

# The value of conceptual knowledge\*

Benjamin Davies<sup>†</sup> and Anirudh Sankar<sup>‡</sup>

Draft version: September 12, 2025

[Click here for latest version](#)

## Abstract

We formalize what it means to have conceptual knowledge about a statistical decision-making environment. Such knowledge tells agents about the structural relationships among unknown, payoff-relevant states. It allows agents to represent states as combinations of features. Conceptual knowledge is more valuable when states are more “reducible”: when their prior variances are explained by fewer features. Its value is non-monotone in the quantity and quality of available data, and vanishes with infinite data. Agents with deeper knowledge can attain the same welfare with less data. This is especially true when states are highly reducible.

*JEL classification:* C44, D83

*Keywords:* Bayesian learning, concepts, dimension reduction, eigenvalues, models, statistical decisions, value of information

---

\*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.

<sup>†</sup>Department of Economics, Stanford University; bldavies@stanford.edu.

<sup>‡</sup>Department of Economics, Stanford University; asankar@stanford.edu.

# 1 Introduction

Humans use “concepts”—mental representations—to make sense of the world and take actions in it (Murphy, 2002). Concepts help us describe the features of, and relationships among, objects in our environment. In this paper, we ask: when and why is conceptual knowledge “valuable?”

For example, suppose a farmer wants to learn the effects of applying fertilizers to his crops. He does so by trialing a combination of fertilizers and observing the resulting yield. The farmer knows the fertilizers supply nitrogen, a nutrient that helps plants grow. This allows him to mentally represent each fertilizer’s effect as the sum of a common “nitrogen effect” and an idiosyncratic effect. The farmer also knows the nitrogen effect explains more of fertilizers’ overall effects than do the idiosyncratic effects. So, in his trial, he combines fertilizers so as to isolate the nitrogen effect and hold constant the idiosyncratic effects. This makes the trial maximally informative: it leads to the largest reduction in his uncertainty about fertilizers’ overall effects.

If the farmer did not know about nitrogen (or, more generally, the concept of a nutrient), then he would not know to isolate the effect of supplying it. Instead he would view fertilizers as “black boxes,” knowing *that* they help plants grow but not knowing *why*. So he would have to learn about each fertilizer separately, making his trials less informative. His conceptual knowledge—his ability to represent fertilizers’ effects as sums of common and idiosyncratic effects—empowers him to run more informative trials. This information is valuable because it helps him make better decisions about which fertilizers to use. In contrast, his conceptual knowledge is valuable because it helps him collect more valuable information. This notion of “value” is what we study.

Quantifying the value of conceptual knowledge is important for designing interventions that give people information. It can be beneficial to give them knowledge that supplements their information and helps them interpret it (Sankar et al., 2025). If we know the value of such knowledge, then we can design better interventions that improve people’s lives more cost-effectively.

Understanding how conceptual knowledge affects inferences is also important for understanding humans’ role in the age of artificial intelligence (AI). Humans (currently) have an edge in using concepts to unite seemingly unrelated phenomena: Newton isolated a concept (gravity) that unites apples falling on earth with the orbits of other planets; Watson and Crick isolated a concept (DNA) that unites crime scene investigation with the origins of domesticated rice; Bernoulli isolated a concept (risk aversion) that unites choices in poker with choices between insurance policies. These and other concepts allow humans to learn from limited data. In contrast, AI relies on recognizing patterns in large, rich sets of data. By studying what conceptual knowledge *is*, and when and why it is valuable, we can better allocate inferential tasks between humans and AI.

**Contributions** This paper makes three major contributions. First, we formalize what it means to have conceptual knowledge about a statistical decision-making environment. Such knowledge refers to understanding the structural relationships among unknown, payoff-relevant states. It is distinct from the information one gains about these states from collecting data; it improves one’s ability to interpret data and makes them more informative. This distinction has been recognized

by others (e.g., Marinacci, 2015), but has not been formalized in a tractable and transparent way.

Second, we use our formalization of conceptual knowledge to define and characterize its value. Our definition builds on that for the instrumental value of information (Howard, 1966; Raiffa and Schlaifer, 1961): whereas information is valuable because it leads to better decisions, knowledge is valuable because it leads to better information. This definition is novel. It leads to a precise yet intuitive measure of the value of conceptual knowledge: the welfare gain from knowing how states relate structurally and using that knowledge to collect data optimally.

Third, we define the “units” of conceptual knowledge as the depth of one’s understanding of structural relationships. This definition is also novel. It allows us to compare the marginal values of “deepening” one’s knowledge and having more data. It also identifies conceptual knowledge as an economic good one may acquire in the same way data are goods one acquires. In this way, we advance the literature on learning and information acquisition that treats structural knowledge as fixed and minimally restrictive (e.g., Bardhi, 2024; Callander, 2011; Schwartzstein, 2014).

This paper also contributes to the literatures on model-based learning, model uncertainty and mis-specification, and human and machine reasoning. We explain these contributions later.

**Overview** Section 2 elaborates on our leading example of a farmer learning about fertilizers.

Section 3 extends the example to a more general setting. We consider a Bayesian agent who makes a statistical decision. His environment contains a collection of unknown, real-valued states. He learns about these states from noisy data. Then he takes real-valued actions. His loss equals the mean squared difference between the actions and states. The agent takes the actions that minimize his expected posterior loss.

The agent’s prior on the states encodes his conceptual knowledge. This knowledge tells him how the states relate structurally. It allows him to represent them as linear combinations of “features.” These features form a basis for the Euclidean space of state vectors. They are eigenvectors of the state vector’s prior variance matrix. The corresponding eigenvalues capture features’ influences on the states: an eigenvalue is larger when the corresponding feature contributes more to the states’ prior variances.

We say states are more “reducible” when they are influenced by fewer features. This happens when the eigenvalues of the prior variance matrix are more spread out. For example, if one eigenvalue is much larger than the others, then the state vector is likely to belong to a one-dimensional subspace of the many-dimensional state space. The more spread out are the eigenvalues, the more the agent can “reduce” states by representing them as low-dimensional combinations of high-dimensional features. This dimension reduction is what makes conceptual knowledge valuable.

Section 4 contains definitions and preliminary results that we draw upon later. For example, we define the “value” of the agent’s data, derive sharp bounds on this value (see Proposition 1), and define what it means for eigenvalues to be “more spread out.”

We define the value of conceptual knowledge in Section 5. First, we suppose the agent collects an “optimal sample” that leads to the least posterior expected loss among all samples of a given

size. This sample contains information about the most influential features only (see Proposition 2). Its design relies on the agent’s knowledge of how the states relate structurally.

Next, we consider a counterfactual agent who does not know how the states relate structurally. This “naïve” agent has no reason to believe the states covary or have different prior variances. So he assumes they do not. Consequently, when he collects an optimal sample, he tries to learn about every state equally. This is because he cannot identify features with different influences on the states, precluding him from focusing on the most influential features.

We compute the difference between the values of the true and naïve agent’s optimal samples, and we call this difference the “value of conceptual knowledge.” It captures the welfare gain from knowing how the states relate structurally and using that knowledge to collect data optimally.

Our first main result (Theorem 1) says that conceptual knowledge is more valuable when states are more reducible. If a few features contribute most of the states’ prior variances, then the agent gains a lot from identifying those features and collecting data on them (i.e., “asking the right questions”). In contrast, if every feature contributes equally, then he gains nothing from identifying those features because he collects the same data that he would if he was naïve.

Our second main result (Theorem 2) says that the value of conceptual knowledge (i) is non-monotone in the agent’s access to data and (ii) vanishes when he has infinite data. If the agent has more data, then he can learn more about the features on which he collects data, raising the gain from identifying the most influential features. However, he can also collect data on more features, lowering the gain from identifying the most influential features. The first effect dominates the second precisely when the sample is sufficiently small. As it becomes arbitrarily 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.

In Section 6, we extend our measure of the value of conceptual knowledge to one of “deeper” knowledge. We suppose the agent can identify some, but not all, of the most influential features, and refer to the “depth” of his knowledge as the number he can identify. Our third main result (Theorem 3) says that deeper knowledge is weakly more valuable. However, if the agent’s knowledge is sufficiently deep, then deepening it further yields no additional value because it does not change the data he collects.

Finally, in Section 7, we study the trade-off between having deeper knowledge or more data. We measure the agent’s welfare by the least posterior expected loss he can attain given the depth of his knowledge and his access to data. Our fourth main result (Theorem 4) says that if he has deeper knowledge or the states are more reducible, then he can attain the same welfare with less data. This is because he can design better samples and extract more value from each observation, lowering the number he needs to attain a given welfare target.

Section 8 discusses our modeling assumptions and related literature. Section 9 concludes. Appendix A contains additional discussions and results. Appendix B contains proofs of our mathematical claims.

## 2 An illustrative example

This section elaborates on the example presented in our introduction. The example is inspired by our empirical work in Uganda, where we study the role that conceptual knowledge plays when farmers learn about fertilizers (Sankar et al., 2025).

**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.<sup>1</sup> 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  $\sigma_u^2 > 0$ . It captures the randomness in  $y$  due to variation in unobserved factors.

**Conceptual knowledge** The farmer knows the two fertilizers supply equal amounts of nitrogen, a nutrient that helps plants grow. This knowledge is purely conceptual: it comes from recognizing that fertilizers supply nutrients and from understanding the mechanisms through which plants grow, rather than from observing data on plants' growth. It allows him to mentally represent each effect  $\theta_k$  as the sum of a common "nitrogen effect" and an idiosyncratic 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  $\theta$ . The common and idiosyncratic effects  $\gamma_1$  and  $\gamma_2$  are the coordinates of  $\theta$  over this basis. The farmer does not know these coordinates, but he knows  $\gamma_1$  contributes more to the prior variances of  $\theta_1$  and  $\theta_2$  than does  $\gamma_2$ . So he assumes  $\gamma_1$  and  $\gamma_2$  are independently distributed with variances  $\lambda_1 = \sigma^2(1 + \rho)$  and  $\lambda_2 = \sigma^2(1 - \rho)$ . The sum

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

---

<sup>1</sup>We interpret negative values of  $w_k$  as using less of fertilizer  $k$  than the farmer uses currently.

of these variances equals the sum of the prior variances of  $\theta_1$  and  $\theta_2$ . The parameter  $\rho \in [0, 1)$  determines the share

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

of this sum contributed by  $\gamma_1$ . This share equals  $1/2$  when  $\rho = 0$ , in which case  $\gamma_1$  and  $\gamma_2$  contribute equally. It equals one in the limit as  $\rho \rightarrow 1$ , in which case only  $\gamma_1$  contributes. The larger is  $\rho$ , the more likely is  $\theta$  to belong to the one-dimensional subspace of  $\mathbb{R}^2$  spanned by  $v_1$ .

The coordinate vector  $\gamma \equiv (\gamma_1, \gamma_2)$  has variance

$$\mathbb{V}(\gamma) = \sigma^2 \begin{bmatrix} 1 + \rho & 0 \\ 0 & 1 - \rho \end{bmatrix},$$

and so the effect vector

$$\theta = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \gamma$$

has prior variance

$$\begin{aligned} \mathbb{V}(\theta) &= \left( \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right) \mathbb{V}(\gamma) \left( \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right)^T \\ &= \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \end{aligned} \tag{1}$$

Thus  $\theta_1$  and  $\theta_2$  have equal prior variances  $\sigma^2$  and correlation  $\rho$ . Intuitively, the more  $\theta_1$  and  $\theta_2$  are determined by the common effect  $\gamma_1$ , 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  $\lambda_k$  equals the prior variance of  $\theta$  in the direction of the corresponding eigenvector  $v_k$ . So  $\theta$  has the most prior variance in the direction of  $v_1$  and the least in the direction of  $v_2$ .

**Value of information** The farmer's data  $\mathcal{S} \equiv \{(w, y)\}$  comprise the vector  $w = (w_1, w_2)$  and outcome  $y$ . These data are valuable insofar as they make the farmer's beliefs about  $\theta$  more precise. We measure the value of  $\mathcal{S}$  via the mean difference

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

between the prior and posterior variances of  $\theta_1$  and  $\theta_2$ . This difference is largest when  $w = \pm v_1$  and is smallest when  $w = \pm v_2$  (see Proposition A2). For example, choosing  $w = v_1$  makes  $y = \gamma_1 + u$  a "pure signal" of  $\gamma_1$ . This makes  $\mathcal{S}$  maximally valuable because it provides information about the component of  $\theta$  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  $\gamma_2$ .

This makes  $\mathcal{S}$  *minimally* valuable because it provides information about the component of  $\theta$  with the *least* prior variance, leading to the *smallest* difference between prior and posterior variances.<sup>2</sup>

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

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

The farmer attains  $\pi^*$  by choosing  $w = \pm v_1$ . Doing so relies on his conceptual knowledge about fertilizers' effects: his ability to identify the common and idiosyncratic components of each effect  $\theta_k$ . If the farmer lacked this ability, then he would have no reason to believe  $\theta_1$  and  $\theta_2$  were correlated. So he would assume  $\rho = 0$  and the data would have maximal value

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

The difference

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

between  $\pi^*$  and  $\pi^{(0)}$  captures the value of the farmer's conceptual knowledge: the value of knowing how  $\theta_1$  and  $\theta_2$  relate structurally, and using this knowledge to collect data optimally.

The value  $\Pi$  of the farmer's conceptual knowledge is larger when  $\rho$  is larger.<sup>3,4</sup> Intuitively, the more likely it is  $\theta$  that belongs to a low-dimensional subspace, the more precision can be gained by identifying that subspace and focusing on it when collecting data. We formalize this intuition in Section 5, and generalize it to a setting in which  $\mathcal{S}$  has arbitrary size and  $\theta$  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). Hence raising  $\rho$  raises  $\Pi$ : 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 makes a statistical decision. This section describes the agent's environment, formalizes his conceptual knowledge about that environment, and presents some specific examples.

---

<sup>2</sup>In general, the data  $\mathcal{S}$  are most valuable when they contain information about the components of  $\theta$  with the most prior variance. We formalize and prove this claim in Sections 4.3 and 4.4.

<sup>3</sup>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 Proposition A2 and the definition of  $\pi^{(0)}$ . So  $\partial \pi^* / \partial \rho > 0$  and  $\partial \pi^{(0)} / \partial \rho = 0$ , from which it follows that  $\partial \Pi / \partial \rho > 0$ .

<sup>4</sup>For example, the correlation  $\rho$  will be close to one when the fertilizers supply nitrogen only, and close to zero when their nutrient profiles are very different.

### 3.1 Environment

**Prior** There is a vector  $\theta \equiv (\theta_1, \dots, \theta_K)$  of unknown, real-valued “states.” The agent’s prior on  $\theta$  is a probability distribution  $\mathbb{P}$  over the  $K$ -dimensional Euclidean space  $\mathbb{R}^K$ . This distribution is normal with mean  $\mu \in \mathbb{R}^K$  and variance  $\Sigma^{K \times K}$ :

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

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

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

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

equal to the sum of  $\theta^T w^{(i)}$  and an independently normally distributed error  $u^{(i)} \sim \mathcal{N}(0, \sigma_u^2)$  with mean zero and variance  $\sigma_u^2 > 0$ .

**Actions and losses** The agent uses his prior  $\mathbb{P}$ , the sample  $\mathcal{S}$ , and Bayes’ rule to form posterior beliefs about  $\theta$ . Then he chooses a  $K$ -vector  $a \equiv (a_1, \dots, a_K)$  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.<sup>5</sup>

Let  $\mathbb{E}$  take expectations with respect to  $\mathbb{P}$ . The agent chooses the action vector that minimizes his posterior expected loss:<sup>6</sup>

$$a \in \arg \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}]. \quad (3)$$

---

<sup>5</sup>Suppose  $p_1, \dots, p_K$  are strictly positive and sum to one. Let  $D$  be the  $K \times K$  diagonal matrix with  $kk^{\text{th}}$  entry  $Kp_k$ , and let  $a' \equiv Da$  and  $\theta' \equiv D\theta$ . Then

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

is a weighted average of the squared differences between the actions and corresponding states. The weights  $p_1, \dots, p_K$  encode the agent’s preferences: the larger is  $p_k$ , the larger is the loss from taking an action  $a_k$  different than the state  $\theta_k$ . We focus on the case with  $p_k = 1/K$  for each  $k$ , which makes  $D$  equal the identity matrix and  $\theta'$  equal  $\theta$ . However, we can easily generalize our analysis to a setting with non-equal weights by replacing  $\theta$  with  $\theta'$ . Then what matters are the eigenvalues and eigenvectors of  $\mathbb{V}(\theta') = D\Sigma D^T$ , rather than those of  $\Sigma$ . This does not change our results or insights substantively.

<sup>6</sup>In Appendix Section A1, we explain how the choice problem (3) is equivalent to a prediction problem that arises in the machine and statistical learning literatures. This equivalence comes from interpreting  $\theta_1, \dots, \theta_K$  as values of an unknown function.



### 3.2 Conceptual knowledge

Our definition of conceptual knowledge stems from psychologists'. They define concepts as mental representations that help us describe objects in our environment and how they relate (Murphy, 2002). Accordingly, our agent's conceptual knowledge tells him how the states  $\theta_1, \dots, \theta_K$  relate. It allows him to represent

$$\theta = \sum_{k=1}^K \gamma_k v_k \quad (4)$$

as a linear combination of orthonormal vectors  $v_1, \dots, v_K \in \mathbb{R}^K$ . These vectors encode structural "features" of  $\theta$ .<sup>7,8</sup> The coefficients  $\gamma_1, \dots, \gamma_K \in \mathbb{R}$  encode features' influences on  $\theta$ . They are akin to "deep parameters" that determine the "reduced-form" states  $\theta_1, \dots, \theta_K$  via the structural relationships embedded in  $v_1, \dots, v_K$  (Lucas, 1976).

The agent knows the vectors  $v_1, \dots, v_K$ . He does not know the coefficients  $\gamma_1, \dots, \gamma_K$ , but he knows some contribute more to the states' prior variances than others. Specifically, he knows each coefficient  $\gamma_k$  is independently distributed with variance  $\lambda_k > 0$  non-decreasing in  $k$ .<sup>9</sup> Then  $\theta$  has prior variance

$$\begin{aligned} \Sigma &= V \Lambda V^T \\ &= \sum_{k=1}^K \lambda_k v_k v_k^T, \end{aligned} \quad (5)$$

where

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

is the  $K \times K$  diagonal matrix with entries  $\lambda_1, \dots, \lambda_K$  and

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

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

<sup>7</sup>The vectors  $v_1, \dots, v_K$  may not correspond to physical features of the agent's environment. Instead they are mental constructs he uses to make sense of his environment. For example, nutrients like nitrogen are mental constructs—no farmer "sees" them. All they can see are the effects of applying fertilizers. This is why we refer to identifying  $v_1, \dots, v_K$  as having "conceptual" knowledge.

<sup>8</sup>In Appendix A1.2, we show how to derive  $v_1, \dots, v_K$  from an approximating "model" that captures the "generalizable structure" of the agent's environment.

<sup>9</sup>It is without loss of generality to assume  $\gamma_1, \dots, \gamma_K$  are independently distributed. This is because  $\Lambda$  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  $\Sigma$  has eigendecomposition  $V' \Lambda' (V')^T$ , so we can carry out our analysis by replacing  $V$  with  $V'$  and  $\Lambda$  with  $\Lambda'$ . Likewise, it is without loss to assume  $\lambda_1 \geq \dots \geq \lambda_K$  because we can permute the indices of the eigenpairs  $(\lambda_k, v_k)$  without changing  $\Sigma$ .

Equation (5) is an eigendecomposition of  $\Sigma$ . The  $k^{\text{th}}$  largest eigenvalue  $\lambda_k = \mathbb{V}(\gamma_k)$  of  $\Sigma$  equals the prior variance of  $\theta$  in the direction of the corresponding unit eigenvector  $v_k$ . The trace

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

of  $\Sigma$  equals the sum of the eigenvalues  $\lambda_1, \dots, \lambda_K$ . So these eigenvalues' mean

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

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

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

In contrast, if  $\lambda_1 / \text{tr}(\Sigma) \approx 1$ , then  $\gamma_1$  contributes most of the states' prior variances.

The distribution of  $\lambda_1, \dots, \lambda_K$  around their mean  $\bar{\lambda} = \text{tr}(\Sigma)/K$  captures the states' "reducibility": they are more "reducible" when they are influenced by fewer features, which happens precisely when  $\lambda_1, \dots, \lambda_K$  are more spread out around  $\bar{\lambda}$ .<sup>10,11</sup> Thus, the agent's conceptual knowledge allows him to "reduce" states by representing them as low-dimensional combinations of higher-dimensional features.

If the agent had no conceptual knowledge, then he would not be able to reduce states because he would not know how to represent them as combinations of features with different influences. So his prior variance matrix

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

would equal the prior variance matrix in the case when  $\lambda_k = \bar{\lambda}$  for each  $k \in \{1, \dots, K\}$ .<sup>12</sup>

We use the matrix  $\Sigma^{(0)}$  to measure how much the agent's conceptual knowledge allows him to reduce states. We consider a counterfactual "naïve" agent whose prior  $\mathbb{P}^{(0)} \equiv \mathcal{N}(\mu, \Sigma^{(0)})$  on  $\theta$  has a different variance than the "true" agent's prior  $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$ . The naïve agent does not how the

<sup>10</sup>We formalize what it means for  $\lambda_1, \dots, \lambda_K$  to be "more spread out" in Section 4.5.

<sup>11</sup>If  $\lambda_2 = \dots = \lambda_K$  (as in Example 1), then the distribution of  $\lambda_1, \dots, \lambda_K$  is fully determined by the leading eigenvalue  $\lambda_1$  and the "spectral gap"  $(\lambda_1 - \lambda_2)$ . This gap appears elsewhere in the statistical literature: it determines Markov chains' mixing times (Levin et al., 2008) and whether principal components can be estimated consistently (Yu et al., 2015).

<sup>12</sup>The naïve prior variance matrix  $\Sigma^{(0)}$  is robust to mis-specification in that it commits as little as possible to any given covariance structure. By spreading variance evenly across all dimensions of  $\mathbb{R}^K$ , it avoids overweighting components of  $\theta$  that later prove irrelevant. However, this robustness comes at a cost: it forfeits any gains that could be obtained by overweighting components of  $\theta$  that later prove essential.

states relate structurally and, thus, assumes  $\theta$  has prior variance  $\Sigma^{(0)}$ . The true agent knows how the states relate structurally and, thus, knows  $\theta$  has prior variance  $\Sigma$ .

Since  $\mathbb{P}$  and  $\mathbb{P}^{(0)}$  are normal distributions with equal means, the Kullback-Leibler (hereafter “KL”) divergence from  $\mathbb{P}$  and  $\mathbb{P}^{(0)}$  equals<sup>13</sup>

$$\begin{aligned}\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) &= \frac{1}{2} \left( \text{tr}((\Sigma^{(0)})^{-1}\Sigma) - K + \ln \left( \frac{\det(\Sigma^{(0)})}{\det(\Sigma)} \right) \right) \\ &= -\frac{1}{2} \sum_{k=1}^K \ln \left( \frac{\lambda_k}{\bar{\lambda}} \right).\end{aligned}\tag{6}$$

The KL divergence (6) measures the “information gain” from using  $\mathbb{P}$  as a prior rather than  $\mathbb{P}^{(0)}$ . This information is purely conceptual: it does not depend on the sample  $\mathcal{S}$ . It comes from knowing how to represent states as low-dimensional combinations of high-dimensional features. We study the value of this dimension reduction in Section 5.

The KL divergence (6) depends on the eigenvalues  $\lambda_1, \dots, \lambda_K$  of the “true” prior variance matrix  $\Sigma$ . It equals zero when  $\lambda_1, \dots, \lambda_K$  are equal (to  $\bar{\lambda}$ ) and is larger when they are more spread out (see Proposition A4).<sup>14</sup> If some features are more influential than others, then identifying the most influential features empowers the agent to reduce  $\theta_1, \dots, \theta_K$ . Intuitively, identifying these features tells him “where to look” for the vector  $\theta$  in the space  $\mathbb{R}^K$  containing it.

### 3.3 Examples

Below are two examples of how the prior variance matrix  $\Sigma$  encodes the structural relationships among the states. The first example generalizes the setting described in Section 2. It builds  $\Sigma$  from first principles, starting with its eigenvalues and eigenvectors. The second example builds  $\Sigma$  from knowledge of how the states are generated, then derives its eigenvalues and eigenvectors.

**Example 1** (Pairwise correlated states). Suppose the agent knows each state  $\theta_k$  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  $\mathbf{1}_K \equiv (1, \dots, 1)$  is the  $K$ -vector of ones. He encodes the idiosyncratic components by unit vectors  $v_2, \dots, v_K$  that are orthogonal to  $v_1$  and each other. The  $k^{\text{th}}$  coefficient  $\gamma_k$  in (4) has prior

<sup>13</sup>See Rasmussen and Williams (2006, Section A.5) for a derivation of (6).

<sup>14</sup>If  $\mathbb{P}^{(0)}$  has mean  $\mu^{(0)} \in \mathbb{R}^K$ , then (6) becomes

$$\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = \frac{1}{2} \left( \frac{1}{\bar{\lambda}} \|\mu - \mu^{(0)}\|^2 - \sum_{k=1}^K \ln \left( \frac{\lambda_k}{\bar{\lambda}} \right) \right).$$

So even if  $\mu^{(0)} \neq \mu$ , the KL divergence from  $\mathbb{P}$  to  $\mathbb{P}^{(0)}$  is non-negative and does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS (see Proposition A4). But it is strictly larger than zero when  $\mu^{(0)} \neq \mu$ , even if  $\lambda_1 = \dots = \lambda_K$ .

variance

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

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

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

of the prior variances of  $\theta_1, \dots, \theta_K$  contributed by the coefficient  $\gamma_1$  on  $v_1$ . This share equals  $1/K$  when  $\rho = 0$ , in which case  $\lambda_k$  is constant in  $k$  and so  $\gamma_1, \dots, \gamma_K$  contribute to the prior variances of  $\theta_1, \dots, \theta_K$  equally. It equals one in the limit as  $\rho \rightarrow 1$ , in which case only  $\gamma_1$  contributes.

Since  $v_1, \dots, v_K$  are orthonormal, the sum

$$\sum_{k=1}^K v_k v_k^T = I_K$$

of their outer products equals the  $K \times K$  identity matrix. Therefore, the prior variance matrix

$$\begin{aligned} \Sigma &= \lambda_1 v_1 v_1^T + \lambda_K (I_K - v_1 v_1^T) \\ &= \rho \sigma^2 \mathbf{1}_K \mathbf{1}_K^T + (1 - \rho) \sigma^2 I_K \\ &= \sigma^2 \begin{bmatrix} 1 & \rho & \cdots \\ \rho & 1 & \\ \vdots & & \ddots \end{bmatrix} \end{aligned} \tag{7}$$

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

**Example 2 (Random walk).** Let  $\nu > 0$ . Suppose the agent knows  $\theta_1, \dots, \theta_K$  are values of a random walk with known initial value  $\theta_0 \in \mathbb{R}$  and unknown, independently distributed increments

$$\theta_k - \theta_{k-1} \sim \mathcal{N}(0, \nu^2).$$

Then the prior variance matrix

$$\Sigma = \nu^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & & \vdots \\ 1 & 2 & \cdots & K \end{bmatrix} \tag{8}$$

has  $j^{\text{th}}$  entry  $\Sigma_{jk} = \nu^2 \min\{j, k\}$ . Fortiana and Cuadras (1997) show that (8) has  $k^{\text{th}}$  largest eigenvalue

$$\lambda_k = \frac{\nu^2}{4} \csc^2 \left( \frac{(2k-1)\pi}{2K+1} \right)$$

and that the corresponding unit eigenvector  $v_k$  has  $j^{\text{th}}$  component

$$[v_k]_j = \frac{2}{\sqrt{2K+1}} \sin \left( \frac{j(2k-1)\pi}{2K+1} \right).$$

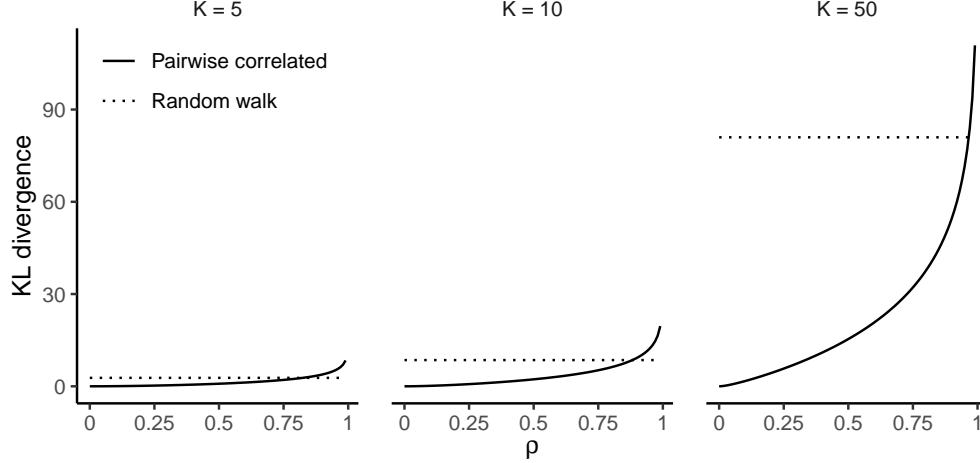


Figure 1: KL divergences  $\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)})$  when states are pairwise correlated (i.e., when  $\theta$  has prior variance (7)) and when they follow a random walk (i.e., when  $\theta$  has prior variance (8) with  $v^2 = 2\sigma^2 / (K + 1)$ )

The eigenvalues of (7) have mean  $\sigma^2$ , whereas the eigenvalues of (8) have mean  $v^2(K + 1)/2$ . Choosing  $v^2 = 2\sigma^2 / (K + 1)$  equates these two means but does not equate the eigenvalues’ distributions, nor the KL divergences  $\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)})$  those distributions imply. We illustrate this fact in Figure 1. It shows that assuming states follow a random walk is equivalent, in terms of how much prior structure it imposes, to assuming a large pairwise correlation.<sup>15</sup> This is especially true when there are many states: if  $K = 5$ , then the equivalent correlation is about 0.82; if  $K = 50$ , then it is about 0.97.

## 4 Preliminaries

This section contains definitions and preliminary results that we draw upon in later sections. First, we characterize the optimal action vector (3) and the posterior expected loss it induces. This loss depends on the posterior variance of the unknown vector  $\theta$  given the sample  $\mathcal{S}$ . We define the “value of  $\mathcal{S}$ ” to be the difference between the prior and posterior expected losses, and derive sharp lower and upper bounds on this value. Finally, we formalize what it means for the eigenvalues of the prior variance matrix to be “more spread out.”

<sup>15</sup>Callander (2011) and others use Brownian motions (the continuous-time analogues of random walks) as tools for modeling “complexity.” They define “complex” environments as those in which only local learning is possible: learning a state provides some information about nearby states but little about distant states (see also Bardhi (2024)). The limiting case is when learning a state provides *no* information about others; in our framework, this happens when the states are uncorrelated. Yet Figure 1 suggests that Brownian motions are as structurally restrictive as assuming states are highly correlated.

## 4.1 Optimal actions and expected losses

Let  $\mathbb{V}$  take variances with respect to the prior distribution  $\mathbb{P}$ . Lemma 1 characterizes the optimal action vector (3) and the posterior expected loss it induces.<sup>16</sup> This vector equals the posterior mean of  $\theta$ . It induces a posterior expected loss equal to the mean of the posterior variances of  $\theta_1, \dots, \theta_K$ .

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

$$\mathbb{E}[L(\theta, a) \mid \mathcal{S}] = \frac{1}{K} \sum_{k=1}^K \mathbb{V}(\theta_k \mid \mathcal{S}). \quad (9)$$

## 4.2 Posterior variance and Gram matrices

The expected loss (9) depends on the trace

$$\text{tr}(\mathbb{V}(\theta \mid \mathcal{S})) = \sum_{k=1}^K \mathbb{V}(\theta_k \mid \mathcal{S})$$

of the posterior variance matrix  $\mathbb{V}(\theta \mid \mathcal{S})$ . Lemma 2 uses the prior variance matrix  $\Sigma$  and the  $K \times K$  “Gram matrix”

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

to characterize  $\mathbb{V}(\theta \mid \mathcal{S})$ .

**Lemma 2.** *We have*

$$\mathbb{V}(\theta \mid \mathcal{S}) = \left( \Sigma^{-1} + \frac{1}{\sigma_u^2} G \right)^{-1}. \quad (10)$$

## 4.3 Value of information

If the agent did not observe the sample  $\mathcal{S}$ , then, by Lemma 1, his expected loss would equal the mean of the prior variances of  $\theta_1, \dots, \theta_K$ . Observing  $\mathcal{S}$  lowers his minimized expected loss by

$$\begin{aligned} \pi(\mathcal{S}) &\equiv \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a')] - \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}] \\ &= \frac{1}{K} \sum_{k=1}^K (\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k \mid \mathcal{S})). \end{aligned} \quad (11)$$

We call  $\pi(\mathcal{S})$  the “value of  $\mathcal{S}$ .”<sup>17</sup> It is non-negative, grows as  $\mathcal{S}$  grows, and shrinks as  $\sigma_u^2$  grows (see Proposition A1). Intuitively, if the agent has more information, then he can take actions that estimate the states more accurately.

<sup>16</sup>Lemma 1 holds even when  $\mathbb{P}$  is not normal. We assume  $\mathbb{P}$  is normal so that we can derive closed-form expressions for the posterior variances of  $\theta_1, \dots, \theta_K$  and, thus, the value (11) of  $\mathcal{S}$ .

<sup>17</sup>Raiffa and Schlaifer (1961, p. 90) define a similar object and call it the “(expected) value of sample information.”

We can write (11) in terms of the traces of the prior and posterior variance matrices:

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

Thus, by Lemma 2, the value of  $\mathcal{S}$  depends on the Gram matrix  $G$ . This matrix 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  $\delta_1 \geq \dots \geq \delta_K \geq 0$  and a  $K \times K$  orthogonal matrix

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

such that

$$\begin{aligned} G &= \Omega \Delta \Omega^T \\ &= \sum_{k=1}^K \delta_k \omega_k \omega_k^T. \end{aligned} \tag{12}$$

Then  $\delta_1, \dots, \delta_K$  are the eigenvalues of  $G$  and  $\omega_1, \dots, \omega_K \in \mathbb{R}^K$  are the corresponding unit eigenvectors. Proposition 1 uses the eigendecompositions (5) and (12) 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*

$$\frac{1}{K} \sum_{k=1}^K \left( \lambda_k - \left( \frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2} \right)^{-1} \right) \stackrel{*}{\leq} \pi(\mathcal{S}) \stackrel{**}{\leq} \frac{1}{K} \sum_{k=1}^K \left( \lambda_k - \left( \frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1} \right), \tag{13}$$

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

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

#### 4.4 Optimal samples

Suppose the eigenvectors of  $\Sigma$  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

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

of the posterior variance matrix falls. This trace depends on the eigenvalues  $\delta_1, \dots, \delta_K$  of  $G$ , which are non-negative, non-increasing, and sum to  $n$ .<sup>18</sup> So  $\pi(\mathcal{S})$  is maximized when  $\delta_1, \dots, \delta_K$  solve

$$\min_{\delta_1, \dots, \delta_K \in \mathbb{R}} \sum_{k=1}^K \left( \frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1} \quad \text{subject to } \delta_1 \geq \dots \geq \delta_K \geq 0 \text{ and } \sum_{k=1}^K \delta_k = n. \quad (14)$$

Proposition 2 describes a solution to (14). It uses the integer

$$R^* \equiv \max \left\{ k \in \{1, \dots, K\} : \sum_{j=1}^k \frac{1}{\lambda_j} + \frac{n}{\sigma_u^2} \geq \frac{k}{\lambda_k} \right\} \quad (15)$$

to provide a sharp upper bound

$$\pi^* \equiv \frac{1}{K} \left( \sum_{k=1}^{R^*} \lambda_k - (R^*)^2 \left( \sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} \right) \quad (16)$$

on the value of  $\mathcal{S}$ .<sup>19</sup>

**Proposition 2.** *Define*

$$\delta_k^* \equiv \begin{cases} \frac{n}{R^*} + \sigma_u^2 \left( \frac{1}{R^*} \sum_{j=1}^{R^*} \frac{1}{\lambda_j} - \frac{1}{\lambda_k} \right) & \text{if } k \leq R^* \\ 0 & \text{if } k > R^* \end{cases} \quad (17)$$

for each  $k \in \{1, \dots, K\}$ . Then  $\pi(\mathcal{S}) \leq \pi^*$  with equality if

$$G = \sum_{k=1}^K \delta_k^* v_k v_k^T. \quad (18)$$

We call the sample “optimal” if it induces the Gram matrix (18). The agent can construct such a sample as follows: for each  $k \in \{1, \dots, K\}$ , collect  $\delta_k^*$  observations with covariate  $v_k$ .<sup>20</sup> Then  $\mathcal{S}$  contains information about the  $R^*$  most influential features only. This is the optimal way for the agent to “spend his data budget”: by learning about the features that matter and ignoring those that do not. The number  $R^*$  of features that “matter” grows as his budget (i.e., the sample size  $n$ ) grows. We call  $R^*$  the “rank” of an optimal sample because it is the rank of the Gram matrix (18).

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

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

---

<sup>18</sup>Indeed

$$\sum_{k=1}^K \delta_k = \text{tr}(G) = \text{tr} \left( \sum_{i=1}^n w^{(i)} (w^{(1)})^T \right) \stackrel{*}{=} \sum_{i=1}^n \text{tr} \left( (w^{(i)})^T w^{(i)} \right) \stackrel{**}{=} n,$$

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

<sup>19</sup>Proposition 2 echoes Liang et al.’s (2022) Theorem 1, which says that if there are two unknown states (which Liang et al. call “attributes”), then one should prioritize learning about the state with more prior variance.

<sup>20</sup>This may be infeasible for two reasons: (i) the eigenvalues  $d_1^*, \dots, d_K^*$  may not be integers; (ii) the agent may not be able to choose  $v_1, \dots, v_K$  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.



and trace<sup>21</sup>

$$\sum_{k=1}^K \left( \frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2} \right)^{-1} = (R^*)^2 \left( \sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} + \sum_{k>R^*} \lambda_k. \quad (19)$$

The eigenvalues of  $\mathbb{V}(\theta \mid \mathcal{S})$  are the posterior variances of the unknown coefficients  $\gamma_1, \dots, \gamma_K$ . So if  $\mathcal{S}$  is optimal, then it equates the posterior variances of  $\gamma_1, \dots, \gamma_{R^*}$  to each other and the posterior variances of  $\gamma_{R^*+1}, \dots, \gamma_K$  to their prior variances.<sup>22</sup> Intuitively, the agent has a target variance and designs  $\mathcal{S}$  so as to bring the posterior variances of  $\gamma_1, \dots, \gamma_K$  below that target.<sup>23</sup> This minimizes the trace (19) given the sample size  $n$ .

#### 4.5 Mean-preserving spreads

Finally, consider the eigenvalues  $\lambda_1, \dots, \lambda_K$  of the prior variance matrix  $\Sigma$ . Let  $F : (0, \infty) \rightarrow [0, 1]$  be their (empirical) cumulative distribution function (hereafter “CDF”):

$$F(z) = \frac{|\{k \in \{1, \dots, K\} : \lambda_k \leq z\}|}{K} \quad (20)$$

for all  $z > 0$ . A “mean-preserving spread” (hereafter “MPS”) of  $F$  is a CDF  $F' : (0, \infty) \rightarrow [0, 1]$  such that

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

$$\int_0^\infty z \, dF(z) = \int_0^\infty z \, dF'(z).$$

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

$$\int_0^z (F'(t) - F(t)) \, dt \geq 0.$$

These are the “integral conditions” from Rothschild and Stiglitz (1970). Condition (ii) says that  $F'$  has more weight in its tails than  $F$ , capturing the idea of eigenvalues being more spread out.

We say  $\lambda_1, \dots, \lambda_K$  “undergo a MPS” when their CDF (20) undergoes a MPS. This changes the trace of the posterior variance matrix without changing the trace of  $\Sigma$ . So if  $\lambda_1, \dots, \lambda_K$  undergo a MPS, then the agent’s posterior expected loss changes but his prior expected loss does not. This makes MPSs useful for analyzing how the value of  $\mathcal{S}$  depends on the distribution of  $\lambda_1, \dots, \lambda_K$ . We discuss this dependence in Section 5 and Appendix Section A2, in which we state results that depend on the following lemma:

<sup>21</sup>The two terms on the RHS of (19) correspond to the “sampling” and “extrapolation” errors discussed in Appendix Section A2.3.

<sup>22</sup>This equality of large eigenvalues and ignorance of small eigenvalues is reminiscent of Arrow’s (1963) theorem on the optimality of deductible insurance contracts. Such contracts second-degree stochastically dominate all other contracts with the same premia (Gollier and Schlesinger, 1996). They provide full coverage against risks above a minimum threshold. Similarly, optimal samples provide “full coverage against posterior variance” above a minimum threshold.

<sup>23</sup>This strategy is called “reverse water-filling” in rate-distortion theory—see Cover and Thomas (2006, Chapter 10). It also appears in Ilut and Valchev’s (2025) model of abstract reasoning.

**Lemma 3.** Let  $\lambda_k > 0$  and  $\lambda'_k > 0$  be non-increasing in  $k \in \{1, \dots, K\}$ , and let  $F$  and  $F'$  be their CDFs defined as in (20). 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  $g : (0, \infty) \rightarrow \mathbb{R}$ .
- (iii)  $\sum_{j=1}^k \lambda'_j \geq \sum_{j=1}^k \lambda_j$  for each  $k \in \{1, \dots, K\}$ , with equality when  $k = K$ .
- (iv)  $\sum_{j=k}^K \lambda'_j \leq \sum_{j=k}^K \lambda_j$  for each  $k \in \{1, \dots, K\}$ , with equality when  $k = 1$ .

For example, consider the prior variance matrix (7) constructed in Example 1. This matrix has eigenvalues  $\lambda_1 = (1 + \rho(K - 1))\sigma^2$  and  $\lambda_2 = \dots = \lambda_K = (1 - \rho)\sigma^2$ . Their  $k^{\text{th}}$  partial sum

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

is increasing in  $\rho$  when  $k < K$  and constant in  $\rho$  when  $k = K$ . Thus, by Lemma 3, the eigenvalues of (7) undergo a MPS when  $\rho$  rises.

## 5 Value of conceptual knowledge

Whereas information is valuable insofar as it helps the agent make better decisions (i.e., take actions  $a_1, \dots, a_K$  that estimate the states  $\theta_1, \dots, \theta_K$  more accurately), conceptual knowledge is valuable insofar as it helps him obtain better information.

We formalize this idea as follows. Suppose the agent collects an optimal sample with value  $\pi^*$  (see Section 4.4). Doing so relies on his conceptual knowledge of how the states relate structurally: on his ability to identify the most influential features and focus on them when collecting data. If he lacked this ability, then he would use the “naïve” prior  $\mathbb{P}^{(0)} = \mathcal{N}(\mu, \Sigma^{(0)})$  described in Section 3.2. The prior variance matrix  $\Sigma^{(0)} = \bar{\lambda}I_K$  would have equal eigenvalues  $\lambda_1^{(0)} = \dots = \lambda_K^{(0)} = \bar{\lambda}$ . So, by analogy to (15) and (16), the agent’s optimal sample would have rank

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

and value

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

where

$$\tau \equiv \frac{n/\sigma_u^2}{1/\bar{\lambda}}$$

indexes the precision of the data relative to the agent's prior. The difference

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

between  $\pi^*$  and  $\pi^{(0)}$  equals the decline in the agent's posterior expected loss from knowing how the states relate and using that knowledge to collect information optimally. Accordingly, we call  $\Pi$  the "value of conceptual knowledge." It depends on the eigenvalues  $\lambda_1, \dots, \lambda_K$  of the "true" prior variance matrix  $\Sigma$  and on the precision parameter  $\tau$  (which jointly determine  $R^*$ ). We characterize this dependence in Theorems 1 and 2.

**Theorem 1.** *The value  $\Pi$  of conceptual knowledge*

- (i) *is non-negative,*
- (ii) *equals zero when  $\lambda_1, \dots, \lambda_K$  are equal, and*
- (iii) *does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.*

**Theorem 2.** *There is a finite threshold  $\tau' \geq 0$  such that  $\Pi$  is increasing in  $\tau$  if and only if  $\tau < \tau'$ . This threshold equals zero if and only if  $\lambda_1, \dots, \lambda_K$  are equal. Moreover,*

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

Theorem 1 says that conceptual knowledge is more valuable when states are more reducible. If a few features contribute most of states' prior variances, then the agent gains a lot from identifying those features and collecting data on them (i.e., "asking the right questions"). In contrast, if every feature contributes equally, then he gains nothing from identifying those features because he collects the same optimal sample that he would collect if he was naïve.

Theorem 2 says that the value of conceptual knowledge is non-monotone in  $\tau$ , which indexes the agent's access to information. This is because raising  $\tau$  has two effects:

1. it gives the agent more information about the features on which he collects data, raising the gain from identifying the most influential features and collecting data on them;
2. it leads the agent to collect data on more features (i.e., it raises  $R^*$ ), lowering the gain from identifying the most influential features.

The first effect dominates the second precisely when  $\tau < \tau'$ . The threshold  $\tau'$  equals zero if and only if  $\lambda_1, \dots, \lambda_K$  are equal, in which case conceptual knowledge has no value because it does not change the optimal sample from what a naïve agent would collect.

Theorem 2 also says that the value of conceptual knowledge vanishes as the agent's access to information grows without bound. This is because the agent's posterior becomes less dependent on his prior as  $\tau$  grows and is independent in the limit as  $\tau \rightarrow \infty$ . Thus, intuitively, having access to unlimited data washes out the benefit of knowing what data to collect. However, this relies on the agent having unrestricted access: he must be able to construct a sample  $\mathcal{S} = \{(w^{(n)}, y^{(n)})\}_{i=1}^n$  such that the covariates  $w^{(1)}, \dots, w^{(n)}$  span the  $K$ -dimensional Euclidean space containing the state vector  $\theta$ . If the covariates do not span  $\mathbb{R}^K$ , then  $\mathcal{S}$  may contain no information about some components of  $\theta$ , and so 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 Appendix Section A2.3.

As an illustration of Theorems 1 and 2, consider the prior variance matrix (7) constructed in Example 1. Its eigenvalues are equal when  $\rho = 0$  and undergo a MPS when  $\rho \in [0, 1)$  rises. So, by Theorem 1, the value  $\Pi$  of conceptual knowledge equals zero when  $\rho = 0$  and is non-decreasing in  $\rho$ . Moreover, by Theorem 2, there is a threshold  $\tau' \geq 0$  such that  $\Pi$  is increasing in  $\tau$  if and only if  $\tau < \tau'$ . We characterize this threshold below. It equals zero when  $\rho = 0$  and rises when  $\rho$  rises, consistent with Theorem 2.

**Proposition 3.** *Suppose  $\theta$  has prior variance (7) with  $\sigma^2 > 0$  and  $\rho \in [0, 1)$ . Then  $\Pi$*

- (i) *equals zero when  $\rho = 0$ ,*
- (ii) *is increasing in  $\rho$ , and*
- (iii) *is increasing in  $\tau$  if and only if*

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

Whereas Theorem 1 implies  $\Pi$  is non-decreasing in  $\rho$ , Proposition 3 says  $\Pi$  is increasing in  $\rho$ . This is because Theorem 1 holds for an arbitrary MPS, which may not affect the largest  $R^*$  eigenvalues and thus may not change the value (16) of an optimal sample. But this is impossible for the MPS induced by raising  $\rho$ , which raises the largest eigenvalue  $\lambda_1 = (1 + \rho(K - 1))\sigma^2$  of (7).

## 6 Deeper knowledge

Now we study the value of “deepening” the agent's conceptual knowledge. We model this process as follows. Suppose the agent knows the trace

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

of the true prior variance matrix  $\Sigma$ , its largest  $J \in \{0, 1, \dots, K\}$  eigenvalues, and the corresponding eigenvectors, but does not know the smallest  $(K - J)$  eigenvalues or the corresponding eigenvectors. Intuitively, he knows how much the features  $v_1, \dots, v_K$  influence the states  $\theta_1, \dots, \theta_K$  overall, and he knows the  $J$  most influential features and how influential they are, but he does not know the  $(K - J)$  least influential features or how influential they are. He assumes the latter features are equally influential. That is, he assumes  $\theta$  has prior variance

$$\Sigma^{(J)} = \sum_{k \leq J} \lambda_k v_k v_k^T + \lambda_K^{(J)} \left( I_K - \sum_{k \leq J} v_k v_k^T \right), \quad (21)$$

where

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

is the mean of the smallest  $(K - J)$  eigenvalues of  $\Sigma$ .<sup>24</sup> The matrix (21) has the same trace as  $\Sigma$  but (possibly) different eigenvalues; its  $k^{\text{th}}$  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  $\Sigma$  if and only if  $k \leq J$ . The eigenvalues of  $\Sigma^{(J)}$  have mean

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

independently of  $J$ . Likewise  $\lambda_K^{(0)} = \bar{\lambda}$  by definition. Thus  $\Sigma^{(0)} = \bar{\lambda} I_K$  is the prior variance matrix used by the “naïve” agent described in Section 3.2. The parameter  $J$  interpolates between  $\Sigma^{(0)}$  and  $\Sigma^{(K)} = \Sigma$ . It captures the “depth” of the agent’s conceptual knowledge: raising  $J$  corresponds to knowing more of the structural features that influence states.<sup>25</sup> Accordingly, we say the agent has “ $J$ -deep conceptual knowledge” if his prior on  $\theta$  has variance  $\Sigma^{(J)}$ .

Suppose the agent has  $J$ -deep conceptual knowledge and collects an optimal sample. Then, by analogy to (15) and (16), this sample has rank

$$R^{(J)} \equiv \max \left\{ k \in \{1, \dots, K\} : \bar{\lambda} \left( \frac{k}{\lambda_k^{(J)}} - \sum_{j=1}^k \frac{1}{\lambda_j^{(J)}} \right) \leq \tau \right\}$$

<sup>24</sup>If  $J = K$ , then we define  $\lambda_K^{(J)} \equiv \lambda_K$ .

<sup>25</sup>Since  $\lambda_1 \geq \dots \geq \lambda_K$  (by assumption), there are non-increasing returns to knowing more features (i.e., increasing  $J$ ), since each additional feature contributes a non-increasing share of the states’ prior variances. This captures the intuitive idea that the agent acquires conceptual knowledge by learning about the most influential features first. For example, he could gain knowledge by taking classes or reading textbooks that provide “high-level summaries” before “digging into the details.”

and value

$$\pi^{(J)} \equiv \frac{\bar{\lambda}}{K} \left( \sum_{k=1}^{R^{(J)}} \frac{\lambda_k^{(J)}}{\bar{\lambda}} - \left( R^{(J)} \right)^2 \left( \sum_{k=1}^{R^{(J)}} \frac{\bar{\lambda}}{\lambda_k^{(J)}} + \tau \right)^{-1} \right),$$

where  $\tau \equiv n\bar{\lambda}/\sigma_u^2$  is the precision index defined in Section 5. For example, letting  $J = 0$  yields the rank  $R^{(0)} = K$  and value

$$\pi^{(0)} = \frac{\bar{\lambda}\tau}{K + \tau}$$

of an optimal sample collected by a naïve agent. 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.” We characterize the relationship between  $R^{(J)}$  and  $J$  in Lemma 4, and the relationship between  $\Pi^{(J)}$  and  $J$  in Theorem 3.

**Lemma 4.** *There is a threshold  $J' \in \{0, \dots, K\}$  such that*

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

for each  $J \in \{0, \dots, K\}$ . This threshold is non-decreasing in  $\tau$ .

**Theorem 3.** *The value  $\Pi^{(J)}$  of  $J$ -deep conceptual knowledge*

- (i) *is non-negative,*
- (ii) *equals zero when  $J = 0$ ,*
- (iii) *is non-decreasing in  $J$ , and*
- (iv) *equals  $\Pi$  when  $J \geq R^*$ .*

Theorem 3 says that deeper knowledge is (weakly) more valuable. Intuitively, the more the agent knows about states’ features, the better he can design samples that provide information about the most influential features.

The value of  $J$ -deep conceptual knowledge is bounded above by the value  $\Pi^{(K)} = \Pi$  of “full” knowledge. It attains this bound when  $J \geq R^*$ . Thus, the agent gains no additional value from deepening his knowledge beyond depth  $R^*$ . This is because he ignores the least influential ( $K - R^*$ ) features when collecting data (since  $\delta_k^* = 0$  for each  $k > R^*$ ), so learning about those features does not change his optimal sample or the value of the information it contains.

For example, suppose the true prior variance matrix  $\Sigma$  has  $k^{\text{th}}$  largest eigenvalue

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

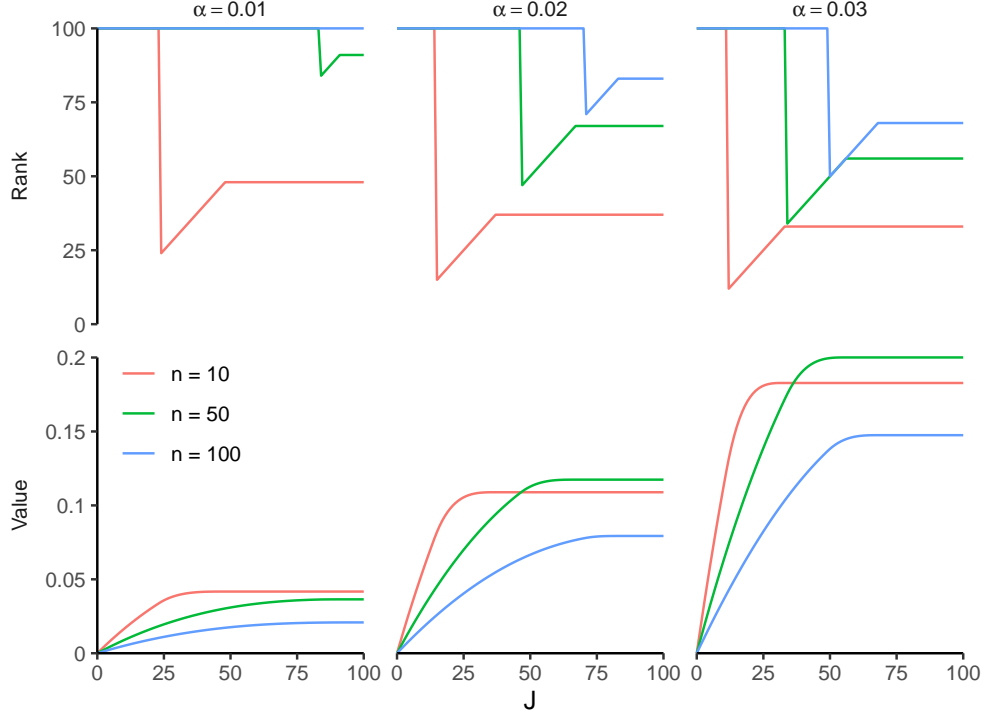


Figure 2: Rank  $R^{(J)}$  and value  $\Pi^{(J)}$  when  $\lambda_{k+1} = (1 - \alpha)\lambda_k$  and  $(K, \bar{\lambda}, \sigma_u^2) = (100, 1, 1)$

with  $0 < \alpha < 1$ . Then  $\lambda_1, \dots, \lambda_K$  are strictly positive, have mean  $\bar{\lambda} = 1$ , are constant in the limit as  $\alpha \rightarrow 0$ , and undergo a MPS as  $\alpha$  rises.<sup>26</sup> This parameter determines the rate

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

at which  $\lambda_k$  decays as  $k$  grows. Intuitively, the larger is  $\alpha$ , the faster features' marginal influences on states diminishes. Thus, if  $\alpha$  is larger, then states are more reducible.

Figure 2 shows how  $R^{(J)}$  and  $\Pi^{(J)}$  depend on  $J$  when  $(K, \bar{\lambda}, \sigma_u^2) = (100, 1, 1)$  and  $\lambda_k$  decays at rate  $\alpha \in \{0.01, 0.02, 0.03\}$ . If  $J$  is sufficiently small, then the agent collects an optimal sample with full rank  $R^{(J)} = 100$ ; otherwise, he collects a sample with rank  $R^{(J)} = \min\{J, R^*\}$ . The threshold depth at which he switches from 100 to  $\min\{J, R^*\}$  rises as the sample size  $n$  rises, consistent with Lemma 4.<sup>27</sup> Intuitively, if the agent has access to more data, then he has less to gain from knowing which features are most influential and focusing on them when he collects data.

<sup>26</sup>For each  $k \in \{1, \dots, K\}$  we have  $\lambda_k \rightarrow 1$  as  $\alpha \rightarrow 0$  by L'Hôpital's rule. Moreover, the partial sum

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

is non-decreasing in  $\alpha$  and is constant in  $\alpha$  when  $k = K$ . Thus, by Lemma 3, the eigenvalues  $\lambda_1, \dots, \lambda_K$  undergo a MPS when  $\alpha$  rises.

<sup>27</sup>Here  $\tau = n$  because  $\bar{\lambda} = \sigma_u^2 = 1$ .

The value  $\Pi^{(J)}$  is increasing in  $J$  when  $J < R^*$  and constant when  $J \geq R^*$ , consistent with Theorem 3. It is increasing in  $\alpha$ , consistent with Theorem 1: conceptual knowledge is more valuable when the eigenvalues  $\lambda_1, \dots, \lambda_K$  are more spread out. Likewise  $\Pi^{(J)}$  is non-monotone in  $n$ , consistent with Theorem 2: raising  $n$  allows the agent to learn more about the “in-sample” features (raising  $\Pi^{(J)}$ ) but also prompts him to expand his sample to include more features (lowering  $\Pi^{(J)}$ ).

## 7 Deeper knowledge or more data?

Finally, we consider the trade-off between deepening the agent’s knowledge and giving him access to more data.<sup>28</sup> Suppose he has  $J$ -deep knowledge and collects an optimal sample of size  $n$ . The value  $\pi^{(J)}$  of this sample indexes his welfare: it is larger when his minimized posterior expected loss is smaller. Lemma 5 says he is better off with deeper knowledge and more data. Intuitively, if he has deeper knowledge, then he can use it to design “better” samples that provide information about the “right” features. If he has access to more data, then he can obtain more information, making his posterior beliefs more accurate and his expected loss lower.

**Lemma 5.** *The value  $\pi^{(J)}$  is*

- (i) *non-decreasing in  $J$  and*
- (ii) *increasing in  $n$ .*

Now suppose the agent has a target value  $\pi_0 \geq 0$ . 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 has deeper knowledge and when states are more reducible:

**Theorem 4.** *Fix  $\pi_0 \geq 0$ . Then  $n_{\pi_0}^{(J)}$*

- (i) *is non-increasing in  $J$  and*
- (ii) *does not rise when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.*

Theorem 4 says that if the agent has deeper knowledge or states are more reducible (holding the depth of his knowledge constant), then he can attain the same welfare with less data. This is because he can design better samples and extract more value from each observation, lowering the number he needs to attain the target  $\pi_0$ .

As in illustration of Theorem 4, suppose the “true” eigenvalues  $\lambda_1, \dots, \lambda_K$  have mean  $\bar{\lambda} = 1$  and decay at rate  $\alpha$  as in Section 6. Figure 3 shows how the minimum sample size  $n_{\pi_0}^{(J)}$  depends on the welfare target  $\pi_0$  and depth  $J$  when  $(K, \sigma_u^2) = (100, 1)$  and  $\alpha \in \{0.01, 0.02, 0.03\}$ . Given  $\pi_0$ , the

---

<sup>28</sup>In contrast, Dominitz and Manski (2017) study the trade-off between having more data and “better” data.



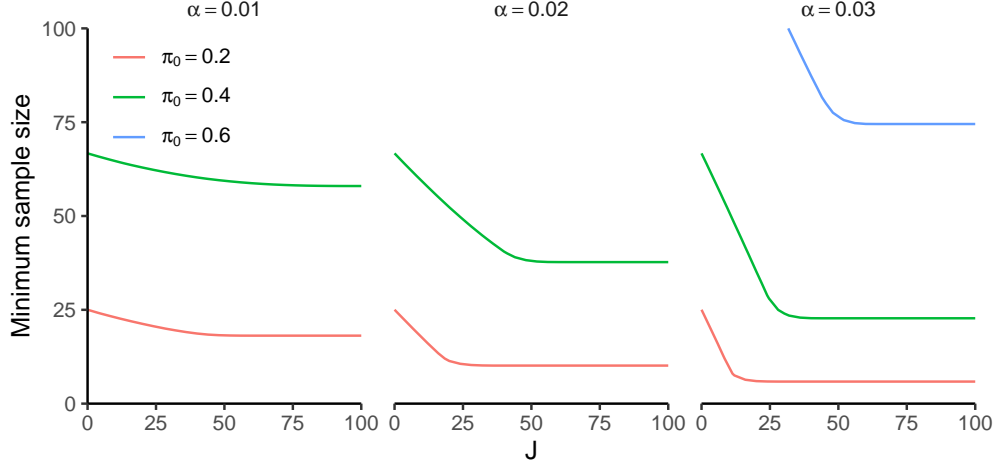


Figure 3: Minimum sample sizes  $n_{\pi_0}^{(J)}$  when  $\lambda_{k+1} = (1 - \alpha)\lambda_k$  and  $(K, \bar{\lambda}, \sigma_u^2) = (100, 1, 1)$

size  $n_{\pi_0}^{(J)}$  is decreasing in  $J$  when  $J < R^*$  and constant in  $J$  when  $J \geq R^*$ . Intuitively, if the agent's knowledge is too shallow, then it constrains his ability to design samples that provide information about the “right” features. Deepening his knowledge relaxes this constraint. It empowers him to design better samples, extract more value from each observation, and require fewer observations to attain  $\pi_0$ . However, once his knowledge is deep enough, deepening it further does not change how he designs samples or the marginal value of each observation. So the only way to collect *more valuable* data is to collect *more* data, thus making  $n_{\pi_0}^{(J)}$  constant in  $J \geq R^*$ .

The curves in Figure 3 are indifference curves: they trace out sets of depth-size pairs  $(J, n)$  that allow the agent to attain different welfare targets  $\pi_0$ . The slope of each curve equals the marginal rate of substitution (hereafter “MRS”) between knowledge and data. Intuitively, this MRS captures the number of observations an additional unit of conceptual knowledge is “worth.” It depends on the depth-size pair  $(J, n)$ , the target  $\pi_0$ , and the parameter  $\alpha$  indexing states' reducibility. Raising this parameter raises the rank  $R^*$  of the optimal sample he would collect if he had full knowledge (see Figure 2). So raising  $\alpha$  can have three effects on the MRS between knowledge and data:

1. If  $J < R^*$  before and after  $\alpha$  rises, then the MRS rises in absolute value;
2. If  $J < R^*$  before  $\alpha$  rises but  $J \geq R^*$  after, then the MRS falls in absolute value (to zero);
3. If  $J \geq R^*$  before and after  $\alpha$  rises, then the MRS remains unchanged (at zero).

Thus, if states are more reducible, then the MRS between knowledge and data may be higher or lower, depending on the depth of the agent's knowledge.

Overall, Figure 3 suggests that if the agent has deeper knowledge, then he can attain the same welfare with less data, especially when states are highly reducible. However, this pattern may be specific to the case in which  $\lambda_1, \dots, \lambda_K$  decay at the constant rate  $\alpha$ . We defer analyzing the general case (in which  $\lambda_1, \dots, \lambda_K$  are distributed arbitrarily) to future research.

## 8 Related literature

We assume states and outcomes are jointly normally distributed under the agent’s prior, and his actions are real-valued and induce quadratic losses. This assumption allows us to derive closed-form expressions for the values of information and conceptual knowledge. It is common in the literature on statistical decisions (Hastie et al., 2009), and in the literature on learning and information acquisition. Within the latter literature, two closely related papers are by Whitmeyer (2024) and Ilut and Valchev (2025).

Whitmeyer (2024) studies the value of information in abstract decision problems. His Theorem 3.1 says that information is more valuable when an agent’s optimal payoff is more convex in their beliefs. This result is analogous to our Theorem 1: an increase in convexity is analogous to eigenvalues undergoing a MPS.

Ilut and Valchev (2025) consider an agent who learns a function via “abstract reasoning” and “integrating experience.” Similarly, we consider an agent who learns a vector by using conceptual knowledge and observing data. 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.

Ilut and Valchev model the cognitive processes that humans use to learn and make decisions. These processes differ from those used by machines: whereas humans can use concepts and causal reasoning, machines currently cannot (without human supervision). Instead they rely on pattern recognition and data-driven prediction (Felin and Holweg, 2024).

This difference between humans and machines motivates papers comparing their predictive performance (e.g., Kleinberg et al., 2018; Kühl et al., 2022; Mullainathan and Obermeyer, 2022). We shed light on *when* humans are likely to outperform “naïve” (in the sense defined in Section 3.2) machines: when the unknowns are highly reducible (see Theorem 1), when data are noisy or scarce (see Theorem 2), and when the sampling frame is limited (see Appendix Section A2.3).

We also identify one reason *why* humans may outperform machines: humans have structural knowledge that empowers us to “ask the right questions.” This knowledge tells us which features of an environment are most important, so we can focus on them when we collect and analyze data. We do not claim this idea as our own: Fessler and Kasy (2019) discuss how structural knowledge can be used to improve econometric estimators; Jackson (2019) discusses how it can guide our experimental designs. However, as far as we know, we are the first to quantify the value of having structural knowledge when learning and making decisions.

Whereas we compare humans and machines implicitly, Iakovlev and Liang (2025) compare them explicitly. They study the “value of context”: how much predictive power one gains from choosing the “right” covariates, relative to an algorithm that cannot make this choice. They show that the value of context vanishes as one’s access to data becomes large. Similarly, we show that the value of conceptual knowledge vanishes as one’s access to data becomes large (see Theorem 2). Together, these results suggest that knowledge and data are substitutes in “big data worlds”: machines can use large, rich samples to learn and make decisions without prior structural knowledge.

However, our Theorem 2 also suggests that knowledge and data are complements in “small data worlds”: humans can make small samples more valuable by leveraging our structural knowledge.

This paper also connects to the literature on model-based learning. Andrews et al. (2025) find that “black box” algorithms outperform models when predicting within domains, but are worse at generalizing across domains.<sup>29</sup> This is consistent with our idea that models embed conceptual knowledge (see Appendix Section A1.2) and that such knowledge boosts out-of-sample predictive performance (see Appendix Section A2.3). Fudenberg et al. (2022) propose a measure of model “completeness”: the share of reducible prediction error that imposing a model reduces. Our paper suggests a different notion of completeness: a model is “more complete” when it embeds deeper conceptual knowledge.<sup>30</sup>

In our analysis of deeper knowledge (see Sections 6 and 7), we assume there exists a “true” model of the agent’s environment (i.e., a true structural relationship among the unknown states) that can be known at different depths. This is in contrast to the literature on model uncertainty and mis-specification, which considers settings where the true model is unknown (e.g., Cerreia-Vioglio et al., 2025; Esponda and Pouzo, 2016; Hansen and Sargent, 2001; Marinacci, 2015).<sup>31</sup> In these settings, many authors advocate a “robust” approach that imposes minimal structural restrictions and provides payoff guarantees across a range of possible models (Gilboa and Schmeidler, 1989; Klibanoff et al., 2005). This is consistent with our “naïve” agent, who lacks conceptual knowledge, assuming a minimally restrictive prior. We add to the model uncertainty and mis-specification literature by quantifying the value of knowing and imposing the correct structural restrictions.

We also assume away any competitive or political forces that may lead the agent to choose one model over another (Dasaratha et al., 2025; Izzo et al., 2023). These forces arise in the literature on “models as narratives” (Aina, 2025; Schwartzstein and Sunderam, 2021; Eliaz et al., 2025). Papers in that literature focus on settings where many models are plausibly “true” (e.g., financial markets and political campaigns). In contrast, many real-world decisions are made in settings where there is an objectively true model one could uncover with enough education and introspection.<sup>32</sup> Such settings are our focus in this paper. For this reason, we do not consider the issues that would arise if the agent’s conceptual knowledge was mis-specified or supplied by a strategic communicator with competing incentives. However, we believe these issues are interesting and worthy of future research.

---

<sup>29</sup>See also Fudenberg and Liang (2019), Peterson et al. (2021), and Peysakhovich and Naecker (2017) for comparisons of model-based and black box predictions.

<sup>30</sup>Mailath and Samuelson (2020) argue that “in practice, people work with models that are deliberately incomplete, including the most salient variables and excluding others.” Indeed, Theorem 3 implies that the agent does not benefit from using models embedding depths greater than the rank  $R^*$  of an optimal sample.

<sup>31</sup>See also Chatfield (1995) for a discussion of model uncertainty in statistics.

<sup>32</sup>For example, there exists an objectively true set of mechanisms through which fertilizers help plants grow. A farmer could discover these mechanisms experimentally or be taught them (see Sankar et al., 2025), or may rely on heuristics that capture some mechanisms but not all (e.g., the farmer may notice that two fertilizers intensify greening, but not know they do so because they supply nitrogen). As another example, a physics student may learn the objectively true model  $E = mc^2$ , then deepen his knowledge by learning about Lorentz factors.

## 9 Conclusion

This paper introduces a simple idea: whereas information is valuable because it leads to better decisions, conceptual knowledge is valuable because it leads to better information. We formalize this idea and study its consequences. Conceptual knowledge is more valuable when the unknown states in an environment are more reducible: when they can be represented as lower-dimensional combinations of high-dimensional features. Its value is non-monotone in the quantity and quality of available data, and vanishes with infinite data. Deeper knowledge is (weakly) more valuable and allows one to attain the same welfare with less data. This trade-off between knowledge and data highlights the importance of interventions that give people knowledge as well as data. It also highlights an important role humans play in human-AI interactions: knowing how to “ask the right questions.”

Stemming from this paper are several avenues for future research. One is to analyze the trade-off between knowledge and data in a consumer choice setting. This would require specifying the “price” of conceptual knowledge vis-à-vis observations of data. Given a price schedule, one could ask many questions about knowledge and data: Are they complements or substitutes? Are they normal or inferior? How do these statuses depend on the states’ reducibility?

Another avenue is to consider competitive and persuasive forces. If there were many ways of interpreting the states’ structural relationships (i.e., many “models” of the agent’s environment), which interpretation would he choose? How could his choice be manipulated by a strategic communicator with competing incentives?

A third avenue is to make our framework dynamic. For example, the agent could take actions that generate outcomes observed by future agents. This would allow us to study how conceptual knowledge is “discovered” and passed down to new generations. This discovery process has been studied by other authors (Carnehl and Schneider, 2025; Gans, 2025); blending their models and insights with ours may bear fruit.

## References

- Aina, C. (2025). Tailored Stories.
- Andrews, I., Fudenberg, D., Lei, L., Liang, A., and Wu, C. (2025). The Transfer Performance of Economic Models.
- 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.
- Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*, 53(5):941–973.
- Bardhi, A. (2024). Attributes: Selective Learning and Influence. *Econometrica*, 92(2):311–353.
- Bartlett, M. S. (1951). An Inverse Matrix Adjustment Arising in Discriminant Analysis. *Annals of Mathematical Statistics*, 22(1):107–111.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.
- Callander, S. (2011). Searching and Learning by Trial and Error. *American Economic Review*, 101(6):2277–2308.
- Carnehl, C. and Schneider, J. (2025). A Quest for Knowledge. *Econometrica*, 93(2):623–659.
- Cerreia-Vioglio, S., Hansen, L. P., Maccheroni, F., and Marinacci, M. (2025). Making Decisions Under Model Misspecification. *Review of Economic Studies*, forthcoming.
- Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, second edition.
- Dasaratha, K., Ortner, J., and Zhu, C. (2025). Markets for Models.
- Davies, B. (2024). Learning about a changing state.
- 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.
- Eliaz, K., Galperti, S., and Spiegler, R. (2025). False Narratives and Political Mobilization. *Journal of the European Economic Association*, 23(3):983–1027.
- Esponda, I. and Pouzo, D. (2016). Berk-Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models. *Econometrica*, 84(3):1093–1130.
- Felin, T. and Holweg, M. (2024). Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, 9(4):346–371.
- Fessler, P. and Kasy, M. (2019). How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions. *Review of Economics and Statistics*, 101(4):681–698.
- Fortiana, J. and Cuadras, C. (1997). A family of matrices, the discretized Brownian bridge, and distance-based regression. *Linear Algebra and its Applications*, 264:173–188.
- Fudenberg, D., Kleinberg, J., Liang, A., and Mullainathan, S. (2022). Measuring the Completeness of Economic Models. *Journal of Political Economy*, 130(4):956–990.
- Fudenberg, D. and Liang, A. (2019). Predicting and Understanding Initial Play. *American Economic Review*, 109(12):4112–4141.

- Gans, J. S. (2025). A Quest for AI Knowledge. NBER Working Paper w33566, National Bureau of Economic Research, Cambridge, MA.
- Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.
- Gollier, C. and Schlesinger, H. (1996). Arrow’s theorem on the optimality of deductibles: A stochastic dominance approach. *Economic Theory*, 7(2):359–363.
- Hansen, L. P. and Sargent, T. J. (2001). Robust Control and Model Uncertainty. *American Economic Review*, 91(2):60–66.
- 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.
- Izzo, F., Martin, G. J., and Callander, S. (2023). Ideological Competition. *American Journal of Political Science*, 67(3):687–700.
- Jackson, M. O. (2019). The Role of Theory in an Age of Design and Big Data. In Laslier, J.-F., Moulin, H., Sanver, M. R., and Zwicker, W. S., editors, *The Future of Economic Design*, pages 523–530. Springer International Publishing, Cham.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics*, 133(1):237–293.
- Klibanoff, P., Marinacci, M., and Mukerji, S. (2005). A Smooth Model of Decision Making under Ambiguity. *Econometrica*, 73(6):1849–1892.
- Kühl, N., Goutier, M., Baier, L., Wolff, C., and Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, 76:78–92.
- Laajaj, R. and Macours, K. (2024). The Complexity of Multidimensional Learning in Agriculture. CEPR Discussion Paper DP19009, CEPR Press, Paris and London.
- Levin, D., Peres, Y., and Wilmer, E. (2008). *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island.
- Liang, A., Mu, X., and Syrgkanis, V. (2022). Dynamically Aggregating Diverse Information. *Econometrica*, 90(1):47–80.
- 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.
- Marinacci, M. (2015). Model Uncertainty. *Journal of the European Economic Association*, 13(6):1022–1100.

- Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *Quarterly Journal of Economics*, 137(2):679–727.
- Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., and Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214.
- Peysakhovich, A. and Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384.
- Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.
- 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.
- Schwartzstein, J. and Sunderam, A. (2021). Using Models to Persuade. *American Economic Review*, 111(1):276–323.
- Whitmeyer, M. (2024). Making Information More Valuable. *Journal of Political Economy*, forthcoming.
- Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

## A Additional material

### A1 Connection to statistical learning

This section connects our paper to the literatures on machine and statistical learning, which study how to derive predictive functions from data.<sup>33</sup> First, we show that our framework (described in Section 3) can be used to study Bayesian learning about real-valued functions. Second, we show how the agent’s prior derives from his “(conceptual) model” of an unknown function.

#### A1.1 Function-state equivalence

Suppose there is a finite set  $\mathcal{X}$  of “inputs” and a square-summable function  $f : \mathcal{X} \rightarrow \mathbb{R}$  belonging to the set

$$\mathcal{F} \equiv \left\{ g \in \mathbb{R}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} (g(x))^2 < \infty \right\}$$

of such functions. Endow  $\mathcal{F}$  with the inner product defined by

$$\langle g, g' \rangle = \sum_{x \in \mathcal{X}} g(x)g'(x)$$

for all pairs  $(g, g') \in \mathcal{F} \times \mathcal{F}$ . Let  $|\mathcal{X}| = K$ , let  $\ell : \mathcal{X} \rightarrow \{1, \dots, K\}$  be a bijection, and define

$$\phi_k(x) \equiv \begin{cases} 1 & \text{if } \ell(x) = k \\ 0 & \text{otherwise} \end{cases}$$

for each  $x \in \mathcal{X}$  and  $k \in \{1, \dots, K\}$ . Then the indicator functions  $\phi_1, \dots, \phi_K$  form an orthonormal basis  $\mathcal{B} \equiv \{\phi_k\}_{k=1}^K$  for the inner product space  $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ . Now let  $\theta_1, \dots, \theta_K$  be the coordinates of  $f$  over  $\{\phi_k\}_{k=1}^K$ :

$$f = \sum_{k=1}^K \theta_k \phi_k.$$

The agent knows  $\phi_1, \dots, \phi_K$  but not  $\theta \equiv (\theta_1, \dots, \theta_K)$ , so learning about  $f$  is equivalent to learning about  $\theta$ .<sup>34</sup> Moreover, suppose the agent draws an input  $x \in \mathcal{X}$  uniformly at random, and predicts the “output”  $y \in \mathbb{R}$  with conditional distribution

$$y \mid x, f \sim \mathcal{N}(f(x), \sigma_u^2)$$

given  $x$  and  $f$ . His “prediction rule”  $\hat{f} \in \mathcal{F}$  maps each realization of  $x$  to a prediction  $\hat{f}(x)$  of  $y$ . This prediction induces a posterior mean squared error (MSE)

$$\mathbb{E}[(y - \hat{f}(x))^2 \mid x, \mathcal{S}] = \mathbb{E}[(f(x) - \hat{f}(x))^2 \mid x, \mathcal{S}] + \sigma_u^2$$

<sup>33</sup>See Bishop (2006) or Hastie et al. (2009) for textbook treatments.

<sup>34</sup>Assuming  $\theta$  is normally distributed is equivalent to assuming  $\{f(x)\}_{x \in \mathcal{X}}$  follows a Gaussian process. Such processes arise in the economic literature on learning and information acquisition—see, e.g., Bardhi (2024), Davies (2024), Ilut and Valchev (2025), or Laajaj and Macours (2024). See also Bishop (2006, Section 6.4) or Rasmussen and Williams (2006) for more information about Gaussian processes and their applications.



given  $x$  and the sample  $\mathcal{S}$ , where  $\mathbb{E}$  takes expectations with respect to the joint prior distribution of input-output pairs. The agent chooses  $\hat{f}$  to minimize the mean posterior MSE across realizations of  $x$ :

$$\hat{f} \in \arg \min_{g \in \mathcal{F}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{E}[(y - g(x))^2 \mid x, \mathcal{S}]. \quad (\text{A1})$$

The optimal actions  $a_1, \dots, a_K$  defined by (3) are precisely the coordinates of  $\hat{f}$  over the basis  $\mathcal{B}$ . By Lemma 3, these coordinates equal the posterior mean coordinates of  $f$ . The minimized mean posterior MSE

$$\min_{g \in \mathcal{F}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{E}[(y - \hat{f}(x))^2 \mid x, \mathcal{S}] = \mathbb{E}[L(\theta, a) \mid \mathcal{S}] + \sigma_u^2$$

equals the expected loss (9) plus a constant  $\sigma_u^2$  that arises due to the irreducible randomness in the outcome  $y$ . Thus, the prediction problem (A1) is equivalent to the choice problem (3).

## A1.2 Conceptual models

Suppose the agent knows about a collection  $\psi_1, \dots, \psi_J \in \mathcal{F}$  of “attributes” that (partially) mediate the relationship between inputs and outputs.<sup>35</sup> These attributes are linearly independent (but not necessary orthogonal) elements of the function space  $\mathcal{F}$ . They map inputs to known, measurable quantities.

The agent uses  $\psi_1, \dots, \psi_J$  to build a “(conceptual) model”  $m \in \mathcal{F}$  that approximates the unknown function  $f$ . This model is a linear combination of attributes: there is a(n unknown) vector  $\beta \equiv (\beta_1, \dots, \beta_J) \in \mathbb{R}^J$  such that

$$m(x) = \sum_{k=1}^J \beta_k \psi_k(x)$$

for each  $x \in \mathcal{X}$ . Then the derivative

$$\frac{\partial m(x)}{\partial \psi_k(x)} = \beta_k$$

of  $m(x)$  with respect to  $\psi_k(x)$  does not depend on the input  $x$ . In this way, the model  $m$  captures the generalizable structure of  $f$  that is common to all inputs. In contrast, the model’s “approximation error”

$$\epsilon \equiv f - m$$

captures the idiosyncrasies specific to each input.

The agent uses his knowledge of  $\psi_1, \dots, \psi_J$  to construct his prior  $\mathbb{P}$  on  $\theta$ . First, he identifies the subspace

$$\mathcal{F}^m \equiv \text{span}\{\psi_1, \dots, \psi_J\}$$

---

<sup>35</sup>This  $J$  is the same as the depth parameter defined in Section 6.

of  $\mathcal{F}$  spanned by the attributes. It corresponds to a subspace

$$\Theta^m \equiv \left\{ \vartheta \in \Theta : \sum_{k=1}^K \vartheta_k \phi_k \in \mathcal{F}^m \right\}$$

of the Euclidean space  $\Theta \equiv \mathbb{R}^K$  containing the unknown coordinate vector  $\theta$ . Concretely, if  $\theta^j \equiv (\theta_1^j, \dots, \theta_K^j)$  contains the (known) coordinates of the  $j^{\text{th}}$  attribute  $\psi_j$  over the orthonormal basis  $\mathcal{B}$ , then

$$\Theta^m = \text{span}\{\theta^1, \dots, \theta^J\}$$

is the subspace of  $\Theta$  spanned by the vectors  $\theta^1, \dots, \theta^J$ .

Next, the agent constructs an orthonormal basis  $\{v_k\}_{k=1}^J$  for  $\Theta^m$  (e.g., by applying the Gram-Schmidt process to  $\theta^1, \dots, \theta^J$ ). If  $J < K$ , then he also constructs an orthonormal basis  $\{v_k\}_{k=J+1}^K$  for the orthogonal complement

$$\Theta^\epsilon \equiv \left\{ \vartheta \in \Theta : \vartheta^T \vartheta' = 0 \text{ for all } \vartheta' \in \Theta^m \right\}$$

of  $\Theta^m$ . Then  $\{v_k\}_{k=1}^K$  is an orthonormal basis for  $\Theta$ . Letting  $\gamma \equiv (\gamma_1, \dots, \gamma_K)$  contain the coordinates of  $\theta$  over  $\{v_k\}_{k=1}^K$  yields (4); the eigendecomposition (5) of  $\Sigma$  follows. Thus, the agent's prior on  $\theta$  derives from his prior on  $\gamma$ , which derives from his knowledge of  $\psi_1, \dots, \psi_J$  (which define the model  $m$ ).

## A2 Value of information

Consider the sample  $\mathcal{S}$ . Proposition A1 says that the value  $\pi(\mathcal{S})$  of  $\mathcal{S}$  is non-negative, grows as  $\mathcal{S}$  grows, and shrinks as  $\sigma_u^2$  grows.

**Proposition A1.** *The value  $\pi(\mathcal{S})$  of the sample  $\mathcal{S}$*

- (i) *is non-negative,*
- (ii) *does not fall when  $\mathcal{S}$  gains observations, and*
- (iii) *falls when  $\sigma_u^2$  rises.*

Sections A2.1–A2.3 discuss the values of samples with specific structures.

### A2.1 Singleton samples

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  $\delta_2 = \dots = \delta_K = 0$ . Substituting them into (13) gives us bounds on the value of  $\mathcal{S}$ :

**Proposition A2.** Suppose  $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$  contains a single observation. Then its value

$$\pi(\mathcal{S}) = \frac{(w^{(1)})^T \Sigma^2 w^{(1)}}{K((w^{(1)})^T \Sigma w^{(1)} + \sigma_u^2)} \quad (\text{A2})$$

satisfies

$$\frac{\lambda_K^2}{K(\lambda_K + \sigma_u^2)} \stackrel{*}{\leq} \pi(\mathcal{S}) \stackrel{**}{\leq} \frac{\lambda_1^2}{K(\lambda_1 + \sigma_u^2)}, \quad (\text{A3})$$

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

The value of  $\{(w^{(1)}, y^{(1)})\}$  is largest when  $w^{(1)}$  is an eigenvector of  $\Sigma$  with corresponding eigenvalue  $\lambda_1 = \max\{\lambda_1, \dots, \lambda_K\}$ . It is smallest when  $w^{(1)}$  is an eigenvector of  $\Sigma$  with corresponding eigenvalue  $\lambda_K = \min\{\lambda_1, \dots, \lambda_K\}$ . Intuitively, the more “weight”  $w^{(1)}$  puts on directions in which the prior variance of  $\theta$  is large, the more valuable it is to observe  $(w^{(1)}, y^{(1)})$  because the larger is the variance reduction it delivers. This is especially true when there are few dimensions (i.e.,  $K$  is small) and when the signal  $y^{(1)}$  is precise (i.e.,  $\sigma_u^2$  is small).

For example, suppose  $\Sigma$  is the matrix (7) constructed in Example 1. Let  $K = 2$  and suppose  $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$  contains a single observation with

$$w^{(1)} = (\sin(\pi t), \cos(\pi t))$$

and  $-1/2 \leq t \leq 1/2$ . Increasing  $t$  from  $-1/2$  to  $1/2$  rotates  $w^{(1)}$  clockwise from  $(-1, 0)$  to  $(1, 0)$ . The value<sup>36</sup>

$$\pi(\mathcal{S}) = \frac{(1 + 2\rho \sin(2\pi t) + \rho^2)\sigma^4}{2((1 + \rho \sin(2\pi t))\sigma^2 + \sigma_u^2)} \quad (\text{A4})$$

of  $\mathcal{S}$  attains its minimum when  $t = -1/4$ , in which case  $w^{(1)} = (-1/\sqrt{2}, 1/\sqrt{2})$  equals the unit eigenvector  $v_2$  of  $\Sigma$  with the smallest corresponding eigenvalue. In contrast, the value of  $\mathcal{S}$  attains its maximum when  $t = 1/4$ , in which case  $w^{(1)} = (1/\sqrt{2}, 1/\sqrt{2})$  equals the unit eigenvector  $v_1$  of  $\Sigma$  with the largest corresponding eigenvalue. Figure A1 shows that  $\pi(\mathcal{S})$  rises monotonically as  $t$  rises from  $-1/4$  to  $1/4$ , which lowers the angle between  $w^{(1)}$  and  $v_1$  from  $90^\circ$  to  $0^\circ$ .

## A2.2 Representative samples

Suppose the covariates  $w^{(1)}, \dots, w^{(n)}$  in  $\mathcal{S}$  are binary vectors: for each  $i \in \{1, \dots, n\}$ , there is an index  $k_i \in \{1, \dots, K\}$  such that  $w^{(i)}$  has  $k_i^{\text{th}}$  component

$$w_k^{(i)} = \begin{cases} 1 & \text{if } k = k_i \\ 0 & \text{otherwise.} \end{cases}$$

Then each outcome

$$y^{(i)} = \theta_{k_i} + u^{(i)}$$

---

<sup>36</sup>We obtain (A4) by substituting  $K = 2$ , the prior variance matrix (7), the covariate  $w^{(1)} = (\sin(\pi t), \cos(\pi t))$ , and the sample size  $n = 1$  into (A2).

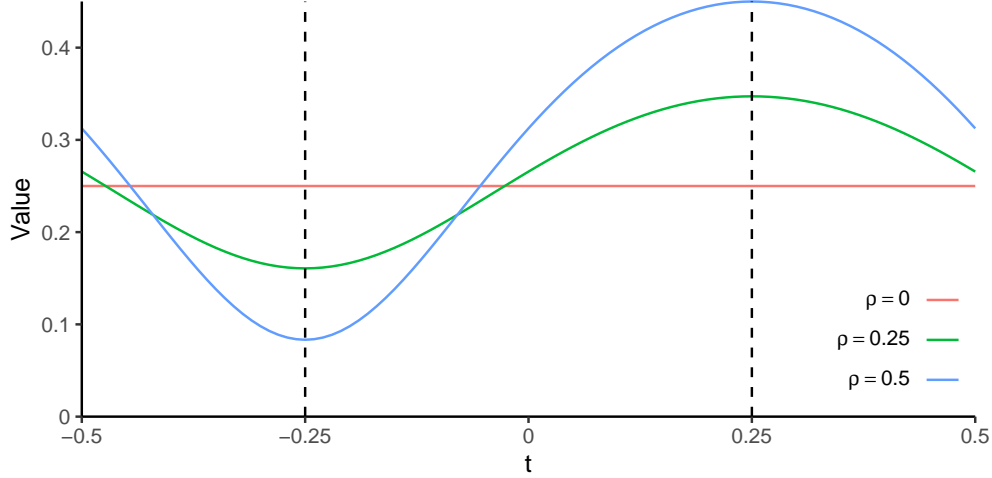


Figure A1: Value (A4) of observing  $\mathcal{S} = \{((\sin(\pi t), \cos(\pi t)), y^{(1)})\}$  when  $\theta$  has prior variance (7) and  $(K, \sigma^2, \sigma_u^2) = (2, 1, 1)$

is a “pure signal” of the state  $\theta_{k_i}$ . Moreover, the Gram matrix  $G$  is diagonal: its  $kk^{\text{th}}$  entry

$$G_{kk} = |\{i \in \{1, \dots, K\} : k_i = k\}|$$

counts the outcomes in  $\mathcal{S}$  that are pure signals of  $\theta_k$ . If  $G_{11}, \dots, G_{KK}$  are equal (to  $n/K$ ), then the eigenvalues of  $G$  are also equal (to  $n/K$ ). These eigenvalues characterize a sample that contains equal information about each state.

Accordingly, we say  $\mathcal{S}$  is “representative” if it induces a Gram matrix with equal eigenvalues. Then the lower and upper bounds in (13) are equal, and so  $\mathcal{S}$  has value

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

This value is larger when the eigenvalues  $\lambda_1, \dots, \lambda_K$  of  $\Sigma$  are more spread out:

**Proposition A3.** *If  $\mathcal{S}$  is representative, then its value  $\pi(\mathcal{S})$  does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.*

If  $\mathcal{S}$  is representative, then it contains equal information about each component of  $\theta$ . But there are diminishing returns to having more information about a given component. So if the prior variances of  $\gamma_1, \dots, \gamma_K$  change in a mean-preserving way, then the increased reduction of the higher variances more than offsets the decreased reduction of the lower variances, thereby raising  $\pi(\mathcal{S})$ .

For example, suppose  $\Sigma$  is the matrix (7) constructed in Example 1. Then the eigenvalues  $\lambda_1 = (1 + \rho(K-1))\sigma^2$  and  $\lambda_2 = \dots = \lambda_K = (1 - \rho)\sigma^2$  of  $\Sigma$  undergo a MPS when  $\rho$  rises. So if  $\mathcal{S}$  is

representative, then its value

$$\pi(\mathcal{S}) = \sigma^2 - \frac{1}{K} \left( \frac{1}{\lambda_1} + \frac{n}{K\sigma_u^2} \right)^{-1} - \left( 1 - \frac{1}{K} \right) \left( \frac{1}{\lambda_K} + \frac{n}{K\sigma_u^2} \right)^{-1}$$

must be non-decreasing in  $\rho$ . Indeed, the derivative

$$\frac{\partial \pi(\mathcal{S})}{\partial \rho} = \left( 1 - \frac{1}{K} \right) \sigma^2 \left( \frac{1}{\lambda_K^2} - \frac{1}{\lambda_1^2} \right) \left( 1 + \frac{n}{K\sigma_u^2} \right)^{-2}$$

of  $\pi(\mathcal{S})$  with respect to  $\rho$  is non-negative because  $\lambda_K \geq \lambda_1$ .

### A2.3 Non-spanning samples

If the sample  $\mathcal{S}$  is *not* representative, then its value can fall when the eigenvalues of  $\Sigma$  undergo a MPS. This happens, for example, when  $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$  is a singleton and  $w^{(1)}$  is an eigenvector of  $\Sigma$  corresponding to an eigenvalue that falls under the MPS.<sup>37</sup> Such a sample is “non-spanning”: the rank

$$R \equiv \max\{k \in \{1, \dots, K\} : \delta_k > 0\}$$

of the Gram matrix  $G$  is strictly less than  $K$ , so there are components of  $\theta$  about which  $\mathcal{S}$  contains no information because they are outside the column space

$$\begin{aligned} \text{col}(G) &\equiv \text{span}\{\omega_1, \dots, \omega_K\} \\ &= \text{span}\{w^{(1)}, \dots, w^{(n)}\} \end{aligned}$$

of  $G$ . The agent cannot learn about these components from  $\mathcal{S}$  directly. But he can learn about them indirectly if he knows how they relate structurally to the components that belong to  $\text{col}(G)$ .

For example, suppose the observations in  $\mathcal{S}$  are pure signals of  $\theta_1, \dots, \theta_R$ . Then

$$\text{col}(G) = \left\{ v \in \mathbb{R}^K : v_k = 0 \text{ for each } k > R \right\}$$

is the subspace of  $\mathbb{R}^K$  spanned by the first  $R$  standard basis vectors. So the first  $R$  components of  $\theta$  are “on-support” but the last  $(K - R)$  components are “off-support.” Let  $\theta_S \equiv (\theta_1, \dots, \theta_R)$  contain the first  $R$  components of  $\theta$  and define

$$\xi_k \equiv \mathbb{V}(\theta_S)^{-1} \begin{bmatrix} \Sigma_{11} \\ \vdots \\ \Sigma_{Rk} \end{bmatrix}$$

for each  $k \in \{1, \dots, K\}$ .<sup>38</sup> Then the posterior variance matrix has trace

$$\text{tr}(\mathbb{V}(\theta \mid \mathcal{S})) = \underbrace{\sum_{k=1}^R \mathbb{V}(\theta_k \mid \mathcal{S}) + \sum_{k>R} \xi_k^T \mathbb{V}(\theta_S \mid \mathcal{S}) \xi_k}_{\text{Sampling error}} + \underbrace{\sum_{k>R} \mathbb{V}(\theta_k \mid \theta_S)}_{\text{Extrapolation error}}, \quad (\text{A5})$$

<sup>37</sup>For example, if  $\theta$  has prior variance (7), then the value of  $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$  is decreasing in  $\rho$  when  $w^{(1)} = v_2$ .

<sup>38</sup>The matrix  $\mathbb{V}(\theta_S)$  is invertible because it is a leading principal submatrix of an invertible matrix.

where

$$\mathbb{V}(\theta_k | \theta_S) = \mathbb{V}(\theta_k) - \tilde{\zeta}_k^T \mathbb{V}(\theta_S) \tilde{\zeta}_k \quad (\text{A6})$$

is the prior variance of  $\theta_k$  left unexplained  $\theta_S$ .<sup>39</sup> The first two terms on the RHS of (A5) are “sampling errors” that depend on how much information  $\mathcal{S}$  contains about the on-support components  $\theta_1, \dots, \theta_R$ . The third term is an “extrapolation error” that depends on how much information these components contain about the off-support components  $\theta_{R+1}, \dots, \theta_K$ . Whereas the sampling error can be reduced by collecting more (or less noisy) data on  $\theta_1, \dots, \theta_R$ , the extrapolation error cannot. It can only be reduced by knowing more about how the on- and off-support components of  $\theta$  relate structurally.

For example, suppose  $\Sigma$  is the matrix (7) constructed in Example 1. Then

$$\mathbb{V}(\theta_k | \theta_S) = \frac{(1 - \rho)(1 + \rho R)\sigma^2}{1 + \rho(R - 1)} \quad (\text{A7})$$

for each  $k > R$  and so the extrapolation error

$$\sum_{k>R} \mathbb{V}(\theta_k | \theta_S) = \frac{(1 - \rho)(1 + \rho R)(K - R)\sigma^2}{1 + \rho(R - 1)}$$

falls as the correlation  $\rho$  rises. It equals  $(K - R)\sigma^2$  when  $\rho = 0$ , in which case  $\theta_1, \dots, \theta_R$  provide no information about  $\theta_{R+1}, \dots, \theta_K$  and so  $\mathbb{V}(\theta_k | \theta_S) = \mathbb{V}(\theta_k) = \sigma^2$  for each  $k > R$ . It equals zero in the limit as  $\rho \rightarrow 1$ , in which case  $\theta_1, \dots, \theta_K$  are fully determined by the coefficient  $\gamma_1$  on their common component and so  $\mathbb{V}(\theta_k | \theta_S) = \mathbb{V}(\theta_k | \gamma_1) = 0$  for each  $k \in \{1, \dots, K\}$ .

If, in addition, the observations in  $\mathcal{S}$  have no noise, then  $\mathbb{V}(\theta_k | \mathcal{S}) = 0$  for each  $k \leq R$  and so  $\mathcal{S}$  has value

$$\begin{aligned} \lim_{\sigma_u^2 \rightarrow 0} \pi(\mathcal{S}) &= \frac{1}{K} \left( \sum_{k=1}^R (\mathbb{V}(\theta_k) - 0) + \sum_{k>R} (\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k | \theta_S)) \right) \\ &= \left( 1 - \frac{(1 - \rho)(1 + \rho R)(K - R)}{(1 + \rho(R - 1))K} \right) \sigma^2. \end{aligned}$$

This value rises as  $\rho$  rises, equals  $R\sigma^2/K$  when  $\rho = 0$ , and equals  $\sigma^2$  in the limit as  $\rho \rightarrow 1$ . Taking the limit as  $K \rightarrow \infty$  gives

$$\lim_{K \rightarrow \infty} \lim_{\sigma_u^2 \rightarrow 0} \pi(\mathcal{S}) = \frac{\rho^2 R \sigma^2}{(1 + \rho(R - 1))},$$

which is bounded away from zero if and only if  $\rho > 0$ . So if there are many states, and the agent has noise-free data but a limited sampling frame, then his sample has value if and only if he knows the states have a common component that explains some of their prior variances.

This example highlights the importance of conceptual knowledge when making out-of-sample predictions. If the agent did not know how the states related structurally (i.e., if  $\rho = 0$ ), then his extrapolation error could be arbitrarily large and the value of his sample could be arbitrarily small.

---

<sup>39</sup>We derive (A5)–(A7) in Appendix B.

Knowing how the states relate allows him to use data on the on-support components of  $\theta$  to learn about the off-support components. This lowers his extrapolation error and ensures his sample has *some* value, even if the sampling frame is limited.

### A3 KL divergences

In Section 3.2, we claim the KL divergence (6) quantifies how much the agent's conceptual knowledge allows him to reduce  $\theta$ . Proposition A4 justifies this claim. It says the KL divergence from the true prior  $\mathbb{P}$  to the naïve prior  $\mathbb{P}^{(0)}$  is (weakly) larger when states are more reducible (i.e., when the eigenvalues  $\lambda_1, \dots, \lambda_K$  of  $\Sigma$  are more spread out).

**Proposition A4.** *The KL divergence from  $\mathbb{P}$  to  $\mathbb{P}^{(0)}$*

- (i) *is non-negative,*
- (ii) *equals zero when  $\lambda_1, \dots, \lambda_K$  are equal, and*
- (iii) *does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.*

For example, suppose  $\Sigma$  is the matrix (7) constructed in Example 1. This matrix has eigenvalues  $\lambda_1 = (1 + \rho(K-1))\sigma^2$  and  $\lambda_2 = \dots = \lambda_K = (1 - \rho)\sigma^2$ , which equal  $\bar{\lambda} = \sigma^2$  when  $\rho = 0$  and undergo a MPS when  $\rho$  rises (see Section 4.5). So, by Proposition A4, the KL divergence

$$\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = -\frac{\ln(1 + \rho(K-1)) + (K-1)\ln(1 - \rho)}{2}$$

from the true prior  $\mathbb{P}$  to the naïve prior  $\mathbb{P}^{(0)}$  must equal zero when  $\rho = 0$  and be non-decreasing in  $\rho$ . Indeed

$$\begin{aligned} \mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) \Big|_{\rho=0} &= -\frac{\ln(1) + (K-1)\ln(1)}{2} \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial}{\partial \rho} \mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) &= \frac{K-1}{2} \left( \frac{1}{1-\rho} - \frac{1}{1+\rho(K-1)} \right) \\ &\geq 0 \end{aligned}$$

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

## B Proofs

### B1 Claims in Section 4

#### B1.1 Proof of Lemma 1

We have

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

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  $k \in \{1, \dots, K\}$ . □

#### B1.2 Proof of Lemma 2

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

**Lemma B1.** *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  $\mu$  and variance  $\Sigma$ . 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  $\mu$  and  $\Sigma$ . If  $\Sigma_{22}$  is invertible, then*

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

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

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

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

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  $\theta$  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 B1 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  $G = W^T W$ . □

### B1.3 Proof of Proposition 1

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

**Lemma B2.** *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  $a_1 \geq \dots \geq a_n$  and  $b_1 \geq \dots \geq b_n$ , and let  $C = A + B$  have eigenvalues  $c_1 \geq \dots \geq c_n$ . 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  $k \in \{1, \dots, n\}$ , with equality when  $k = n$ .

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

*Proof of Proposition 1.* Now

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

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

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

and hence

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

by the orthogonality of  $V$  and the cyclic property of matrix traces. So (13) is equivalent to

$$\sum_{k=1}^K \left( \frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1} \stackrel{\star\star}{\leq} \text{tr} \left( \left( \Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T \right)^{-1} \right) \stackrel{\star}{\leq} \sum_{k=1}^K \left( \frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2} \right)^{-1}. \quad (\text{B1})$$

Now  $\Lambda^{-1}$  is real, symmetric, and positive definite. It has  $k^{\text{th}}$  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  $k^{\text{th}}$  largest eigenvalue  $b_k \equiv \delta_k/\sigma_u^2 \geq 0$ . Define

$$\begin{aligned} c_k^{**} &\equiv a_k + b_{K-k+1} \\ &= \frac{1}{\lambda_{K-k+1}} + \frac{\delta_{K-k+1}}{\sigma_u^2} \\ &> 0 \end{aligned}$$

and

$$\begin{aligned} c_k^* &\equiv a_k + b_k \\ &= \frac{1}{\lambda_{K-k+1}} + \frac{\delta_k}{\sigma_u^2} \\ &> 0 \end{aligned}$$

for each  $k \in \{1, \dots, K\}$ , and consider the matrix  $C \equiv \Lambda^{-1} + B$  with  $k^{\text{th}}$  largest eigenvalue  $c_k$ . This matrix is positive definite and so  $c_k > 0$  for each  $k$ . Moreover, by Lemma B2, we have

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

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

Now define  $g(z) \equiv 1/z$  for all  $z > 0$ . Then  $g : (0, \infty) \rightarrow \mathbb{R}$  is convex. So, by Lemma 3, we have

$$\sum_{k=1}^K \frac{1}{c_k^{**}} \leq \sum_{k=1}^K \frac{1}{c_k} \leq \sum_{k=1}^K \frac{1}{c_k^*}. \quad (\text{B2})$$

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  $c_1^{**}, \dots, c_K^{**}$  and  $c_1^*, \dots, c_K^*$ , and

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

by the definition of  $C$ . Substituting these expressions into (B2) yields (B1), from which (13) follows.

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

Suppose  $\omega_k = v_{K-k+1}$  for each  $k \in \{1, \dots, K\}$ . Then

$$\begin{aligned} Z &= \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix}^T \begin{bmatrix} v_K & \cdots & v_1 \end{bmatrix} \\ &= \begin{bmatrix} & & 1 \\ & \ddots & \\ 1 & & \end{bmatrix} \end{aligned}$$

is the  $K \times K$  anti-diagonal matrix with  $jk^{\text{th}}$  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

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

has trace

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

and thus  $\star$  holds with equality.

Now suppose  $\omega_k = v_k$  for each  $k \in \{1, \dots, K\}$ . Then  $Z$  equals 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

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

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

#### B1.4 Proof of Proposition 2

Consider the constrained minimization problem (14). We can ignore the constraint that  $\delta_k$  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  $\eta \in \mathbb{R}$  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  $(j, k) \in \{1, \dots, K\}^2$ , from which it follows that  $\mathcal{L}$  is convex in the vector  $(\delta_1, \dots, \delta_K)$  whenever it has non-negative components. So if  $\delta_1^*, \dots, \delta_K^*$  solve (14), then they satisfy the first-order conditions (FOCs)

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

the complementary slackness conditions  $0 = \eta_k \delta_k^*$ , and the sum constraint  $\delta_1^* + \dots + \delta_K^* = n$ .

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

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

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

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

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

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

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

and therefore

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

Then the sum constraint implies

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

Thus

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

for each  $k \leq k_0$  and  $\delta_k^* = 0$  for each  $k > k_0$ . Then

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

is non-increasing in  $k_0$  when  $k_0 \leq R^*$ . Thus, the eigenvalues  $\delta_1^*, \dots, \delta_K^*$  defined by (17) solve (14). They are non-increasing because  $\lambda_1, \dots, \lambda_K$  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 (18) holds. □

### B1.5 Proof of Lemma 3

The result follows from establishing three equivalences:

1. (i)  $\iff$  (ii). Rothschild and Stiglitz (1970, Theorem 2) show that (i) is equivalent to  
(ii')  $\int_0^\infty g(z) dF'(z) \geq \int_0^\infty g(z) dF(z)$  for all convex functions  $g : (0, \infty) \rightarrow \mathbb{R}$ ,  
which is equivalent to (ii) by the definitions of  $F$  and  $F'$ .
2. (ii)  $\iff$  (iii). Consider the  $K$ -vectors  $\lambda' \equiv (\lambda'_1, \dots, \lambda'_K)$  and  $\lambda \equiv (\lambda_1, \dots, \lambda_K)$ . Arnold (1987, Theorem 2.9) shows that (ii) holds precisely when  $\lambda'$  majorizes  $\lambda$ . But the components of  $\lambda'$  and  $\lambda$  are non-increasing, and so  $\lambda'$  majorizes  $\lambda$  if and only if (iii) holds.
3. (iii)  $\iff$  (iv). For each  $k \in \{1, \dots, K\}$  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.  $\square$

## B2 Claims in Sections 5–7

### B2.1 Proof of Theorem 1

Our proof of Theorem 1 invokes the following lemma.

**Lemma B3.** *The value  $\pi^*$  of an optimal sample does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.*

*Proof.* Now

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

from the proof of Proposition 2. So if  $\lambda_1, \dots, \lambda_K$  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 (16).

Let  $\lambda'_1 \geq \dots \geq \lambda'_K > 0$  be the eigenvalues after the MPS. By Lemma 3, the difference

$$\eta \equiv \sum_{k=1}^{R^*} \lambda'_k - \sum_{k=1}^{R^*} \lambda_k \tag{B3}$$

is non-negative. The MPS raises the first bracketed term on the RHS of (16) by  $\eta$ . So it suffices to show that the MPS lowers the second bracketed term by at most  $\eta$ :

$$\underbrace{(R^*)^2 \left( \sum_{k=1}^{R^*} \frac{1}{\lambda'_k} + \frac{n}{\sigma_u^2} \right)^{-1}}_{S'} - \underbrace{(R^*)^2 \left( \sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1}}_S \leq \eta. \tag{B4}$$

Consider the first term  $S'$  on the LHS of (B4). 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  $k \in \{1, \dots, K\}$  gives

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

and  $\eta_1 + \dots + \eta_{R^*} = \eta$ . Lemma 3 implies

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

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

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

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

$$\begin{aligned} & \min_{\eta_1, \dots, \eta_{R^*} \in \mathbb{R}} \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k} \\ & \text{subject to } \lambda_k + \eta_k > 0 \text{ for each } k \in \{1, \dots, R^*\}, \\ & \sum_{j=1}^k \eta_j \geq 0 \text{ for each } k \in \{1, \dots, R^*\}, \\ & \text{and } \sum_{k=1}^{R^*} \eta_k = \eta. \end{aligned} \tag{B5}$$

Setting  $\lambda'_k = \lambda_k + \eta_k^*$  for each  $k \in \{1, \dots, R^*\}$  yields the “worst-case” MPS that maximizes the first term  $S'$  on the LHS of (B4) given the difference (B3).

The differences  $\eta_1^*, \dots, \eta_{R^*}^*$  that solve (B5) are non-negative. To see why, notice that  $\eta_1 < 0$  is infeasible and assume towards a contradiction that  $\eta_\ell^* < 0 \leq \min\{\eta_1^*, \dots, \eta_{\ell-1}^*\}$  for some  $\ell > 1$ . Then

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

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

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

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

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

for each  $k \in \{1, \dots, R^*\}$ , from which it follows that  $\eta_1^\dagger, \dots, \eta_{R^*}^\dagger$  are feasible. But  $\lambda_{\ell'} \geq \lambda_\ell$  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

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

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

Finally, we use the non-negativity of  $\eta_1^*, \dots, \eta_{R^*}^*$  to establish the upper bound (B4) on  $(S' - S)$ . Let  $k \in \{1, \dots, R^*\}$  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  $\lambda_k$ . This derivative is non-negative. It is also bounded above by one, since

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

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

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

where  $\star$  uses the maximality of  $S^*$  (induced by the minimality of  $H^*$ ) and  $\star\star$  uses the Lipschitz property. But  $\eta_1^*, \dots, \eta_{R^*}^*$  are non-negative and sum to  $\eta$ , from which the bound (B4) follows:

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

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

Consider (ii). If  $\lambda_1, \dots, \lambda_K$  are equal, then  $\lambda_k = \bar{\lambda}$  for each  $k \in \{1, \dots, K\}$ , 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 the naïve agent's optimal sample depends on  $\lambda_1, \dots, \lambda_K$  via their mean  $\bar{\lambda}$  only. It does not change when  $\lambda_1, \dots, \lambda_K$  undergo a MPS. Since  $\pi^*$  does not fall under the MPS (by Lemma B3), neither does  $\Pi \equiv \pi^* - \pi^{(0)}$ .  $\square$



## B2.2 Proof of Theorem 2

Define

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

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

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

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

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

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

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

Each piece  $\Pi_k$  is continuous in  $\tau$ . Moreover, for each  $k < K$  the difference

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

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

$$\frac{\partial \Pi_k}{\partial \tau} = \frac{\bar{\lambda}}{K} \left( k^2 \left( \sum_{j=1}^k \frac{\bar{\lambda}}{\lambda_j} + \tau \right)^{-2} - \left( \frac{K}{K + \tau} \right)^2 \right) \quad (\text{B6})$$

on each piece  $\Pi_k$ .

Consider the final piece

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

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

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

for all  $\tau \geq 0$ . Whereas if  $\lambda_1, \dots, \lambda_K$  are not equal (i.e., if  $\lambda_1 > \lambda_K$ ), then

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

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

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

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

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

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

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

Suppose  $\tau \in [\tau_k, \tau_{k+1})$  for some  $k < K$ . Then  $\Pi$  is increasing in  $\tau$  if and only if (B6) exceeds zero, which happens precisely when

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

So  $\Pi_k$  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

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

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

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

So defining

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

for each  $k \in \{1, \dots, K\}$  gives  $\tau'_k \in [\tau_k, \tau_{k+1})$  if and only if  $K/\bar{\lambda} \in [\eta_k, \eta_{k+1})$ . But  $\eta_k$  is non-increasing 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  $\lambda_1, \dots, \lambda_K$  are not equal. To see why, notice that

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

is strictly positive when  $\lambda_1 > \bar{\lambda}$ , which holds precisely when  $\lambda_1, \dots, \lambda_K$  are not equal, in which case the value  $\Pi$  is decreasing in  $\tau$  whenever  $\tau > \tau_K$ . So  $\Pi$  is initially increasing in  $\tau$  and eventually decreasing in  $\tau$ , which, by continuity, means its derivative with respect to  $\tau$  changes sign at least once. Therefore, if  $\lambda_1, \dots, \lambda_K$  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.  $\square$

### B2.3 Proof of Proposition 3

Our proof of Proposition 3 invokes the following lemma.

**Lemma B4.** *Suppose  $\theta$  has prior variance (7) with  $\sigma^2 > 0$  and  $\rho \in [0, 1)$ .*

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

$$R^* = \begin{cases} K & \text{if } \rho \leq \rho' \\ 1 & \text{if } \rho > \rho'. \end{cases} \quad (\text{B7})$$

(ii) *The value  $\pi^*$  of an optimal sample rises when  $\rho$  rises.*

*Proof.* Consider (i). If  $\lambda_1 \geq \lambda_2 = \dots = \lambda_K$ , 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 (7) has eigenvalues  $\lambda_1 = (1 + \rho(K-1))\sigma^2$  and  $\lambda_2 = \dots = \lambda_K = (1 - \rho)\sigma^2$ . So  $R^* = K$  if and only if

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

The bracketed term on the RHS is continuous and decreasing in  $\rho$ , strictly positive when  $\rho = 0$ , and unbounded below as  $\rho \rightarrow 1$ . So, by the intermediate value theorem, there exists  $\rho' \in (0, 1)$  such that (B7) holds.

Now consider (ii). Substituting (B7) into (16) gives

$$\begin{aligned} \pi^* &= \frac{1}{K} \begin{cases} \sum_{k=1}^K \lambda_k - K^2 \left( \sum_{k=1}^K \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} & \text{if } \rho \leq \rho' \\ \lambda_1 - \left( \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} \right)^{-1} & \text{if } \rho > \rho' \end{cases} \\ &= \frac{\sigma^2}{K} \begin{cases} K - K^2 \left( \frac{1}{1 + \rho(K-1)} + \frac{K-1}{1-\rho} + \tau \right)^{-1} & \text{if } \rho \leq \rho' \\ 1 + \rho(K-1) - \left( \frac{1}{1 + \rho(K-1)} + \tau \right)^{-1} & \text{if } \rho > \rho', \end{cases} \end{aligned} \quad (\text{B8})$$

which is piecewise increasing in  $\rho$ :

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

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

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

*Proof of Proposition 3.* Suppose the sample  $\mathcal{S}$  is optimal. Then its value  $\pi^*$  equals

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

when  $\rho = 0$ . Now  $\pi^*$  is increasing in  $\rho$  (by Lemma B4), whereas  $\pi^{(0)}$  is constant in  $\rho$ . So  $\Pi = \pi^* - \pi^{(0)}$  equals zero when  $\rho = 0$  and is increasing in  $\rho$ .

It remains to prove (iii). Now (7) has eigenvalues  $\lambda_1 = (1 + \rho(K - 1))\sigma^2$  and  $\lambda_2 = \dots = \lambda_K = (1 - \rho)\sigma^2$ , which have mean  $\bar{\lambda} = \sigma^2$ . Defining

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

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 (16) 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  $\tau$ : differentiating it with respect to  $\tau$  gives

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

which is strictly positive if and only if

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

In contrast, our proof of Theorem 2 shows that the second piece (with  $\tau \geq \tau_K$ ) is non-increasing in  $\tau$ . But  $\tau' \leq \tau_K$ , from which (iii) follows.  $\square$

#### B2.4 Proof of Lemma 4

Define

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

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

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

as in the proof of Theorem 2. Fix  $J \in \{0, \dots, K\}$  and define

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

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

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

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

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

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

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

But if  $J < K$ , then

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

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

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

exists and

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

Clearly  $J'$  is non-decreasing in  $\tau$ . □

### B2.5 Proof of Theorem 3

Now (i) follows from (ii)–(iv), while (ii) follows from the definition of  $\Pi^{(J)} = \pi^{(J)} - \pi^{(0)}$  and (iii) follows from Lemma 5(i). For (iv), suppose  $J \geq R^*$ . Then  $R^{(J)} = R^*$  by Lemma 4, so

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

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

### B2.6 Proof of Lemma 5

We prove (i) and (ii) separately:

- (i) It suffices to show that  $\lambda_1^{(J)}, \dots, \lambda_K^{(J)}$  undergo a MPS when  $J$  rises. Then (i) follows from an argument similar to that used to prove Lemma B3.

Fix  $J < K$ . For each  $k \in \{1, \dots, K\}$  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

$$\begin{aligned} \sum_{j=1}^k \lambda_j^{(J+1)} - \sum_{j=1}^k \lambda_j^{(J)} &= \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ \lambda_{J+1} - \lambda_K^{(J)} + (k - (J+1))(\lambda_K^{(J+1)} - \lambda_K^{(J)}) & \text{if } k > J+1 \end{cases} \\ &= \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ (K-k)(\lambda_K^{(J)} - \lambda_K^{(J+1)}) & \text{if } k > J+1. \end{cases} \end{aligned} \tag{B9}$$

Now

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

because  $\lambda_{J+1} \geq \dots \geq \lambda_K$ . 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 (B9) is non-negative and equals zero when  $k = K$ . Thus, by Lemma 3, the eigenvalues  $\lambda_1^{(J)}, \dots, \lambda_K^{(J)}$  undergo a MPS when  $J$  rises.

(ii) Fix  $J \in \{0, \dots, K\}$  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  $k \in \{1, \dots, K\}$ . Then

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

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  $n$ . □

## B2.7 Proof of Theorem 4

Our proof of Theorem 4 invokes the following Lemma:

**Lemma B5.** Fix  $J \in \{0, \dots, K\}$ . Then  $\lambda_1^{(J)}, \dots, \lambda_K^{(J)}$  undergo a MPS when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.

*Proof.* Fix  $k \in \{1, \dots, K\}$ . By Lemma 3, the cumulative sum

$$\sum_{j=1}^{\min\{k, J\}} \lambda_j$$

does not fall when  $\lambda_1, \dots, \lambda_K$  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 3.  $\square$

*Proof of Theorem 4.* We prove (i) and (ii) separately:

- (i) Fix  $n \geq 0$ . Now  $\pi^{(J)}$  is non-increasing in  $J$  (by Lemma 5), 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  $\lambda_1, \dots, \lambda_K$  undergo a MPS. Then, since  $\pi^{(J)}$  is increasing in  $n$  (by Lemma 5), 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 B3 implies that  $\pi^{(J)}$  does not fall when  $\lambda_1^{(J)}, \dots, \lambda_K^{(J)}$  undergo a MPS, which, by Lemma B5, happens when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.  $\square$

## B3 Claims in Appendix Sections A2 and A3

### B3.1 Proof of Proposition A1

Let  $\mathcal{S}'$  be a superset of  $\mathcal{S}$ . Then

$$\begin{aligned}\mathbb{V}(\theta_k) &= \mathbb{E}[\mathbb{V}(\theta_k | \mathcal{S})] + \mathbb{V}(\mathbb{E}[\theta_k | \mathcal{S}]) \\ &\geq \mathbb{V}(\theta_k | \mathcal{S}) \\ &= \mathbb{E}[\mathbb{V}(\theta_k | \mathcal{S}, \mathcal{S}') | \mathcal{S}] + \mathbb{V}(\mathbb{E}[\theta_k | \mathcal{S}, \mathcal{S}') | \mathcal{S}) \\ &\geq \mathbb{V}(\theta_k | \mathcal{S}, \mathcal{S}') \\ &= \mathbb{V}(\theta_k | \mathcal{S}')\end{aligned}$$

for each  $k \in \{1, \dots, K\}$ , where the first two equalities hold by the law of total variance, the inequalities hold because the posterior variance of  $\theta_k$  is non-negative and non-random (by Lemma 2), and the last equality holds because  $\mathcal{S}'$  is a superset of  $\mathcal{S}$ . It follows that  $0 \leq \pi(\mathcal{S}) \leq \pi(\mathcal{S}')$ , thereby establishing (i) and (ii).

Now consider (iii). Differentiating the posterior variance matrix (10) with respect to  $\sigma_u^2$  gives

$$\begin{aligned}\frac{\partial}{\partial \sigma_u^2} \mathbb{V}(\theta | \mathcal{S}) &= -\left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right) \left(\frac{\partial}{\partial \sigma_u^2} \left[\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right]\right) \left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right) \\ &= \frac{1}{\sigma_u^4} \Sigma^{-1} G \Sigma^{-1} + \frac{2}{\sigma_u^6} G^2 \Sigma^{-1} + \frac{1}{\sigma_u^8} G^3,\end{aligned}$$

which is the sum of three matrices with strictly positive traces. Thus

$$\text{tr}\left(\frac{\partial}{\partial \sigma_u^2} \mathbb{V}(\theta \mid \mathcal{S})\right) > 0$$

and therefore

$$\begin{aligned} \frac{\partial \pi(\mathcal{S})}{\partial \sigma_u^2} &= -\frac{1}{K} \text{tr}\left(\frac{\partial}{\partial \sigma_u^2} \mathbb{V}(\theta \mid \mathcal{S})\right) \\ &< 0 \end{aligned}$$

because traces are linear operators. □

### B3.2 Proof of Proposition A2

Our proof of Proposition A2 uses the following fact about rank-one updates of invertible matrices.

**Lemma B6** (Sherman-Morrison formula). *Let  $n \geq 1$  be an integer, let  $A \in \mathbb{R}^{n \times n}$  be invertible, and let  $u \in \mathbb{R}^n$  and  $v \in \mathbb{R}^n$ . If  $v^T A^{-1} u \neq -1$ , then*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

*Proof.* See Bartlett (1951).

*Proof of Proposition A2.* Suppose  $\mathcal{S}$  contains a single observation and let  $w \equiv w^{(1)}$  for convenience. Then, by Lemmas 2 and B6, we have

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

Thus

$$\begin{aligned} K\pi(\mathcal{S}) &= \text{tr}(\Sigma - \mathbb{V}(\theta \mid \mathcal{S})) \\ &= \text{tr}\left(\frac{\Sigma w w^T \Sigma}{w^T \Sigma w + \sigma_u^2}\right) \\ &= \frac{w^T \Sigma^2 w}{w^T \Sigma w + \sigma_u^2}, \end{aligned}$$

where the last equality holds by the linearity and cyclic property of matrix traces. Equation (A2) follows. The inequalities (A3) follow from Proposition 1, as do the choices of  $w^{(1)}$  that make  $\star$  and  $\star\star$  hold with equality. □

### B3.3 Proof of Proposition A3

Define  $g(z) \equiv nz^2/K(nz + K\sigma_u^2)$  for all  $z > 0$ . Then  $g : (0, \infty) \rightarrow \mathbb{R}$  is convex. So if  $\mathcal{S}$  is representative, then, by Lemma 3, its value

$$\pi(\mathcal{S}) = \sum_{k=1}^K g(\lambda_k)$$

does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.  $\square$

### B3.4 Derivation of (A5)–(A7)

*Derivation of (A5) and (A6).* Let  $k > R$ . Now  $\theta_1, \dots, \theta_R, \theta_k$  are jointly normally distributed under the agent's prior, and so Lemma B1 implies that  $\theta_k$  is conditionally normally distributed with mean

$$\mathbb{E}[\theta_k \mid \theta_S] = \mathbb{E}[\theta_k] + \zeta_k^T (\theta_S - \mathbb{E}[\theta_S]) \quad (\text{B10})$$

and variance (A6) given  $\theta_S$ . So

$$\begin{aligned} \mathbb{V}(\theta_k \mid \mathcal{S}) &= \mathbb{V}(\mathbb{E}[\theta_k \mid \mathcal{S}, \theta_S] \mid \mathcal{S}) + \mathbb{E}[\mathbb{V}(\theta_k \mid \mathcal{S}, \theta_S) \mid \mathcal{S}] \\ &= \mathbb{V}(\mathbb{E}[\theta_k \mid \theta_S] \mid \mathcal{S}) + \mathbb{E}[\mathbb{V}(\theta_k \mid \theta_S) \mid \mathcal{S}] \\ &= \zeta_k^T \mathbb{V}(\theta_S \mid \mathcal{S}) \zeta_k + \mathbb{V}(\theta_k \mid \theta_S), \end{aligned}$$

where the first equality holds by the law of total variance, the second holds because  $\theta_k$  is conditionally independent of  $\mathcal{S}$  given  $\theta_S$ , and the third uses (B10) and the non-randomness of (A6). So

$$\begin{aligned} \text{tr}(\mathbb{V}(\theta \mid \mathcal{S})) &= \sum_{k=1}^K \mathbb{V}(\theta_k \mid \mathcal{S}) \\ &= \sum_{k=1}^R \mathbb{V}(\theta_k \mid \mathcal{S}) + \sum_{k>R} \left( \zeta_k^T \mathbb{V}(\theta_S \mid \mathcal{S}) \zeta_k + \mathbb{V}(\theta_k \mid \theta_S) \right). \end{aligned} \quad \square$$

*Derivation of (A7).* Suppose  $\theta$  has prior variance (7). Then

$$\mathbb{V}(\theta_S) = (\rho \mathbf{1}_R \mathbf{1}_R + (1 - \rho) I_R) \sigma^2$$

has inverse

$$\begin{aligned} \mathbb{V}(\theta_S)^{-1} &= \frac{1}{(1 - \rho) \sigma^2} \left( I_R + \frac{\rho}{1 - \rho} \mathbf{1}_R \mathbf{1}_R^T \right)^{-1} \\ &= \frac{1}{(1 - \rho) \sigma^2} \left( I_R - \frac{\rho}{1 - \rho + \rho \mathbf{1}_R^T \mathbf{1}_R} \mathbf{1}_R \mathbf{1}_R^T \right) \end{aligned}$$

by Lemma B6. Now if  $k > R$ , then

$$\begin{bmatrix} \Sigma_{11} \\ \vdots \\ \Sigma_{Rk} \end{bmatrix} = \rho \sigma^2 \mathbf{1}_R$$

and so

$$\begin{aligned}
\zeta_k^T \mathbb{V}(\theta_S) \zeta_k &= \begin{bmatrix} \Sigma_{11} \\ \vdots \\ \Sigma_{Rk} \end{bmatrix}^T \mathbb{V}(\theta_S)^{-1} \begin{bmatrix} \Sigma_{11} \\ \vdots \\ \Sigma_{Rk} \end{bmatrix} \\
&= (\rho\sigma^2 \mathbf{1}_R)^T \left( \frac{1}{(1-\rho)\sigma^2} \left( I_R - \frac{\rho}{1-\rho + \rho \mathbf{1}_R^T \mathbf{1}_R} \mathbf{1}_R \mathbf{1}_R^T \right) \right) (\rho\sigma^2 \mathbf{1}_R) \\
&= \frac{\rho^2 R \sigma^2}{1 + \rho(R-1)}
\end{aligned}$$

because  $\mathbf{1}_R^T \mathbf{1}_R = R$ . Substituting this expression and  $\mathbb{V}(\theta_k) = \sigma^2$  into (A6) yields (A7).  $\square$

### B3.5 Proof of Proposition A4

It suffices to prove (ii) and (iii), which together imply (i).

If  $\lambda_1, \dots, \lambda_K$  are equal, then  $\lambda_k = \bar{\lambda}$  for each  $k \in \{1, \dots, K\}$  and so

$$\begin{aligned}
\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) &= -\frac{1}{2} \sum_{k=1}^K \ln(1) \\
&= 0,
\end{aligned}$$

thus establishing (ii). For (iii), consider the function  $g : (0, \infty) \rightarrow \mathbb{R}$  defined by

$$g(z) \equiv \frac{\ln(\bar{\lambda}) - \ln(z)}{2}.$$

This function is convex on its domain. So, by Lemma 3, the KL divergence

$$\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = \sum_{k=1}^K g(\lambda_k)$$

from  $\mathbb{P}$  to  $\mathbb{P}^{(0)}$  does not fall when  $\lambda_1, \dots, \lambda_K$  undergo a MPS.  $\square$