and (iv) are equivalent. □

A.1 Proof of Lemma 1

We have

$$\begin{aligned}\mathbb{E}[L(\theta, a') \mid \mathcal{S}] &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big[(\theta_k - a'_k)^2 \mid \mathcal{S}\big] \\ &= \frac{1}{K}\sum_{k=1}^{K}\Big((\mathbb{E}[\theta_k \mid \mathcal{S}] - a'_k)^2 + \mathbb{V}(\theta_k \mid \mathcal{S})\Big)\end{aligned}$$

for all $a' \in \mathbb{R}^K$. So $\mathbb{E}[L(\theta, a') \mid \mathcal{S}]$ attains its minimum value

$$\min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{V}(\theta_k \mid \mathcal{S})$$

when $a'_k = \mathbb{E}[\theta_k \mid \mathcal{S}]$ for each $k \in \{1, \ldots, K\}$. □

A.2 Proof of Lemma 2

Our proof of Lemma 2 uses a well-known property of normally distributed random variables.

Lemma A2. *Let $n_1 \geq 1$ and $n_2 \geq 1$ be integers, and let $z \in \mathbb{R}^{n_1+n_2}$ be normally distributed with mean μ and variance Σ. Partition $z = (z_1, z_2)$ into vectors $z_1 \in \mathbb{R}^{n_1}$ and $z_2 \in \mathbb{R}^{n_2}$, and let $\mu = (\mu_1, \mu_2)$ and*

$$\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

be the corresponding partitions of μ and Σ. If Σ_{22} is invertible, then

$$z_1 \mid z_2 \sim \mathcal{N}\big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big).$$

Proof. See Bishop (2006, p. 87) or DeGroot (2004, p. 55).

Proof of Lemma 2. Let $y \equiv (y^{(1)}, \ldots, y^{(n)})$ and $u \equiv (u^{(1)}, \ldots, u^{(n)})$ be the n-vectors of outcomes and errors, and let

$$W \equiv \begin{bmatrix}w^{(1)} & \cdots & w^{(n)}\end{bmatrix}^T$$

be the $n \times K$ design matrix. Then we can write (2) in vector form as

$$y \equiv W\theta + u.$$

Consider the concatenation of θ and y. It is normally distributed with variance

$$\mathbb{V}\left(\begin{bmatrix}\theta\\ y\end{bmatrix} \mid W\right) = \begin{bmatrix}\Sigma & \Sigma W^T\\ W\Sigma & W\Sigma W^T + \sigma_u^2 I_n\end{bmatrix}$$

under the agent's prior. Since observing $\mathcal{S}$ is equivalent to observing W and y, Lemma A2 implies

$$\begin{aligned}\mathbb{V}(\theta \mid \mathcal{S}) &= \mathbb{V}(\theta \mid W, y)\\ &= \Sigma - \Sigma W^T\left(W\Sigma W^T + \sigma_u^2 I_n\right)^{-1} W\Sigma\\ &= \left(\Sigma^{-1} + \frac{1}{\sigma_u^2}G\right)^{-1}\end{aligned}$$

because $G = W^T W$. □

A.3 Proof of Proposition 1

Our proof of Proposition 1 uses the following fact about sums of real, symmetric matrices.

Lemma A3. *Let $n \geq 1$ be an integer, let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be symmetric matrices with eigenvalues $a_1 \geq \cdots \geq a_n$ and $b_1 \geq \cdots \geq b_n$, and let $C = A + B$ have eigenvalues $c_1 \geq \cdots \geq c_n$. Then*

$$\sum_{j=1}^{k}(a_j + b_{n-j+1}) \leq \sum_{j=1}^{k} c_j \leq \sum_{j=1}^{k}(a_j + b_j)$$

for each $k \in \{1, \ldots, n\}$, with equality when $k = n$.

Proof. See Horn and Johnson (2012, Theorem 4.3.47).

Proof of Proposition 1. Now

$$\pi(\mathcal{S}) = \frac{1}{K}\left(\operatorname{tr}(\Sigma) - \operatorname{tr}\left(\left(\Sigma^{-1} + \frac{1}{\sigma_u^2}G\right)^{-1}\right)\right)$$

by Lemma 2. Moreover, defining $Z \equiv V^T\Omega$ gives

$$\begin{aligned}\left(\Sigma^{-1} + \frac{1}{\sigma_u^2}G\right)^{-1} &= \left(V\Lambda^{-1}V^T + \frac{1}{\sigma_u^2}VV^T\Omega\Delta\Omega^T VV^T\right)^{-1}\\ &= V\left(\Lambda^{-1} + \frac{1}{\sigma_u^2}Z\Delta Z^T\right)^{-1}V^T\end{aligned}$$

and hence

$$\mathrm{tr}\left(\left(\Sigma^{-1}+\frac{1}{\sigma_u^2}G\right)^{-1}\right)=\mathrm{tr}\left(\left(\Lambda^{-1}+\frac{1}{\sigma_u^2}Z\Delta Z^T\right)^{-1}\right)$$

by the orthogonality of V and the cyclic property of matrix traces. So (15) is equivalent to

$$\sum_{k=1}^{K}\left(\frac{1}{\lambda_k}+\frac{\delta_k}{\sigma_u^2}\right)^{-1}\overset{\star\star}{\leq}\mathrm{tr}\left(\left(\Lambda^{-1}+\frac{1}{\sigma_u^2}Z\Delta Z^T\right)^{-1}\right)\overset{\star}{\leq}\sum_{k=1}^{K}\left(\frac{1}{\lambda_k}+\frac{\delta_{K-k+1}}{\sigma_u^2}\right)^{-1}. \quad \text{(A1)}$$

Now Λ^{-1} is real, symmetric, and positive definite. It has k^{th} largest eigenvalue $a_k \equiv 1/\lambda_{K-k+1} > 0$. Moreover, since Z is orthogonal, the matrix

$$B \equiv \frac{1}{\sigma_u^2}Z\Delta Z^T$$

is real, symmetric, and positive semi-definite. It has k^{th} largest eigenvalue $b_k \equiv \delta_k/\sigma_u^2 \geq 0$. Define $c_k^{**} \equiv a_k + b_{K-k+1} > 0$ and $c_k^{*} \equiv a_k + b_k > 0$ for each $k \in \{1,\ldots,K\}$, and consider the matrix $C \equiv \Lambda^{-1} + B$ with k^{th} largest eigenvalue c_k. This matrix is positive definite and so $c_k > 0$ for each k. Moreover, by Lemma A3, we have

$$\sum_{j=1}^{k}c_j^{**} \leq \sum_{j=1}^{k}c_j \leq \sum_{j=1}^{k}c_j^{*}$$

for each $k \in \{1,\ldots,K\}$, with equality when $k = K$.

Now define $g(z) \equiv 1/z$ for all $z > 0$. Then $g : (0,\infty) \to \mathbb{R}$ is convex. So, by Lemma A1, we have

$$\sum_{k=1}^{K}\frac{1}{c_k^{**}} \leq \sum_{k=1}^{K}\frac{1}{c_k} \leq \sum_{k=1}^{K}\frac{1}{c_k^{*}}. \quad \text{(A2)}$$

But

$$\sum_{k=1}^{K}\frac{1}{c_k^{**}}=\sum_{k=1}^{K}\left(\frac{1}{\lambda_k}+\frac{\delta_k}{\sigma_u^2}\right)^{-1}$$

and

$$\sum_{k=1}^{K}\frac{1}{c_k^{*}}=\sum_{k=1}^{K}\left(\frac{1}{\lambda_k}+\frac{\delta_{K-k+1}}{\sigma_u^2}\right)^{-1}$$

by the definitions of $c_1^{**},\ldots,c_K^{**}$ and $c_1^{*},\ldots,c_K^{*}$, and

$$\begin{aligned}\sum_{k=1}^{K}\frac{1}{c_k} &= \mathrm{tr}\big(C^{-1}\big)\\ &= \mathrm{tr}\left(\left(\Lambda^{-1}+\frac{1}{\sigma_u^2}Z\Delta Z^T\right)^{-1}\right)\end{aligned}$$

by the definition of C. Substituting these expressions into (A2) yields (A1), from which (15) follows.

It remains to show when the bounds $\star$ and $\star\star$ hold with equality.

Suppose $\omega_k = v_{K-k+1}$ for each $k \in \{1, \ldots, K\}$. Then $Z = V^T\Omega$ is the $K \times K$ anti-diagonal matrix with jk^{th} entry

$$Z_{jk} = \begin{cases} 1 & \text{if } j + k = K + 1 \\ 0 & \text{if } j + k \neq K + 1. \end{cases}$$

So the inverse of

$$\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T = \Lambda^{-1} + \frac{1}{\sigma_u^2} \begin{bmatrix} \delta_K & & \\ & \ddots & \\ & & \delta_1 \end{bmatrix}$$

has trace

$$\operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T \right)^{-1} \right) = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2} \right)^{-1}$$

and thus $\star$ holds with equality.

Now suppose $\omega_k = v_k$ for each $k \in \{1, \ldots, K\}$. Then Z is the $K \times K$ identity matrix. So the inverse of

$$\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T = \Lambda^{-1} + \frac{1}{\sigma_u^2} \Delta$$

has trace

$$\operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T \right)^{-1} \right) = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1}$$

and thus $\star\star$ holds with equality. □

A.4 Proof of Corollary 1

If $n = 1$, then the Gram matrix has eigenvalues $\delta_1 = 1$ and $\delta_2 = \cdots = \delta_K = 0$. So

$$\begin{aligned} \sum_{k=1}^{K} \left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2} \right)^{-1} \right) &= \sum_{k=1}^{K-1} \left(\lambda_k - \left(\frac{1}{\lambda_k} + 0 \right)^{-1} \right) + \left(\lambda_K - \left(\frac{1}{\lambda_K} + \frac{1}{\sigma_u^2} \right)^{-1} \right) \\ &= 0 + \frac{\lambda_K^2}{\lambda_K + \sigma_u^2} \end{aligned}$$

and

$$\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1}\right) = \left(\lambda_1 - \left(\frac{1}{\lambda_1} + \frac{1}{\sigma_u^2}\right)^{-1}\right) + \sum_{k=2}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + 0\right)^{-1}\right)$$
$$= \frac{\lambda_1^2}{\lambda_1 + \sigma_u^2} + 0.$$

Substituting these expressions into (15) yields (16). Moreover, Proposition 1 implies that $\star$ holds when $w^{(1)}$ is an eigenvector of Σ with corresponding eigenvalue λ_K, while $\star\star$ holds when it is an eigenvector with corresponding eigenvalue λ_1. □

A.5 Proof of Proposition 2

Consider the constrained minimization problem (17). We can ignore the constraint that δ_k is non-increasing in k because it does not bind (see below). So the problem has Lagrangian

$$\mathcal{L} \equiv \sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1} - \sum_{k=1}^{K}\eta_k\delta_k - \eta\left(n - \sum_{k=1}^{K}\delta_k\right),$$

where $\eta_k \geq 0$ is the Lagrange multiplier on the non-negativity constraint $\delta_k \geq 0$ and $\eta > 0$ is the multiplier on the sum constraint. Now

$$\frac{\partial^2\mathcal{L}}{\partial\delta_j\,\partial\delta_k} = \begin{cases} \frac{2}{\sigma_u^4}\left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-3} & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

for each pair $(j,k) \in \{1,\ldots,K\}^2$, from which it follows that $\mathcal{L}$ is convex in the vector $(\delta_1,\ldots,\delta_K)$ whenever it has non-negative components. So if $\delta_1^\dagger,\ldots,\delta_K^\dagger$ solve (17), then they satisfy the first-order conditions (FOCs)

$$0 = \frac{\partial\mathcal{L}}{\partial\delta_k}$$
$$= -\frac{1}{\sigma_u^2}\left(\frac{1}{\lambda_k} + \frac{\delta_k^\dagger}{\sigma_u^2}\right)^{-2} - \eta_k + \eta,$$

complementary slackness conditions $0 = \eta_k\delta_k^\dagger$, and sum constraint $\delta_1^\dagger + \cdots + \delta_K^\dagger = n$.

Suppose the non-negativity constraint on δ_k binds. Then the FOCs and complementary slackness conditions imply

$$0 < \eta_k$$
$$= \eta - \frac{\lambda_k^2}{\sigma_u^2},$$

which holds if and only if $\lambda_k < \sigma_u\sqrt{\eta}$. But η is strictly positive and λ_k is non-increasing in k. So there is an integer $k_0 \in \{1, \ldots, K\}$ such that $\delta_k^\dagger > 0$ if and only if $k \leq k_0$.

Suppose $k \leq k_0$. Then $\eta_k = 0$ and so the FOCs imply

$$\frac{\sigma_u^2}{\sqrt{\eta}} = \frac{\sigma_u^2}{\lambda_k} + \delta_k^\dagger.$$

The left-hand side is constant in k, from which it follows that

$$\frac{\sigma_u^2}{\lambda_1} + \delta_1^\dagger = \frac{\sigma_u^2}{\lambda_k} + \delta_k^\dagger$$

and therefore

$$\delta_k^\dagger = \delta_1^\dagger + \sigma_u^2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right).$$

Then the sum constraint implies

$$\begin{aligned} n &= \sum_{k=1}^{k_0}\left(\delta_1^\dagger + \sigma_u^2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right)\right) \\ &= k_0\delta_1^\dagger + \sigma_u^2\sum_{k=1}^{k_0}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right). \end{aligned}$$

Thus

$$\begin{aligned} \delta_k^\dagger &= \frac{1}{k_0}\left(n - \sigma_u^2\sum_{j=1}^{k_0}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_j}\right)\right) + \sigma_u^2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right) \\ &= \frac{n}{k_0} + \sigma_u^2\left(\frac{1}{k_0}\sum_{j=1}^{k_0}\frac{1}{\lambda_j} - \frac{1}{\lambda_k}\right) \end{aligned} \tag{A3}$$

for each $k \leq k_0$ and $\delta_k^\dagger = 0$ for each $k > k_0$. Then the candidate solution $\delta_1^\dagger, \ldots, \delta_K^\dagger$ yields minimized objective

$$\begin{aligned} \sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k^\dagger}{\sigma_u^2}\right)^{-1} &= \sum_{k=1}^{k_0}\left(\frac{1}{\lambda_k} + \frac{1}{\sigma_u^2}\left(\frac{n}{k_0} + \sigma_u^2\left(\frac{1}{k_0}\sum_{j=1}^{k_0}\frac{1}{\lambda_j} - \frac{1}{\lambda_k}\right)\right)\right)^{-1} + \sum_{k>k_0}\left(\frac{1}{\lambda_k} + 0\right)^{-1} \\ &= k_0^2\left(\sum_{j=1}^{k_0}\frac{1}{\lambda_j} + \frac{n}{\sigma_u^2}\right)^{-1} + \sum_{k>k_0}\lambda_k. \end{aligned} \tag{A4}$$

The right-hand side (hereafter "RHS") depends on k_0, which determines $\delta_1^\dagger, \ldots, \delta_K^\dagger$ via (A3). We show that the agent cannot do better than choose $k_0 = R^*$. To see why, define

$$g(k) \equiv \sum_{j=1}^{k}\frac{1}{\lambda_j} + \frac{n}{\sigma_u^2}$$

and

$$h(k) \equiv \frac{k^2}{g(k)} + \sum_{j>k} \lambda_j$$

for each $k \in \{1, \ldots, R^*\}$. Then $g(k) = g(k-1) + \lambda_k^{-1} > \lambda_k^{-1}$ and so

$$\begin{aligned} h(k) - h(k-1) &= \frac{k^2}{g(k-1) + \lambda_k^{-1}} - \frac{(k-1)^2}{g(k-1)} - \lambda_k \\ &= -\frac{\big(g(k) - k\lambda_k^{-1}\big)^2}{\big(g(k) - \lambda_k^{-1}\big) g(k) \lambda_k^{-1}} \end{aligned}$$

is non-positive. Thus $h(k)$ is non-increasing in k. But the RHS of (A4) equals $h(k_0)$. So (A4) is smallest at the largest k_0 that induces a feasible solution $\delta_1^\dagger, \ldots, \delta_K^*$ via (A3). Feasibility requires $\delta_k^\dagger \geq 0$ with equality if and only if $k > k_0$. This holds when $k_0 = R^*$ but is violated when $k_0 > R^*$; in the latter case, we have

$$\begin{aligned} \delta_{k_0}^\dagger &= \frac{n}{k_0} + \sigma_u^2 \left(\frac{1}{k_0} \sum_{j=1}^{k_0} \frac{1}{\lambda_j} - \frac{1}{\lambda_{k_0}} \right) \\ &\leq \frac{n}{R^*} + \sigma_u^2 \left(\frac{1}{R^*} \sum_{j=1}^{R^*} \frac{1}{\lambda_j} - \frac{1}{\lambda_{R^*}} \right) \\ &< \frac{n}{R^*} + \sigma_u^2 \left(\frac{1}{R^*} \left(\frac{R^*}{\lambda_{R^*}} - \frac{n}{\sigma_u^2} \right) - \frac{1}{\lambda_{R^*}} \right) \\ &= 0, \end{aligned}$$

where the weak inequality holds because λ_k is non-increasing in k and the strict inequality holds by the definition of R^*. Thus, the largest feasible choice is $k_0 = R^*$. Then $\delta_k^\dagger = \delta_k^*$ for each $k \in \{1, \ldots, K\}$; that is, the eigenvalues $\delta_1^*, \ldots, \delta_K^*$ defined by (20) solve (17). They are non-increasing because $\lambda_1, \ldots, \lambda_K$ are non-increasing. Moreover, Proposition 1 implies

$$\begin{aligned} \pi(\mathcal{S}) &\leq \frac{1}{K} \sum_{k=1}^{K} \left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2} \right)^{-1} \right) \\ &= \frac{1}{K} \left(\sum_{k=1}^{K} \lambda_k - \left((R^*)^2 \left(\sum_{j=1}^{R^*} \frac{1}{\lambda_j} + \frac{n}{\sigma_u^2} \right)^{-1} + \sum_{k>R^*} \lambda_k \right) \right) \\ &= \pi^*, \end{aligned}$$

with equality if (21) holds. □

A.6 Proof of Theorem 1

Our proof of Theorem 1 invokes the following lemma.

Lemma A4. *The value π^* of an optimal sample does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS.*

Proof. Now

$$\begin{aligned} R^* &\in \underset{k_0 \in \{1,\ldots,K\}}{\arg\min} \left(k_0 \left(\frac{1}{k_0} \left(\sum_{k=1}^{k_0} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right) \right)^{-1} + \sum_{k>k_0} \lambda_k \right) \\ &= \underset{k_0 \in \{1,\ldots,K\}}{\arg\max} \left(\sum_{k=1}^{k_0} \lambda_k - k_0^2 \left(\sum_{k=1}^{k_0} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} \right) \end{aligned}$$

from the proof of Proposition 2. So if $\lambda_1, \ldots, \lambda_K$ undergo a MPS, then R^* changes only if doing so makes $\mathcal{S}$ more valuable. So it suffices to show that for fixed R^*, the MPS does not lower the RHS of (19).

Let $\lambda'_1 \geq \cdots \geq \lambda'_K > 0$ be the eigenvalues after the MPS. By Lemma A1, the difference

$$\eta \equiv \sum_{k=1}^{R^*} \lambda'_k - \sum_{k=1}^{R^*} \lambda_k \tag{A5}$$

is non-negative. The MPS raises the first bracketed term on the RHS of (19) by η. So it suffices to show that the MPS lowers the second bracketed term by at most η:

$$\underbrace{(R^*)^2 \left(\sum_{k=1}^{R^*} \frac{1}{\lambda'_k} + \frac{n}{\sigma_u^2} \right)^{-1}}_{S'} - \underbrace{(R^*)^2 \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1}}_{S} \leq \eta. \tag{A6}$$

Consider the first term S' on the LHS of (A6). This term is largest when the harmonic sum

$$H' \equiv \sum_{k=1}^{R^*} \frac{1}{\lambda'_k}$$

is smallest. Defining $\eta_k \equiv \lambda'_k - \lambda_k$ for each $k \in \{1, \ldots, K\}$ gives

$$H' \equiv \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k}$$

and $\eta_1 + \cdots + \eta_{R^*} = \eta$. Lemma A1 implies

$$\sum_{j=1}^{k} \eta_j \geq 0$$

for each $k \in \{1, \ldots, R^*\}$. Thus

$$H' \geq H^* \equiv \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k^*},$$

where $\eta_1^*, \ldots, \eta_{R^*}^*$ solve the constrained minimization problem

$$\begin{aligned} \min_{\eta_1, \ldots, \eta_{R^*} \in \mathbb{R}} & \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k} \\ \text{subject to} \quad & \lambda_k + \eta_k > 0 \text{ for each } k \in \{1, \ldots, R^*\}, \\ & \sum_{j=1}^{k} \eta_j \geq 0 \text{ for each } k \in \{1, \ldots, R^*\}, \\ \text{and} \quad & \sum_{k=1}^{R^*} \eta_k = \eta. \end{aligned} \tag{A7}$$

Setting $\lambda_k' = \lambda_k + \eta_k^*$ for each $k \in \{1, \ldots, R^*\}$ yields the "worst-case" MPS that maximizes the first term S' on the LHS of (A6) given the difference (A5).

The differences $\eta_1^*, \ldots, \eta_{R^*}^*$ that solve (A7) are non-negative. To see why, notice that $\eta_1 < 0$ is infeasible and assume towards a contradiction that $\eta_\ell^* < 0 \leq \min\{\eta_1^*, \ldots, \eta_{\ell-1}^*\}$ for some $\ell > 1$. Then

$$\ell' \equiv \max\{k \in \{1, \ldots, \ell - 1\} : \eta_k^* > 0\}$$

must exist, for otherwise $\eta_1^*, \ldots, \eta_{R^*}^*$ would violate the constraint

$$\sum_{j=1}^{\ell} \eta_k^* \geq 0.$$

Defining

$$\eta_k^\dagger \equiv \begin{cases} \eta_{\ell'}^* + \eta_\ell^* & \text{if } k = \ell' \\ 0 & \text{if } \ell' < k = \ell \\ \eta_k^* & \text{otherwise} \end{cases}$$

gives

$$\sum_{j=1}^{k} \eta_j^\dagger = \begin{cases} \sum_{j=1}^{\ell} \eta_j^* & \text{if } \ell' \le k \le \ell \\ \sum_{j=1}^{k} \eta_j^* & \text{otherwise} \end{cases}$$

for each $k \in \{1, \ldots, R^*\}$, from which it follows that $\eta_1^\dagger, \ldots, \eta_{R^*}^\dagger$ are feasible. But $\lambda_{\ell'} \ge \lambda_\ell$ and $\eta_{\ell'}^* > 0$, and so $\lambda_{\ell'} + \eta_{\ell'}^* > \lambda_\ell > 0$. Thus

$$\frac{1}{\lambda_{\ell'} + \eta_{\ell'}^* + \eta_\ell^*} + \frac{1}{\lambda_\ell} < \frac{1}{\lambda_{\ell'} + \eta_{\ell'}^*} + \frac{1}{\lambda_\ell + \eta_\ell^*}$$

because $g(z) \equiv 1/z$ is a strictly decreasing and convex function of $z > 0$. But then

$$\begin{aligned} \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k^\dagger} &= \sum_{k<\ell'} \frac{1}{\lambda_k + \eta_k^*} + \frac{1}{\lambda_{\ell'} + \eta_{\ell'}^* + \eta_\ell^*} + \frac{1}{\lambda_\ell} + \sum_{k>\ell} \frac{1}{\lambda_k + \eta_k^*} \\ &< \frac{1}{\lambda_{\ell'} + \eta_{\ell'}^*} + \frac{1}{\lambda_\ell + \eta_\ell^*} + \sum_{k \notin \{\ell', \ell\}} \frac{1}{\lambda_k + \eta_k^*} \\ &= \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k^*}, \end{aligned}$$

contradicting the optimality of $\eta_1^*, \ldots, \eta_{R^*}^*$. So they must be non-negative as ℓ cannot exist.

Finally, we use the non-negativity of $\eta_1^*, \ldots, \eta_{R^*}^*$ to establish the upper bound (A6) on $(S' - S)$. Let $k \in \{1, \ldots, R^*\}$ and consider the derivative

$$\frac{\partial S}{\partial \lambda_k} = \left(\frac{R^*}{\lambda_k} \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} \right)^2$$

of S with respect to λ_k. This derivative is non-negative. It is also bounded above by one, since

$$\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \ge \frac{R^*}{\lambda_k}$$

by the definition of R^*. So S is a 1-Lipschitz function of $\lambda_1, \ldots, \lambda_{R^*}$: changing λ_k by η_k changes S by at most $|\eta_k|$. Letting S^* be the value of S that obtains from changing λ_k by η_k^* gives

$$\begin{aligned} S' - S &\overset{\star}{\le} S^* - S \\ &\le |S^* - S| \\ &\overset{\star\star}{\le} \sum_{k=1}^{K} |\eta_k^*|, \end{aligned}$$

where $\star$ uses the maximality of S^* (induced by the minimality of H^*) and $\star\star$ uses the Lipschitz property. But $\eta_1^*, \ldots, \eta_{R^*}^*$ are non-negative and sum to η, from which the bound (A6) follows:

$$\begin{aligned} S' - S &\leq \sum_{k=1}^{K} \eta_k^* \\ &= \eta. \end{aligned}$$

□

Proof of Theorem 1. It suffices to prove (ii) and (iii), which together imply (i). This is because every distribution of $\lambda_1, \ldots, \lambda_K$ is a MPS of the degenerate distribution under which they are equal (to their mean $\overline{\lambda}$).

Consider (ii). If $\lambda_1, \ldots, \lambda_K$ are equal, then $\lambda_k = \overline{\lambda}$ for each $k \in \{1, \ldots, K\}$, and so $R^* = R^{(0)}$ and $\pi^* = \pi^{(0)}$ by definition. Thus $\Pi \equiv \pi^* - \pi^{(0)} = 0$.

Now consider (iii). The value $\pi^{(0)}$ of a naïve agent's optimal sample depends on the eigenvalues $\lambda_1, \ldots, \lambda_K$ via their mean $\overline{\lambda}$ only. It does not change when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. Since π^* does not fall under the MPS (by Lemma A4), neither does $\Pi \equiv \pi^* - \pi^{(0)}$. □

A.7 Proof of Theorem 2

First let $\tau \equiv (n/\sigma_u^2)/(1/\overline{\lambda}) = n\overline{\lambda}/\sigma_u^2$ index the precision of the data relative to the agent's prior. Then, by definition, his optimal sample has rank

$$R^* = \max\left\{ k \in \{1, \ldots, K\} : \overline{\lambda}\left(\frac{k}{\lambda_k} - \sum_{j=1}^{k} \frac{1}{\lambda_j} \right) \leq \tau \right\}$$

and value

$$\pi^* = \frac{\overline{\lambda}}{K}\left(\sum_{k=1}^{R^*} \frac{\lambda_k}{\overline{\lambda}} - (R^*)^2 \left(\sum_{k=1}^{R^*} \frac{\overline{\lambda}}{\lambda_k} + \tau \right)^{-1} \right).$$

Moreover, the optimal sample collected by a naïve agent has rank $R^{(0)} = K$ and value

$$\begin{aligned} \pi^{(0)} &= \overline{\lambda} - \left(\frac{1}{\overline{\lambda}} + \frac{\tau}{K\overline{\lambda}} \right)^{-1} \\ &= \frac{\overline{\lambda}\tau}{K + \tau}. \end{aligned}$$

So the value of conceptual knowledge is

$$\begin{aligned}\Pi &= \pi^* - \pi^{(0)} \\ &= \frac{\overline{\lambda}}{K}\left(\sum_{k=1}^{R^*}\frac{\lambda_k}{\overline{\lambda}} - (R^*)^2\left(\sum_{k=1}^{R^*}\frac{\overline{\lambda}}{\lambda_k} + \tau\right)^{-1} - \frac{K\tau}{K+\tau}\right)\end{aligned}$$

Since τ is proportional to n, Theorem 2 is equivalent to the following result.

Theorem A1. *Define $\tau \equiv n\overline{\lambda}/\sigma_u^2$. There is a finite threshold $\tau' \geq 0$ such that Π is increasing in τ if and only if $\tau < \tau'$. Moreover,*

$$\lim_{\tau\to\infty}\Pi = 0.$$

Proof. Define

$$\tau_k \equiv \overline{\lambda}\left(\frac{k}{\lambda_k} - \sum_{j=1}^{k}\frac{1}{\lambda_j}\right)$$

for each $k \in \{1,\ldots,K\}$. Then $\tau_1 = 0$, and for each $k < K$ the difference

$$\tau_{k+1} - \tau_k = k\overline{\lambda}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)$$

is non-negative because $\overline{\lambda} > 0$ and $\lambda_{k+1} \leq \lambda_k$. So τ_k is non-decreasing in k and hence

$$R^* = \max\{k \in \{1,\ldots,K\} : \tau_k \leq \tau\}$$

is non-decreasing in τ. Now define $\tau_{K+1} \equiv \infty$ and suppose $\tau \in [\tau_k, \tau_{k+1})$ for some $k \in \{1,\ldots,K\}$. Then $R^* = k$ and so

$$\begin{aligned}\Pi &= \Pi_k \\ &\equiv \frac{\overline{\lambda}}{K}\left(\sum_{j=1}^{k}\frac{\lambda_j}{\overline{\lambda}} - k^2\left(\sum_{j=1}^{k}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - \frac{K\tau}{K+\tau}\right).\end{aligned}$$

Each piece Π_k is continuous in τ. Moreover, for each $k < K$ the difference

$$\Pi_{k+1} - \Pi_k = -\frac{\overline{\lambda}}{K}\left((k+1)^2\left(\sum_{j=1}^{k+1}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - k^2\left(\sum_{j=1}^{k}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - \frac{\lambda_{k+1}}{\overline{\lambda}}\right)$$

between consecutive pieces converges to zero as $\tau \to \tau_{k+1}$. It follows that Π is continuous in τ. So to determine whether Π is increasing or decreasing in τ, it suffices to analyze its derivative

$$\frac{\partial \Pi_k}{\partial \tau} = \frac{\overline{\lambda}}{K}\left(k^2\left(\sum_{j=1}^{k}\frac{\overline{\lambda}}{\lambda_j}+\tau\right)^{-2} - \left(\frac{K}{K+\tau}\right)^2\right) \tag{A8}$$

on each piece Π_k.

Consider the final piece

$$\begin{aligned}\Pi_K &= \frac{\overline{\lambda}}{K}\left(\sum_{j=1}^{K}\frac{\lambda_j}{\overline{\lambda}} - K^2\left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j}+\tau\right)^{-1} - \frac{K\tau}{K+\tau}\right)\\ &= K\overline{\lambda}\left(\frac{1}{K+\tau} - \left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j}+\tau\right)^{-1}\right).\end{aligned}$$

If $\lambda_1, \ldots, \lambda_K$ are equal (i.e., if $\lambda_1 = \lambda_K$), then $\lambda_k = \overline{\lambda}$ and $\tau_k = 0$ for each $k \in \{1, \ldots, K\}$, and so

$$\begin{aligned}\Pi\Big|_{\lambda_1=\lambda_K} &= \Pi_K\Big|_{\lambda_1=\lambda_K}\\ &= K\overline{\lambda}\left(\frac{1}{K+\tau} - \left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\overline{\lambda}}+\tau\right)^{-1}\right)\\ &= 0\end{aligned}$$

for all $\tau \geq 0$. On the other hand, if $\lambda_1, \ldots, \lambda_K$ are not equal (i.e., if $\lambda_1 > \lambda_K$), then

$$\begin{aligned}\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j} &> \frac{K\overline{\lambda}}{\frac{1}{K}\sum_{j=1}^{K}\lambda_j}\\ &= K\end{aligned}$$

by Jensen's inequality and the definition of $\overline{\lambda}$, from which it follows that

$$\frac{\partial \Pi_K}{\partial \tau}\Big|_{\lambda_1>\lambda_K} = K\overline{\lambda}\left(\left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j}+\tau\right)^{-2} - \left(\frac{1}{K+\tau}\right)^2\right)$$

is strictly negative. Thus Π is non-increasing in τ whenever $\tau \geq \tau_K$. Moreover,

$$\begin{aligned}\lim_{\tau\to\infty} \Pi &= \lim_{\tau\to\infty} \Pi_K \\ &= K\overline{\lambda} \lim_{\tau\to\infty} \left(\frac{1}{K+\tau} - \left(\sum_{j=1}^{K} \frac{\overline{\lambda}}{\lambda_j} + \tau \right)^{-1} \right) \\ &= 0.\end{aligned}$$

So if $\lambda_1, \ldots, \lambda_K$ are equal, then $\tau_K = 0$ and the result follows from letting $\tau' = 0$.

It remains to show that if $\lambda_1, \ldots, \lambda_K$ are *not* equal, then there exists $\tau' \in (0, \tau_K)$ such that Π is increasing in τ if and only if $\tau < \tau'$.

Suppose $\tau \in [\tau_k, \tau_{k+1})$ for some $k < K$. Then Π is increasing in τ if and only if (A8) exceeds zero, which happens precisely when

$$\begin{aligned}&\tau < \tau_k' \\ &\quad \equiv \frac{K}{K-k} \left(k \left(1 - \frac{\overline{\lambda}}{\lambda_k} \right) + \tau_k \right).\end{aligned}$$

So Π_k is decreasing in $\tau \in [\tau_k, \tau_{k+1})$ if $\tau_k' < \tau_k$, increasing if $\tau_k' \geq \tau_{k+1}$, and increasing-and-then-decreasing if $\tau_k \leq \tau_k' < \tau_{k+1}$. Now $\tau_k' \geq \tau_k$ if and only if

$$\frac{K-k}{\lambda_k} + \sum_{j=1}^{k} \frac{1}{\lambda_j} \leq \frac{K}{\overline{\lambda}},$$

whereas $\tau_k' < \tau_{k+1}$ if and only if

$$\frac{K}{\overline{\lambda}} < \frac{K-(k+1)}{\lambda_{k+1}} + \sum_{j=1}^{k+1} \frac{1}{\lambda_j}.$$

So defining

$$\eta_k \equiv \frac{K-k}{\lambda_k} + \sum_{j=1}^{k} \frac{1}{\lambda_j}$$

for each $k \in \{1, \ldots, K\}$ gives $\tau_k' \in [\tau_k, \tau_{k+1})$ if and only if $K/\overline{\lambda} \in [\eta_k, \eta_{k+1})$. But η_k is non-decreasing in k because $\lambda_{k+1} \leq \lambda_k$ and therefore

$$\begin{aligned}\eta_{k+1} - \eta_k &= (K-k) \left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} \right) \\ &\geq 0.\end{aligned}$$

It follows that $\tau_k' \in [\tau_k, \tau_{k+1})$ for at most one $k < K$. But there is at least one such k when $\lambda_1, \ldots, \lambda_K$ are not equal. To see why, notice that

$$\begin{aligned}\lim_{\tau \to 0} \frac{\partial \Pi}{\partial \tau} &= \lim_{\tau \to 0} \frac{\partial \Pi_1}{\partial \tau} \\ &= \frac{\overline{\lambda}}{K} \lim_{\tau \to 0} \left(\left(\frac{\overline{\lambda}}{\lambda_1} + \tau \right)^{-2} - \left(\frac{K}{K + \tau} \right)^2 \right) \\ &= \frac{\overline{\lambda}}{K} \left(\left(\frac{\lambda_1}{\overline{\lambda}} \right)^2 - 1 \right)\end{aligned}$$

is strictly positive when $\lambda_1 > \overline{\lambda}$, which holds precisely when $\lambda_1, \ldots, \lambda_K$ are not equal, in which case the value Π is decreasing in τ whenever $\tau > \tau_K$. So Π is initially increasing in τ and eventually decreasing in τ, which, by continuity, means its derivative with respect to τ changes sign at least once. Therefore, if $\lambda_1, \ldots, \lambda_K$ are not equal, then there is a unique $k < K$ such that $\tau_k' \in [\tau_k, \tau_{k+1})$. Letting $\tau' = \tau_k' > 0$ completes the proof. □

A.8 Proof of Proposition 3

Define $\tau \equiv n\overline{\lambda}/\sigma_u^2$ as in the proof of Theorem 2. The eigenvalues of (11) have mean $\overline{\lambda} = \sigma^2$. So Proposition 3 is equivalent to the following result.

Proposition A1. *Define $\tau \equiv n\sigma^2/\sigma_u^2$. Suppose the states have equal prior variances $\sigma^2 > 0$ and pairwise correlation $\rho \in [0, 1)$. Then the value of conceptual knowledge*

(i) equals zero when $\rho = 0$,

(ii) is increasing in ρ, and

(iii) is increasing in the precision index τ if and only if

$$\tau < \frac{\rho K}{1 + \rho(K - 1)}.$$

Our proof of Proposition A1 invokes the following lemma.

Lemma A5. *Suppose θ has prior variance* (11) *with $\sigma^2 > 0$ and $\rho \in [0, 1)$.*

(i) There is a threshold $\rho' \in (0, 1)$ such that

$$R^* = \begin{cases} K & \text{if } \rho \leq \rho' \\ 1 & \text{if } \rho > \rho'. \end{cases} \tag{A9}$$

(ii) The value π^* of an optimal sample rises when ρ rises.

Proof. Consider (i). If $\lambda_1 \geq \lambda_2 = \cdots = \lambda_K$, then

$$R^* = \begin{cases} 1 & \text{if } \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} < \frac{1}{\lambda_2} \\ K & \text{otherwise.} \end{cases}$$

Now (11) has eigenvalues $\lambda_1 = (1+\rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$. So $R^* = K$ if and only if

$$\begin{aligned} 0 &\leq \frac{1}{(1+\rho(K-1))\sigma^2} - \frac{1}{(1-\rho)\sigma^2} + \frac{n}{\sigma_u^2} \\ &= \frac{1}{\sigma^2}\left(\frac{1}{1+\rho(K-1)} - \frac{1}{1-\rho} + \tau\right), \end{aligned}$$

where $\tau \equiv n\sigma^2/\sigma_u^2$. The bracketed term on the RHS is continuous and decreasing in ρ, strictly positive when $\rho = 0$, and unbounded below as $\rho \to 1$. So, by the intermediate value theorem, there exists $\rho' \in (0,1)$ such that (A9) holds.

Now consider (ii). Substituting (A9) into (19) gives

$$\begin{aligned} \pi^* &= \frac{1}{K}\begin{cases} \sum_{k=1}^K \lambda_k - K^2\left(\sum_{k=1}^K \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1} & \text{if } \rho \leq \rho' \\ \lambda_1 - \left(\frac{1}{\lambda_1} + \frac{n}{\sigma_u^2}\right)^{-1} & \text{if } \rho > \rho' \end{cases} \\ &= \frac{\sigma^2}{K}\begin{cases} K - K^2\left(\frac{1}{1+\rho(K-1)} + \frac{K-1}{1-\rho} + \tau\right)^{-1} & \text{if } \rho \leq \rho' \\ 1+\rho(K-1) - \left(\frac{1}{1+\rho(K-1)} + \tau\right)^{-1} & \text{if } \rho > \rho', \end{cases} \end{aligned} \tag{A10}$$

which is piecewise increasing in ρ:

$$\begin{aligned} \frac{\partial}{\partial\rho}\left[\pi^*\big|_{\rho\leq\rho'}\right] &= K(K-1)\left(\frac{1}{1+\rho(K-1)} + \frac{K-1}{1-\rho} + \tau\right)^{-2}\left(\frac{1}{(1-\rho)^2} - \frac{1}{(1+\rho(K-1))^2}\right) \\ &\geq 0 \end{aligned}$$

with equality if and only if $\rho = 0$, and

$$\begin{aligned} \frac{\partial}{\partial\rho}\left[\pi^*\big|_{\rho>\rho'}\right] &= \frac{(K-1)\sigma^2}{K}\left(1 + \left(\frac{1}{1+\tau(1+\rho(K-1))}\right)^2\right) \\ &> 0. \end{aligned}$$

□

Proof of Proposition A1. Suppose the sample $\mathcal{S}$ is optimal. Then its value π^* equals

$$\pi^{(0)} = \frac{\sigma^2\tau}{K+\tau}$$

when $\rho = 0$. Now π^* is increasing in ρ (by Lemma A5), whereas $\pi^{(0)}$ is constant in ρ. So $\Pi = \pi^* - \pi^{(0)}$ equals zero when $\rho = 0$ and is increasing in ρ.

It remains to prove (iii). Now (11) has eigenvalues $\lambda_1 = (1+\rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$, which have mean $\overline{\lambda} = \sigma^2$. Defining

$$\begin{aligned}\tau_K &\equiv \overline{\lambda}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \\ &= \frac{\rho K}{(1-\rho)(1+\rho(K-1))}\end{aligned}$$

gives

$$\begin{aligned}R^* &= \begin{cases} 1 & \text{if } \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} < \frac{1}{\lambda_K} \\ K & \text{if } \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} \geq \frac{1}{\lambda_K} \end{cases} \\ &= \begin{cases} 1 & \text{if } \tau < \tau_K \\ K & \text{if } \tau \geq \tau_K, \end{cases}\end{aligned}$$

which when substituted into (19) gives

$$\Pi = \frac{\sigma^2}{K}\begin{cases} 1+\rho(K-1) - \left(\frac{1}{1+\rho(K-1)} + \tau\right)^{-1} - \frac{K\tau}{K+\tau} & \text{if } \tau < \tau_K \\ K^2\left((K+\tau)^{-1} - \left(\sum_{k=1}^K \frac{\sigma^2}{\lambda_k} + \tau\right)^{-1}\right) & \text{if } \tau \geq \tau_K. \end{cases}$$

The first piece is (weakly) concave in τ: differentiating it with respect to τ gives

$$\frac{\partial}{\partial\tau}\left[\Pi\big|_{\tau<\tau_K}\right] = \frac{\sigma^2}{K}\left(\left(\frac{1}{1+\rho(K-1)} + \tau\right)^{-2} - \left(\frac{K}{K+\tau}\right)^2\right),$$

which is strictly positive if and only if

$$\tau < \tau' \equiv \frac{\rho K}{1+\rho(K-1)}.$$

In contrast, our proof of Theorem A1 shows that the second piece (with $\tau \geq \tau_K$) is non-increasing in τ. But $\tau' \leq \tau_K$, from which (iii) follows. □

A.9 Proof of Proposition 4

Our proof of Proposition 4 invokes the following two lemmas.

Lemma A6. *The value $\pi^{(J)}$ of an optimal sample designed by an agent with J-deep conceptual knowledge is*

(i) non-decreasing in J and

(ii) increasing in the sample size n.

Proof. We prove (i) and (ii) separately:

(i) It suffices to show that $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS when J rises. Then (i) follows from an argument similar to that used to prove Lemma A4.

Fix $J < K$. For each $k \in \{1, \ldots, K\}$ we have

$$\lambda_k^{(J+1)} - \lambda_k^{(J)} = \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ \lambda_K^{(J+1)} - \lambda_K^{(J)} & \text{if } k > J+1 \end{cases}$$

and hence

$$\begin{aligned} \sum_{j=1}^{k} \lambda_j^{(J+1)} - \sum_{j=1}^{k} \lambda_j^{(J)} &= \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ \lambda_{J+1} - \lambda_K^{(J)} + (k-(J+1))\left(\lambda_K^{(J+1)} - \lambda_K^{(J)}\right) & \text{if } k > J+1 \end{cases} \\ &= \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ (K-k)\left(\lambda_K^{(J)} - \lambda_K^{(J+1)}\right) & \text{if } k > J+1. \end{cases} \end{aligned} \tag{A11}$$

because $\lambda_{J+1} = (K-J)\lambda_K^{(J)} - (K-(J+1))\lambda_K^{(J+1)}$. We also have

$$\begin{aligned} \lambda_{J+1} &= \frac{1}{K-J} \sum_{k>J} \lambda_{J+1} \\ &\geq \frac{1}{K-J} \sum_{k>J} \lambda_k \\ &= \lambda_K^{(J)} \end{aligned}$$

because $\lambda_{J+1} \geq \cdots \geq \lambda_K$. Likewise $\lambda_{J+1} \geq \lambda_K^{(J+1)}$ and so

$$\begin{aligned}
\lambda_K^{(J)} - \lambda_K^{(J+1)} &= \frac{1}{K-J}\sum_{k>J}\lambda_k - \frac{1}{K-(J+1)}\sum_{k>J+1}\lambda_k \\
&= \frac{1}{K-J}\lambda_{J+1} + \left(\frac{1}{K-J} - \frac{1}{K-(J+1)}\right)\sum_{k>J+1}\lambda_k \\
&= \frac{1}{K-J}\lambda_{J+1} - \frac{1}{K-J}\lambda_K^{(J+1)} \\
&\geq 0.
\end{aligned}$$

So the difference (A11) is non-negative and equals zero when $k = K$. Thus, by Lemma A1, the eigenvalues $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS when J rises.

(ii) Fix $J \in \{0, \ldots, K\}$ and define

$$t_k^{(J)} \equiv k^2\left(\sum_{j=1}^{k}\frac{1}{\lambda_j^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-1} + \sum_{j>k}\lambda_j^{(J)}$$

for each $k \in \{1, \ldots, K\}$. Then

$$\pi^{(J)} = \overline{\lambda} - \frac{1}{K}\min\left\{t_k^{(J)} : k \in \{1, \ldots, K\}\right\}$$

from the proof of Proposition 2. But

$$\frac{\partial t_k^{(J)}}{\partial n} = -\frac{k^2}{\sigma_u^2}\left(\sum_{j=1}^{k}\frac{1}{\lambda_j^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-2}$$

is strictly negative, from which it follows that $\pi^{(J)}$ is increasing in n. □

Lemma A7. *There is a threshold $J' \in \{0, \ldots, K\}$ such that for each $J \in \{0, \ldots, K\}$, the optimal sample collected by an agent with J-deep conceptual knowledge has rank*

$$R^{(J)} = \begin{cases} K & \text{if } J \leq J' \\ J & \text{if } J' < J < R^* \\ R^* & \text{if } J \geq R^*. \end{cases}$$

Proof. Define

$$\tau_k \equiv \overline{\lambda}\left(\frac{k}{\lambda_k} - \sum_{j=1}^{k}\frac{1}{\lambda_j}\right)$$

for each $k \in \{1, \ldots, K\}$ so that

$$R^* = \max\{k \in \{1, \ldots, K\} : \tau_k \leq \tau\}$$

as in the proof of Theorem A1. Fix $J \in \{0, \ldots, K\}$ and define

$$\begin{aligned}
\tau_k^{(J)} &\equiv \overline{\lambda}\left(\frac{k}{\lambda_k^{(J)}} - \sum_{j=1}^{k} \frac{1}{\lambda_j^{(J)}}\right) \\
&= \overline{\lambda}\begin{cases} 0 & \text{if } J = 0 \\ \frac{k}{\lambda_k} - \sum_{j=1}^{k} \frac{1}{\lambda_j} & \text{if } J > 0 \text{ and } k \leq J \\ \frac{J}{\lambda_K^{(J)}} - \sum_{j=1}^{J} \frac{1}{\lambda_j} & \text{if } J > 0 \text{ and } k > J. \end{cases}
\end{aligned}$$

for each $k \in \{1, \ldots, K\}$. Then $\tau_1^{(J)} = 0$, and for each $k < K$ the difference

$$\tau_{k+1}^{(J)} - \tau_k^{(J)} = \overline{\lambda}\begin{cases} 0 & \text{if } J = 0 \\ k\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right) & \text{if } J > 0 \text{ and } k \leq J - 1 \\ J\left(\frac{1}{\lambda_K^{(J)}} - \frac{1}{\lambda_J}\right) & \text{if } J > 0 \text{ and } k = J \\ 0 & \text{if } J > 0 \text{ and } k > J. \end{cases}$$

is non-negative because $\lambda_{k+1} \leq \lambda_k$ and $\lambda_K^{(J)} \leq \lambda_J$. So $\tau_k^{(J)}$ is non-decreasing in k and

$$R^{(J)} = \max\left\{k \in \{1, \ldots, K\} : \tau_k^{(J)} \leq \tau\right\}.$$

Define $\tau_0^{(0)} \equiv 0$ and notice $\tau_J^{(J)} = \cdots = \tau_K^{(J)}$. So if $\tau \geq \tau_J^{(J)}$, then $R^{(J)} = K$; if $\tau < \tau_J^{(J)}$, then

$$\begin{aligned}
R^{(J)} &= \max\left\{k \in \{1, \ldots, J\} : \tau_k^{(J)} \leq \tau\right\} \\
&= \min\{J, \max\{k \in \{1, \ldots, K\} : \tau_k \leq \tau\}\} \\
&= \min\{J, R^*\}.
\end{aligned}$$

But if $J < K$, then

$$\begin{aligned}
\tau_{J+1}^{(J+1)} - \tau_J^{(J)} &= \overline{\lambda}\left(\left(\frac{J+1}{\lambda_{J+1}} - \sum_{j=1}^{J+1} \frac{1}{\lambda_j}\right) - \left(\frac{J}{\lambda_J} - \sum_{j=1}^{J} \frac{1}{\lambda_j}\right)\right) \\
&= J\overline{\lambda}\left(\frac{1}{\lambda_{J+1}} - \frac{1}{\lambda_J}\right)
\end{aligned}$$

is non-negative because $\lambda_{J+1} \leq \lambda_J$. So $\tau_J^{(J)}$ is non-decreasing in J, from which it follows that

$$J' \equiv \max\Big\{k \in \{0,\ldots,K\} : \tau_j^{(j)} \leq \tau \text{ for each } j \in \{0,\ldots,k\}\Big\}$$

exists and

$$R^{(J)} = \begin{cases} K & \text{if } J \leq J' \\ \min\{J, R^*\} & \text{if } J > J' \end{cases}$$

as claimed. □

Proof of Proposition 4. Part (i) follows from parts (ii)–(iv). Part (ii) follows from the definition of $\Pi^{(J)} = \pi^{(J)} - \pi^{(0)}$. Part (iii) follows from Lemma A6(i).

For part (iv), suppose $J \geq R^*$. Then $\lambda_k^{(J)} = \lambda_k$ for each $k \leq R^*$ by definition and $R^{(J)} = R^*$ by Lemma A7. Thus

$$\begin{aligned} \pi^{(J)} &= \frac{1}{K}\left(\sum_{k=1}^{R^*} \lambda_k^{(J)} - (R^*)^2\left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \\ &= \frac{1}{K}\left(\sum_{k=1}^{R^*} \lambda_k - (R^*)^2\left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \\ &= \pi^* \end{aligned}$$

and therefore $\Pi^{(J)} = \Pi$ by definition. □

A.10 Proof of Theorem 3

Our proof of Theorem 3 invokes the following lemma.

Lemma A8. *Fix $J \in \{0,\ldots,K\}$. Then $\lambda_1^{(J)},\ldots,\lambda_K^{(J)}$ undergo a MPS when $\lambda_1,\ldots,\lambda_K$ undergo a MPS.*

Proof. Fix $k \in \{1,\ldots,K\}$. By Lemma A1, the cumulative sum $\sum_{j=1}^{\min\{k,J\}} \lambda_j$ does not fall when $\lambda_1,\ldots,\lambda_K$ undergo a MPS, while the tail sum $\sum_{j>J} \lambda_j$ does not rise under the MPS. So the MPS does not lower

$$\begin{aligned} \sum_{j=1}^{k} \lambda_j^{(J)} &= \begin{cases} \sum_{j=1}^{k} \lambda_j & \text{if } k \leq J \\ \sum_{j=1}^{J} \lambda_j + (k-J)\lambda_K^{(J)} & \text{if } k > J \end{cases} \\ &= \begin{cases} \sum_{j=1}^{\min\{k,J\}} \lambda_j & \text{if } k \leq J \\ \sum_{j=1}^{K} \lambda_j - \frac{K-k}{K-J}\sum_{j>J} \lambda_j & \text{if } k > J \end{cases} \end{aligned}$$

and leaves it unchanged when $k = K$. The result follows from Lemma A1. □

Proof of Theorem 3. We prove (i) and (ii) separately:

(i) Now $\pi^{(J)}$ is non-decreasing in J (by Lemma A6), so if $\pi^{(J)} \geq \pi_0$ then $\pi^{(J+1)} \geq \pi_0$. Thus

$$\{n \geq 0 : \pi^{(J)} \geq \pi_0\} \subseteq \{n \geq 0 : \pi^{(J+1)} \geq \pi_0\}$$

and therefore $n_{\pi_0}^{(J)} \geq n_{\pi_0}^{(J+1)}$.

(ii) It suffices to show that $\pi^{(J)}$ does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. Then, since $\pi^{(J)}$ is increasing in n (by Lemma A6), the MPS expands $\{n \geq 0 : \pi^{(J)} \geq \pi_0\}$ and so cannot raise $n_{\pi_0}^{(J)}$. But the argument used to prove Lemma A4 implies that $\pi^{(J)}$ does not fall when $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS, which, by Lemma A8, happens when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. □

References

Arnold, B. C. (1987). *Majorization and the Lorenz Order: A Brief Introduction*, volume 43 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Athey, S. and Levin, J. (2018). The value of information in monotone decision problems. *Research in Economics*, 72(1):101–116.

Bardhi, A. (2024). Attributes: Selective Learning and Influence. *Econometrica*, 92(2):311–353.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Blackwell, D. (1951). Comparison of Experiments. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press.

Blackwell, D. (1953). Equivalent Comparisons of Experiments. *Annals of Mathematical Statistics*, 24(2):265–272.

Brooks, B., Frankel, A., and Kamenica, E. (2024). Comparisons of Signals. *American Economic Review*, 114(9):2981–3006.

Cabrales, A., Gossner, O., and Serrano, R. (2013). Entropy and the Value of Information for Investors. *American Economic Review*, 103(1):360–377.

Callander, S. (2011). Searching and Learning by Trial and Error. *American Economic Review*, 101(6):2277–2308.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419.

Cox, D. R. (1990). Role of Models in Statistical Analysis. *Statistical Science*, 5(2).

Davies, B. and Sankar, A. (2026). The value of conceptual knowledge. arXiv preprint 2509.09170v2.

DeGroot, M. H. (2004). *Optimal Statistical Decisions*. Wiley, first edition.

Dominitz, J. and Manski, C. F. (2017). More Data or Better Data? A Statistical Decision Problem. *Review of Economic Studies*, 84(4):1583–1605.

Esponda, I. and Pouzo, D. (2016). Berk-Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models. *Econometrica*, 84(3):1093–1130.

Frankel, A. and Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10):3650–3680.

Fudenberg, D., Kleinberg, J., Liang, A., and Mullainathan, S. (2022). Measuring the Completeness of Economic Models. *Journal of Political Economy*, 130(4):956–990.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive computation and machine learning. MIT press, Cambridge, MA.

Halevy, A., Norvig, P., and Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2):8–12.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, second edition.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, New York, NY, second edition.

Howard, R. (1966). Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26.

Iakovlev, A. and Liang, A. (2025). The Value of Context: Human versus Black Box Evaluators.

Ilut, C. and Valchev, R. (2025). Learning Optimal Behavior Through Reasoning and Experiences.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Number 6 in Cognitive science series. Harvard University Press, Cambridge, MA.

Koopmans, T. C. (1947). Measurement Without Theory. *Review of Economics and Statistics*, 29(3):161.

Lehmann, E. L. (1988). Comparing Location Experiments. *Annals of Statistics*, 16(2).

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

Mailath, G. J. and Samuelson, L. (2020). Learning under Diverse World Views: Model-Based Inference. *American Economic Review*, 110(5):1464–1501.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer-Verlag, New York.

Marinacci, M. (2015). Model Uncertainty. *Journal of the European Economic Association*, 13(6):1022–1100.

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press.

Persico, N. (2000). Information Acquisition in Auctions. *Econometrica*, 68(1):135–148.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, 2(3):225–243.

Sankar, A., Dulin, R., Davies, B., Nourani, V., Rudder, J., Salomon, A., and Taulya, G. (2025). How mechanistic explanations reshape learning and behavior: Evidence from a fertilizer choice experiment in Eastern Uganda.

Schwartzstein, J. (2014). Selective Attention and Learning. *Journal of the European Economic Association*, 12(6):1423–1452.

Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations. *Quarterly Journal of Economics*, 131(3):1243–1290.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285.

Whitmeyer, M. (2026). Making Information More Valuable. *Journal of Political Economy*, 134(3):978–1016.

Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390.

Wolpin, K. I. (2013). *The Limits of Inference without Theory*. Tjalling C. Koopmans Memorial Lectures. MIT Press.