Information Overload under Rational Learning*

Jonghyun Park† Sudipta Sarangi‡ Hector Tzavellas§

October 2026

Abstract

We develop a rational theory of information overload in which a larger signal environment can raise the cost of implementing an estimate. An agent observes all signals and minimizes mean squared error plus processing cost. With costless constant estimates and convex, lower semicontinuous costs, the optimal estimate converges in mean square to the prior mean when sustaining nonvanishing estimate variance becomes prohibitively costly. Under these conditions, overload occurs if and only if optimal inference achieves a strict accuracy gain at some finite dimension. We derive affine optimality when the posterior mean is affine and costs are weakly increasing in estimate variance. In a Gaussian–quadratic benchmark with positive processing costs, the optimal rule attenuates the Bayesian posterior mean while preserving relative signal weights. Independent, equally precise signals yield single-troughed mean squared error. In a correlation-neglect application, attenuation makes differences in estimates and true total loss vanish as the environment expands, despite persistent covariance misspecification.

JEL Classifications: D81, D83, D87, D91

Keywords: Information Overload, Cognitive Costs, Affine Learning, Correlation Neglect

*We would like to thank Patrick Allmis, Christophe Bravard, Niloy Bose, Promit Chaudhuri, Sumit Joshi, Matthew Kovach, Ahmed Saber Mahmud, and Gerelt Tserenjigmid for their insightful comments and suggestions.

†Department of Economics, Virginia Tech. E-mail: jongh2419@vt.edu

‡Department of Economics, Virginia Tech. E-mail: ssarangi@vt.edu

§Department of Economics, Virginia Tech. E-mail: hectortz@vt.edu

1 Introduction

Can access to more useful evidence make a rational decision-maker less accurate? Consider a financial analyst forecasting a firm's earnings from a growing set of indicators. Additional indicators can improve prediction, but producing a forecast that responds to the evidence requires resources. Empirical and experimental research documents that larger information sets can alter processing strategies and reduce decision quality across consumer, accounting, and organizational settings (Eppler and Mengis, 2004; Roetzel, 2019). The challenge for economic theory is to explain why an optimizing analyst might lose accuracy gains achieved with a smaller information set, even though the original indicators remain available. This paper studies how an expanding information environment can raise the cost of implementing a responsive forecast and make those earlier accuracy gains too costly to sustain.

Under costless Bayesian inference, additional signals cannot increase minimum expected prediction loss because every rule based on the original signals remains feasible (Blackwell, 1953). When implementing rules is costly, the same reasoning guarantees that minimum total loss, including processing costs, cannot increase if earlier rules remain available at unchanged cost. Our model preserves this statistical feasibility but relaxes the invariance of implementation costs. A rule that ignores the new signals produces exactly the same estimate as before, yet implementing it within a larger information environment may be more expensive. Ignoring additional signals therefore need not preserve an earlier accuracy gain at its original implementation cost.

Experimental evidence suggests that disregarding additional information need not be costless. Payne (1976) shows that more complex decision tasks change information-search strategies and induce greater reliance on selective evaluation. Dummel et al. (2016) show that cues excluded by a simple decision heuristic can nevertheless affect processing, while Chadd et al. (2021) find that irrelevant information increases decision time and generates demand for simpler choice environments. These findings motivate allowing implementation costs to depend on the surrounding information environment. One possible interpretation is that producing a forecast from selected indicators requires isolating those inputs from a growing set of competing cues. We capture such implementation demands through a reduced-form processing technology in which maintaining a response to evidence can become more costly as the information environment expands.

We consider an agent who estimates an unknown state from a fully observed vector of $(d+1)$ signals. She knows their joint distribution with the state and chooses a measurable learning rule ℓ to minimize mean squared error plus processing cost. The signal vectors are nested, so expanding the environment preserves the original evidence. Processing costs may depend

on both the induced estimate and the size of the surrounding signal vector. We call the variance of the induced estimate its *responsiveness*. Our central cost condition requires that any sequence of rules whose responsiveness remains bounded away from zero become arbitrarily costly as the environment expands. Constant estimates remain costless, giving the agent an option that avoids the burden of responding to evidence.

Our main result characterizes information overload, defined as the existence of a finite information environment whose optimal estimate is strictly more accurate than the estimates chosen in every richer environment. Under the cost conditions above and convex, lower semicontinuous processing costs, overload occurs if and only if optimal inference achieves a strict accuracy gain over the prior-mean estimate at some finite dimension. The costless prior-mean estimate bounds optimal total loss by the prior variance. Processing expenditure is therefore bounded, forcing optimal responsiveness to vanish as the environment expands. We show that the optimal estimate then converges in mean square to the prior mean, and its mean squared error returns to the prior variance. This characterization requires neither a parametric signal structure nor a restriction to linear learning rules.

We next characterize how the agent adjusts her inference rule. When the Bayesian posterior mean is affine in the signals and processing costs are weakly increasing in estimate variance, the optimal learning rule is also affine. We then specialize to a Gaussian–quadratic environment in which processing cost is proportional to the number of signals times the variance of the estimate. The optimal rule shrinks the Bayesian posterior mean toward the prior mean by a common attenuation factor. With independent signals of equal precision, additional observations improve Bayesian precision while reducing the fraction of the Bayesian adjustment that the agent implements. When processing costs are positive, mean squared error is single-troughed and converges to the prior variance; its minimum may occur at the smallest feasible information dimension.

The benchmark generates overload while preserving the relative Bayesian weights on the signals. Every signal receives a positive weight at each finite dimension. With heterogeneous signal precisions, the optimal rule continues to multiply all Bayesian weights by the same attenuation factor, leaving their relative magnitudes unchanged. An agent can therefore combine evidence in the correct proportions while losing accuracy gains through attenuation of her overall response. We also characterize the preferred information dimension when the agent can choose the number of available signals. This choice minimizes prediction error plus processing cost, so the preferred dimension need not coincide with the dimension that maximizes statistical accuracy.

Finally, we apply the framework to correlation neglect, in which an agent treats positively

correlated signal noises as independent (Enke and Zimmermann, 2019). She evaluates both prediction error and processing cost under this misspecified covariance model. We compare her estimate with the estimate chosen under the true covariance structure. With positive processing costs, both estimates converge in mean square to the prior mean, and the additional true total loss from correlation neglect vanishes. Without processing costs, the behavioral and loss gaps persist. The convergence reflects reduced responsiveness while covariance beliefs remain misspecified. Similar forecasts can therefore conceal persistent differences in statistical beliefs when both agents respond weakly to evidence.

Related Literature

The paper is closely related to rational explanations of information overload through search and information extraction. Branco et al. (2016) study costly consumer search and show how additional product information can lower the average usefulness of what consumers examine. Lu (2022) develops a rational-inattention model in which more detailed financial reports can make useful information harder to extract; with complementary investment decisions, additional detail can also reduce welfare through a failure to coordinate attention. Blankespoor et al. (2020) review the costs of becoming aware of disclosures, acquiring their content, and analyzing it. Our model takes the signal vector as fully observed and studies the cost of implementing an estimate. We characterize when an expanding information environment eliminates earlier prediction gains and show how overload can arise through attenuation that preserves relative Bayesian signal weights.

The paper also relates to rational inattention more broadly. Sims (2003) studies limits on information processing, while Matějka and McKay (2015) derive discrete-choice behavior from optimal information acquisition under processing costs. Maćkowiak et al. (2023) review this approach and its applications. Deterministic responses arise in Saint-Paul (2017), where an information-capacity constraint leads to piecewise-constant decision rules. We share the premise that limited responsiveness can reflect optimization. Our focus is on a processing technology under which the cost of an unchanged estimate can increase when additional signals enter the environment. Earlier learning rules remain statistically feasible, but retaining them need not preserve their implementation costs.

Our processing technology connects to research on cognitive simplification and rule implementation. Gabaix (2014) derives selective responses through simplified representations that retain important variables. Oprea (2020) provides experimental evidence that implementing decision rules is costly and that these costs depend on the characteristics and representation of the rule. Avoyan and Schotter (2020) show that attention allocated to a game depends on

the other game faced simultaneously, highlighting the influence of the surrounding decision environment. These studies motivate considering implementation demands alongside statistical information. We represent those demands through reduced-form costs that depend on the induced estimate and the size of the information environment.

Empirical research provides further motivation. Studies of consumer choice and financial-distress judgments document settings in which greater information loads are associated with lower decision quality (Chewning Jr and Harrell, 1990; Jacoby et al., 1974; Malhotra, 1982). Chadd et al. (2021) show that irrelevant information increases decision time and generates demand for simpler choice environments. Related choice-overload experiments examine how larger menus affect participation and decision outcomes (Iyengar and Lepper, 2000). Our analysis holds the prediction problem fixed and varies the number of available signals. It provides a rational mechanism consistent with the possibility of declining accuracy, while treating the dimension dependence of processing costs as a modeling assumption rather than a technology identified by these experiments.

Finally, the correlation-neglect application relates to research on the use of dependent evidence. Enke and Zimmermann (2019) document insufficient adjustment for correlation across signals, and Dasaratha and He (2021) find that denser observation networks can impair sequential learning, consistent with neglect of correlation among observed decisions. We take the covariance mistake as given and examine how a separate processing friction changes its consequences. Attenuation reduces differences in realized estimates and true total loss even though covariance beliefs remain different. The application thus identifies a reason why persistent statistical misspecification may become difficult to detect from observed forecasts.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 characterizes information overload and the prior-reversion mechanism. Section 4 establishes affine optimality. Section 5 develops the Gaussian analysis, and Section 6 applies it to correlation neglect. Section 7 concludes.

2 Model

This section develops a model of inference under dimension-dependent processing costs. At each information dimension, the signal vector is given and fully observed. The agent's chooses how to transform these signals into an estimate of the state. We separate the statistical content of the signal vector from the cost of implementing a responsive inference rule. The dimension d indexes the size of the information environment, with $d+1$ signals available at dimension d. Whether a larger environment is more burdensome is determined by the

processing technology introduced below.

2.1 Information Environment

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting a real-valued state θ and a sequence of real-valued signals $(S_k)_{k\geq 0}$. The state has a finite, nondegenerate second moment:

$$\mu_\theta := \mathbb{E}[\theta], \qquad 0 < \sigma_\theta^2 := \mathrm{Var}(\theta) < \infty.$$

At information dimension $d \in \mathbb{N}_0$, the agent observes signals with

$$S^{(d)} := (S_0, S_1, \ldots, S_d)^\top \in \mathbb{R}^{d+1}.$$

Uppercase letters denote random variables, while $s^{(d)} := (s_0, s_1, \ldots, s_d)^\top$ denotes a realization of $S^{(d)}$. Unless otherwise stated, all expectations, variances, and covariances are evaluated under the joint law $\mathbb{P}$ of the state and signals.[1]

We initially impose no parametric restriction on the joint distribution of $(\theta, S^{(d)})$. The signals may be biased, correlated, heterogeneous in quality, redundant, partly informative, or completely uninformative.[2]

The information environments are nested. Every coordinate available at dimension d remains available at every dimension $d' > d$, and an increase in d leaves the joint distribution of the previously available state and signals unchanged. A larger d therefore expands the statistical information available to the agent without removing any previously feasible use of the original coordinates.

2.2 Learning Rules and Processing Costs

At dimension d, a learning rule is a Borel-measurable function

$$\ell : \mathbb{R}^{d+1} \to \mathbb{R}$$

[1] Equivalently, one may begin with a projectively consistent family of finite-dimensional joint laws for $(\theta, S^{(d)})$. The common-probability-space representation makes the nesting of the information environments explicit.

[2] One may interpret S_0 as the agent's own signal and $S_1, \ldots, S_d$ as reports received from additional sources. None of the general results depends on this interpretation.

that maps the realized signal vector into an estimate $\ell(S^{(d)})$ of θ. The set of feasible learning rules is

$$\mathcal{L}^{(d)} := \left\{ \ell : \mathbb{R}^{d+1} \to \mathbb{R} : \ell \text{ is Borel measurable and } \mathbb{E}\big[\ell(S^{(d)})^2\big] < \infty \right\}.$$

Square integrability ensures that both the mean squared error and the variance of the induced estimate are well-defined.

Implementing a learning rule may require cognitive or organizational resources. At dimension d, let

$$C(\ell; d) \in [0, +\infty]$$

denote the processing cost of a rule $\ell \in \mathcal{L}^{(d)}$. The value $+\infty$ allows prohibitive implementation costs. We assume that, at a given dimension, processing costs depend on the induced estimate rather than on the particular functional representation of the rule. Thus, if $\ell, \widetilde{\ell} \in \mathcal{L}^{(d)}$ satisfy

$$\ell(S^{(d)}) = \widetilde{\ell}(S^{(d)}) \quad \text{almost surely,}$$

then $C(\ell; d) = C(\widetilde{\ell}; d)$. We therefore identify learning rules that induce the same estimate almost surely. The economically relevant object is the estimate produced from the observed signals.

An important property of a learning rule is how much its estimate responds to the observed information. We call

$$\mathrm{Var}\big(\ell(S^{(d)})\big)$$

the *responsiveness* of the learning rule. Responsiveness measures how much the estimate varies across signal realizations. A constant rule has zero responsiveness, while an estimate that varies nontrivially with the signals has positive responsiveness. Responsiveness is not itself a measure of accuracy or informativeness: an estimate may vary because it responds to useful information, to statistical noise, or to both.

Because the information environments are nested, every estimate available at a smaller dimension remains statistically feasible at a larger one. Fix $d' > d$ and $\ell \in \mathcal{L}^{(d)}$. Define the lifted rule

$$\ell^{+}_{d'}(s_0, \ldots, s_{d'}) := \ell(s_0, \ldots, s_d).$$

This rule ignores the additional coordinates and satisfies

$$\ell^{+}_{d'}(S^{(d')}) = \ell(S^{(d)}) \quad \text{almost surely.}$$

It therefore produces exactly the same estimate, mean squared error, and responsiveness as the original rule. A larger information environment does not eliminate any previously feasible estimate.

The implementation cost of that estimate, however, need not remain unchanged. Although processing costs depend only on the induced estimate at a fixed dimension, they may also depend directly on the size of the information environment. Consequently, the lifted rule may satisfy

$$C(\ell^{+}_{d'}; d') \neq C(\ell; d).$$

Even when the agent ignores every additional signal, maintaining the same response to the original information may become more expensive. Thus, expanding the information environment preserves the statistical feasibility of earlier learning strategies but need not preserve their implementation costs.

At each dimension d, the agent chooses a learning rule to minimize mean squared error plus processing cost:

$$\inf_{\ell \in \mathcal{L}^{(d)}} \left\{ \mathbb{E}\Big[\big(\ell(S^{(d)}) - \theta \big)^2 \Big] + C(\ell; d) \right\}. \tag{P_d}$$

The first term measures statistical inaccuracy. The second measures the resources required to implement the corresponding degree of responsiveness in an environment of dimension d.

We impose two substantive restrictions on the processing technology. The first provides a costless benchmark that does not respond to the realized signals. The second describes how the cost of sustaining responsiveness behaves as the surrounding information environment expands.

Assumption 1. *For every $d \geq 0$ and every $c \in \mathbb{R}$, if $\ell(s^{(d)}) \equiv c$, then $C(\ell; d) = 0$.*

Assumption 1 gives the agent a costless no-learning option. Under squared-error loss, the best such option is the prior-mean rule $\ell(s^{(d)}) \equiv \mu_\theta$. The agent can therefore avoid all responsiveness costs by reporting the prior mean independently of the realized signals.

Assumption 2. *For every sequence of learning rules $\{\ell_d\}_{d \geq 0}$ with $\ell_d \in \mathcal{L}^{(d)}$, if*

$$\liminf_{d \to \infty} \operatorname{Var}\big(\ell_d(S^{(d)}) \big) > 0,$$

then

$$\lim_{d \to \infty} C(\ell_d; d) = +\infty.$$

Assumption 2 states that nonvanishing responsiveness cannot be sustained at uniformly bounded cost as the information environment becomes arbitrarily large. It does not re-

quire the cost of a fixed rule to increase after every addition of a signal, nor does it require processing costs to be monotone in d at all finite dimensions. It imposes only an asymptotic restriction by stating that every sequence of induced estimates whose variance remains bounded away from zero must eventually become prohibitively costly.

The assumption permits the lifted version of a fixed lower-dimensional rule to become more costly in a larger environment. Depending on the application, implementing a responsive estimate may require maintaining a larger reporting system, reconciling additional inputs, monitoring interactions among variables, verifying that excluded cues remain irrelevant, or allocating attention across a broader set of potential considerations. Even when the final estimate responds directly to only a subset of the coordinates, preserving that response within a larger information system may require additional cognitive or organizational resources. Evidence that larger cue sets alter information-processing strategies and that irrelevant information can increase decision time motivates this reduced-form dimension dependence (Chadd et al., 2021; Dummel et al., 2016; Payne, 1976). Assumption 2 characterizes the consequences of a processing technology under which maintaining a nontrivial response becomes increasingly demanding as the environment expands.

2.3 Well-Posedness

To establish existence and uniqueness of the optimal estimate, fix $d \geq 0$ and let

$$\mathcal{H}_d := L^2\big(\Omega, \sigma(S^{(d)}), \mathbb{P}\big)$$

denote the Hilbert space of square-integrable estimates measurable with respect to the observed signal vector. Its inner product and norm are

$$\langle X, Y \rangle_d := \mathbb{E}[XY], \qquad \|X\|_2 := \sqrt{\mathbb{E}[X^2]}.$$

Every learning rule $\ell \in \mathcal{L}^{(d)}$ induces an estimate $X = \ell(S^{(d)}) \in \mathcal{H}_d$. Conversely, because $S^{(d)}$ takes values in the standard Borel space $\mathbb{R}^{d+1}$, every $X \in \mathcal{H}_d$ can be represented as $X = \ell(S^{(d)})$ for some Borel-measurable function ℓ. Since rules that induce the same estimate almost surely are identified, problem (P_d) is equivalently a choice over $\mathcal{H}_d$.

For $X = \ell(S^{(d)})$, define

$$C_d(X) := C(\ell; d).$$

The objective becomes

$$J_d(X) := \mathbb{E}\big[(X-\theta)^2\big] + C_d(X), \qquad X \in \mathcal{H}_d.$$

Assumption 3. *For every $d \geq 0$, the functional $C_d : \mathcal{H}_d \to [0, +\infty]$ is:*

(i) proper*: $C_d(X) < +\infty$ for at least one $X \in \mathcal{H}_d$;*

(ii) convex*: for every $X, Y \in \mathcal{H}_d$ and every $\alpha \in [0,1]$,*

$$C_d\big(\alpha X + (1-\alpha)Y\big) \leq \alpha C_d(X) + (1-\alpha) C_d(Y);$$

(iii) lower semicontinuous*: if $X_n \to X$ in L^2, then*

$$C_d(X) \leq \liminf_{n\to\infty} C_d(X_n).$$

Properness rules out a processing technology under which every estimate has infinite cost. Under Assumption 1, properness is automatic because every constant estimate has zero cost. Convexity means that combining two induced estimates does not cost more than the corresponding convex combination of their costs. Lower semicontinuity rules out upward jumps in cost at limits of nearby estimates.

The following result shows the existence of a unique estimate of the decision problem.

Proposition 1. *Under Assumption 3, problem (P_d) admits a unique optimal induced estimate*

$$X_d^* = \ell_d^*(S^{(d)}) \in \mathcal{H}_d$$

for every $d \geq 0$, where uniqueness is understood up to almost-sure equality.

We may consequently define the statistical accuracy of optimal inference without selecting among multiple optimal estimates:

$$\mathrm{MSE}(d) := \mathbb{E}\Big[\big(\ell_d^*(S^{(d)}) - \theta\big)^2\Big].$$

A lower value of $\mathrm{MSE}(d)$ represents greater statistical accuracy. This object should be distinguished from the minimized value of problem (P_d), which also includes the processing cost of implementing the optimal rule.

Assumption 1 provides a dimension-independent benchmark. The prior-mean rule satisfies

$$\mathbb{E}\big[(\mu_\theta - \theta)^2\big] + C(\mu_\theta; d) = \sigma_\theta^2.$$

Optimality therefore implies

$$\mathrm{MSE}(d) + C(\ell_d^*; d) \leq \sigma_\theta^2.$$

Because processing costs are nonnegative,

$$\mathrm{MSE}(d) \leq \sigma_\theta^2 \qquad \text{for every } d \geq 0. \tag{1}$$

Thus, the agent can always attain the prior benchmark by declining to respond to the observed signals. The substantive question is whether she chooses a more accurate responsive estimate and whether any such accuracy gain survives as dimensionality increases.

3 Learning Collapse and Overload

Because the implementation cost of a learning rule may depend on ambient dimension, the statistical nesting of the signal vectors does not determine the path of optimal accuracy. A richer environment preserves every lower-dimensional estimate as a mathematical mapping, but it may change the cost of maintaining that estimate or any other responsive rule. This section establishes two implications of the processing technology.

The first is *prior reversion*: under Assumption 2, the optimal estimate converges to the prior mean as dimensionality grows. The second is *information overload*: a finite information dimension produces an estimate that is strictly more accurate than the optimal estimate at every richer dimension. Overload therefore requires both a finite-dimensional gain from learning and the eventual erosion of that gain.

By Proposition 1, the optimal induced estimate is uniquely defined at each dimension under Assumption 3. Recall that

$$\mathrm{MSE}(d) = \mathbb{E}\left[\left(\ell_d^*(S^{(d)}) - \theta\right)^2\right].$$

Information overload is defined in terms of this statistical accuracy, rather than the full objective in problem (P_d).

Definition 1. *Information overload occurs if there exists a finite $d_1 \geq 0$ such that*

$$\mathrm{MSE}(d) > \mathrm{MSE}(d_1)$$

for every $d > d_1$.

Definition 1 requires more than diminishing marginal returns to information. It requires a finite-dimensional environment whose optimal estimate is strictly more accurate than the

optimal estimate in every richer environment. The definition does not require $\mathrm{MSE}(d)$ to increase after each successive addition of a signal. Accuracy may be locally flat or nonmonotone before and after d_1. In the Gaussian-quadratic benchmark later, we obtain the sharper single-troughed path commonly associated with information overload.

The following proposition establishes the resulting limit on optimal inference.

Proposition 2. *Suppose Assumptions 1, 2, and 3 hold. Then*

$$\lim_{d\to\infty} \mathbb{E}\Big[\big(\ell_d^*(S^{(d)}) - \mu_\theta\big)^2\Big] = 0. \tag{2}$$

Consequently,

$$\mathrm{Var}\big(\ell_d^*(S^{(d)})\big) \to 0, \qquad \mathbb{E}\big[\ell_d^*(S^{(d)})\big] \to \mu_\theta,$$

and

$$\lim_{d\to\infty} \mathrm{MSE}(d) = \sigma_\theta^2. \tag{3}$$

The proposition identifies the limiting response to an expanding information environment. The induced estimate converges in mean square to the prior mean. Its variation across signal realizations vanishes, its expectation approaches the prior mean, and its accuracy returns to the level attained by ignoring the signals altogether.

The economic force behind the result is the interaction of the two cost assumptions. Because the prior-mean rule is always available at zero cost, the optimal value of problem (P_d) is bounded above by σ_θ^2 at every dimension. The processing cost paid by the optimal rule is therefore uniformly bounded. Assumption 2 then rules out any sequence of optimal estimates whose responsiveness remains bounded away from zero. Once the variance of the optimal estimate vanishes, the squared-error bound in (1) forces its mean to approach μ_θ.

No parametric signal structure is required for this conclusion. The signals may become increasingly precise or numerous, and the Bayesian posterior based on the full signal vector may remain highly informative. Proposition 2 concerns the estimate the agent optimally implements after accounting for processing costs. The statistical environment remains important for the value of learning at finite dimensions and for whether Assumption 2 is economically appropriate in a particular application.

This convergence result does not, by itself, establish information overload. An agent who always reports the prior mean has no accuracy gain to lose. To distinguish overload from such persistent nonuse of information, we must ask whether responding to the signals improves the accuracy of optimal inference at some finite dimension.

Proposition 3. *Suppose Assumptions 1, 2, and 3 hold. Information overload occurs if and only if there exists a finite $d_0 \geq 0$ such that*

$$\mathrm{MSE}(d_0) < \sigma_\theta^2. \tag{4}$$

Proposition 3 identifies when the loss of responsiveness becomes economically meaningful as information overload. Under the maintained cost assumptions, it is enough that the agent obtains a strict accuracy gain at some finite dimension. She need not initially learn the state with high precision, and the gain need not be large. What matters is that useful inference occurs at a finite scale and that its accuracy advantage cannot be sustained as the surrounding information environment expands.

Condition (4) is stronger than statistical informativeness alone. It evaluates the estimate chosen after the agent trades off accuracy against processing cost. Signals may contain useful information about the state while the cost of implementing a responsive rule makes the prior-mean estimate optimal. The condition therefore distinguishes information that could improve prediction from information that the agent finds worthwhile to use.

Once such a gain exists, additional signals need not be misleading or less informative to generate overload. The original evidence remains available, and new signals may further improve the frictionless Bayesian estimate. What fails to survive is the agent's willingness to maintain an estimate that responds sufficiently to that evidence. A larger information environment can therefore improve the statistical opportunities for inference while reducing the accuracy of the estimate actually implemented. The lost accuracy reflects an optimal response to implementation costs rather than a deterioration in the underlying evidence.

The converse distinguishes overload from an environment in which learning never improves the agent's estimate. For example, suppose the signals never change the conditional mean of the state:

$$\mathbb{E}[\theta \mid S^{(d)}] = \mu_\theta \qquad \text{for every } d.$$

Responding to these signals cannot improve squared-error prediction. Because the prior-mean rule is costless and processing costs are nonnegative, the optimal estimate is the prior mean and $\mathrm{MSE}(d) = \sigma_\theta^2$ at every dimension. There is no information overload because there is no accuracy gain to lose. The same absence of a gain can occur when signals are useful for prediction but too costly to use.

Overload is also a comparison with a smaller information environment, not a claim that the agent performs worse than if she ignored all signals. The costless prior-mean option ensures that $\mathrm{MSE}(d) \leq \sigma_\theta^2$ at every dimension. An agent may remain strictly more accurate than

this benchmark at every finite dimension and still experience overload because the accuracy gain becomes smaller in sufficiently large environments.

The next section derives conditions under which optimal inference can be represented by an affine rule. The Gaussian–quadratic specialization then makes the adjustment in responsiveness explicit, characterizes the path of accuracy, and distinguishes the accuracy-maximizing dimension from the dimension preferred by the agent.

4 Affine Optimality

The preceding section characterized the limiting behavior of optimal inference while allowing the agent to choose any measurable function of the observed signals. We now identify conditions under which this infinite-dimensional problem reduces to a finite-dimensional choice of coefficients. The key question is whether nonlinear transformations of the signal vector can improve the agent's objective once both statistical accuracy and the cost of responsiveness are taken into account.

Under squared-error loss, the Bayesian posterior mean contains all of the signal-dependent variation that is useful for predicting the state. If that posterior mean is affine in the primitive signals, any nonaffine component of an estimate is orthogonal to the statistically relevant affine component. When processing costs are also weakly increasing in the variance of the induced estimate, such a nonaffine component raises both prediction error and responsiveness. It is therefore never optimal.

For the remainder of this section, suppose that

$$\mathbb{E}\big[\|S^{(d)}\|^2\big] < \infty$$

for every $d \geq 0$. Define the subspace of estimates that are affine in the primitive signals by

$$\mathcal{A}^{(d)} := \big\{c + \omega^\top S^{(d)} : \omega \in \mathbb{R}^{d+1},\ c \in \mathbb{R}\big\} \subseteq \mathcal{H}_d.$$

Because $\mathcal{A}^{(d)}$ is the linear span of the finite collection

$$1, S_0, \ldots, S_d,$$

it is finite-dimensional and therefore closed in $\mathcal{H}_d$.

Let

$$m_d := \mathbb{E}[\theta \mid S^{(d)}]$$

denote the Bayesian posterior mean. Under quadratic loss and in the absence of processing costs, m_d is the unique optimal estimate of θ in $\mathcal{H}_d$.

Assumption 4. *For every $d \geq 0$,*

$$m_d = \mathbb{E}[\theta \mid S^{(d)}] \in \mathcal{A}^{(d)}.$$

Assumption 4 requires the statistically optimal predictor to be affine in the primitive signal coordinates. It is satisfied when $(\theta, S^{(d)})$ is jointly Gaussian and, more generally, in environments with a linear conditional expectation. The assumption does not restrict the learning rule chosen by the agent; it restricts only the form of the frictionless Bayesian benchmark.

Our second restriction concerns the processing technology. At a fixed information dimension, estimates that vary more across signal realizations are weakly more costly to implement.

Assumption 5. *For every $d \geq 0$ and every $X, Y \in \mathcal{H}_d$,*

$$\mathrm{Var}(X) \geq \mathrm{Var}(Y) \quad \Longrightarrow \quad C_d(X) \geq C_d(Y).$$

Thus, within a fixed information environment, responsiveness is sufficient for ranking processing costs. The cost may still depend separately on d, so the same degree of responsiveness can have different implementation costs in environments of different dimensions.

Proposition 4. *Suppose Assumptions 3, 4, and 5 hold. Then, for every $d \geq 0$, the unique optimal induced estimate belongs to $\mathcal{A}^{(d)}$. Equivalently, there exist*

$$\omega_d^* \in \mathbb{R}^{d+1} \qquad \text{and} \qquad c_d^* \in \mathbb{R}$$

such that

$$\ell_d^*(S^{(d)}) = (\omega_d^*)^\top S^{(d)} + c_d^*.$$

The induced estimate is unique up to almost-sure equality.

The proposition does not impose affine behavior as a restriction on the agent's strategy set. It derives affine optimality from the interaction between the statistical environment and the processing technology. To see the force of the result, take any candidate estimate and project it onto $\mathcal{A}^{(d)}$. Because the posterior mean belongs to $\mathcal{A}^{(d)}$, the nonaffine residual contains no component that improves prediction of the state. Removing it strictly lowers mean squared error whenever the residual is nonzero. The projection also weakly lowers the variance of the estimate and therefore cannot increase its processing cost. Every genuinely nonaffine estimate is consequently dominated.

Proposition 4 converts the original optimization over all measurable learning rules into a finite-dimensional coefficient problem.

Corollary 1. *Suppose Assumptions 3, 4, and 5 hold. Define*

$$\widetilde{C}_d(\omega, c) := C_d\big(c + \omega^\top S^{(d)}\big).$$

Then problem (P_d) *is equivalent to the finite-dimensional convex program*

$$\min_{(\omega,c)\in\mathbb{R}^{d+1}\times\mathbb{R}} \left\{\mathbb{E}\Big[\big(c + \omega^\top S^{(d)} - \theta\big)^2\Big] + \widetilde{C}_d(\omega, c)\right\}.$$

Every minimizing coefficient pair induces the unique optimal estimate $\ell_d^*(S^{(d)})$.

Corollary 1 shows that all minimizing coefficient pairs generate the same random estimate problem.

Propositions 2–4 establish the paper's general theory. Proposition 2 identifies conditions under which optimal learning collapses toward the prior mean, Proposition 3 characterizes when this collapse constitutes information overload, and Proposition 4 identifies conditions under which optimal inference reduces to affine rules. We next impose a Gaussian signal structure and a quadratic responsiveness cost. Under this environment, we can derive the closed-form coefficients and specify the tradeoff between statistical precision and processing costs explicitly with economic implications.

5 Gaussian Specialization

We now specialize the general theory developed above to a Gaussian–quadratic environment. This specialization yields explicit characterizations of learning collapse, information overload, and optimal responsiveness. The optimal estimate takes the form of a shrinkage transformation of the Bayesian posterior mean, making the tradeoff between statistical precision and processing costs explicit. The resulting expressions also permit a sharp distinction between two economic questions. When the information dimension is exogenous, we study the statistical accuracy of the optimally chosen rule. When the dimension can itself be selected, we study the number of signals that minimizes total loss, including processing costs.

5.1 The Benchmark and the Cognitive Deflator

Fix $d \geq 0$ and write

$$m := d + 1$$

for the number of available signals. We specialize the processing cost to

$$C(\ell; d) := \lambda m \operatorname{Var}\big(\ell(S^{(d)})\big), \qquad \lambda \geq 0. \tag{5}$$

The parameter λ converts responsiveness into processing cost, while the factor m captures the scale of the surrounding information environment. If an induced estimate has variance v, its processing cost is $\lambda m v$. Thus, holding the induced estimate fixed, adding one signal raises its implementation cost by λv.

The specification in (5) satisfies Assumptions 1, 3, and 5. When $\lambda > 0$, it also satisfies Assumption 2: any sequence of estimates whose responsiveness remains bounded away from zero has a processing cost that diverges with m. When $\lambda = 0$, the model reduces to the frictionless Bayesian benchmark.

We impose the following statistical environment.

Assumption 6. *The state satisfies*

$$\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \qquad \sigma_\theta^2 > 0,$$

and the signals satisfy

$$S_k = \theta + \varepsilon_k, \qquad k \geq 0,$$

where $(\varepsilon_k)_{k\geq 0}$ are independent and identically distributed according to

$$\varepsilon_k \sim \mathcal{N}(0, \sigma_\varepsilon^2), \qquad \sigma_\varepsilon^2 > 0,$$

and are independent of θ.

Assumption 6 makes the sample average a sufficient statistic for the state and ensures that the posterior mean is affine. Define

$$\overline{S}^{(d)} := \frac{1}{m} \sum_{k=0}^{d} S_k.$$

The Bayesian posterior mean is

$$\mathbb{E}[\theta \mid S^{(d)}] = \mu_\theta + \beta_m\big(\overline{S}^{(d)} - \mu_\theta\big), \qquad \beta_m := \frac{\sigma_\theta^2 m}{\sigma_\varepsilon^2 + \sigma_\theta^2 m}.$$

The coefficient $\beta_m \in (0, 1)$ is the Bayesian response to the sample average. It increases toward one as m grows because averaging additional independent signals reduces measurement noise.

Proposition 5. *Under Assumption 6 and the processing cost in (5), the unique optimal induced estimate is*

$$\ell_d^*(S^{(d)}) = \mu_\theta + \kappa_m\big(\overline{S}^{(d)} - \mu_\theta\big), \tag{6}$$

where

$$\kappa_m = \underbrace{\frac{1}{1+\lambda m}}_{\text{cognitive deflator}} \quad \underbrace{\frac{\sigma_\theta^2 m}{\sigma_\varepsilon^2 + \sigma_\theta^2 m}}_{\text{Bayesian responsiveness}} = \frac{1}{1+\lambda m}\beta_m. \tag{7}$$

Equivalently,

$$\ell_d^*(S^{(d)}) = (\omega_d^*)^\top S^{(d)} + c_d^*,$$

where

$$\omega_d^* = \alpha_m \mathbf{1}_m, \qquad \alpha_m = \frac{1}{1+\lambda m}\frac{\sigma_\theta^2}{\sigma_\varepsilon^2 + \sigma_\theta^2 m}, \qquad c_d^* = \big(1 - m\alpha_m\big)\mu_\theta.$$

The optimal estimate can be written directly as a deflated posterior mean:

$$\ell_d^*(S^{(d)}) = \mu_\theta + \frac{1}{1+\lambda m}\left(\mathbb{E}[\theta \mid S^{(d)}] - \mu_\theta\right).$$

The term λm therefore reduces the magnitude of the posterior-mean adjustment toward or away from the prior mean.

Equation (7) separates two effects of a larger information environment. The Bayesian responsiveness β_m increases with m, reflecting greater statistical precision. The cognitive deflator $(1+\lambda m)^{-1}$ decreases with m, reflecting the higher cost of maintaining a responsive estimate. When $\lambda = 0$, the optimal rule is the Bayesian posterior mean. When $\lambda > 0$, Bayesian responsiveness converges to one, but the cognitive deflator converges to zero. The overall response κ_m therefore eventually vanishes.

5.2 Accuracy and the Preferred Information Dimension

We first treat d as exogenous and characterize the statistical accuracy of the optimal rule. From (6),

$$\ell_d^*(S^{(d)}) - \mu_\theta = \kappa_m\left(\theta - \mu_\theta + \frac{1}{m}\sum_{k=0}^{d}\varepsilon_k\right).$$

It follows that

$$\mathrm{Var}\big(\ell_d^*(S^{(d)})\big) = \kappa_m^2\left(\sigma_\theta^2 + \frac{\sigma_\varepsilon^2}{m}\right) = \frac{\sigma_\theta^4 m}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)\,(1+\lambda m)^2}. \tag{8}$$

The corresponding mean squared error is

$$\mathrm{MSE}(d) = \big(1-\kappa_m\big)^2\sigma_\theta^2 + \kappa_m^2\frac{\sigma_\varepsilon^2}{m} \tag{9}$$

$$= \sigma_\theta^2 - \frac{\sigma_\theta^4 m(1+2\lambda m)}{(\sigma_\varepsilon^2+\sigma_\theta^2 m)\,(1+\lambda m)^2}. \tag{10}$$

The first term in (9) is the error generated by anchoring the estimate toward the prior mean. The second is the residual measurement error transmitted through the sample average. Averaging reduces the raw noise variance σ_ε^2/m, but the endogenous response κ_m is itself attenuated by the processing cost. The path of accuracy is determined by the interaction of these two forces.

Proposition 6. *Suppose Assumption 6 and the processing cost in (5) hold.*

(i) *If $\lambda = 0$, then*

$$\mathrm{MSE}(d) = \frac{\sigma_\theta^2\sigma_\varepsilon^2}{\sigma_\varepsilon^2+\sigma_\theta^2(d+1)},$$

so $\mathrm{MSE}(d)$ is strictly decreasing in d and converges to zero.

(ii) *If $\lambda > 0$, define the continuous extension*

$$\mathcal{M}(m) := \sigma_\theta^2 - \frac{\sigma_\theta^4 m(1+2\lambda m)}{(\sigma_\varepsilon^2+\sigma_\theta^2 m)\,(1+\lambda m)^2}, \qquad m > 0.$$

There exists a unique $\widehat{m}_{\mathrm{MSE}} > 0$ such that $\mathcal{M}$ is strictly decreasing on $(0, \widehat{m}_{\mathrm{MSE}})$ and strictly increasing on $(\widehat{m}_{\mathrm{MSE}}, \infty)$. Consequently, the discrete sequence $d \mapsto \mathrm{MSE}(d) = \mathcal{M}(d+1)$ is single-troughed, and its set of minimizers consists of either one integer or two adjacent integers. Moreover,

$$\mathrm{MSE}(d) < \sigma_\theta^2 \quad \textit{for every finite } d, \qquad \lim_{d\to\infty}\mathrm{MSE}(d) = \sigma_\theta^2.$$

Information overload therefore occurs whenever $\lambda > 0$.

For every finite d, the optimal responsive estimate is strictly more accurate than the prior mean. The size of this improvement nevertheless converges to zero. Depending on the parameters, accuracy may initially improve before deteriorating, or the feasible sequence may begin on the increasing side of the continuous MSE curve. In either case, some finite dimension is strictly more accurate than every sufficiently large environment.

The limiting behavior in Proposition 6 is an explicit instance of Proposition 2. When $\lambda > 0$,

$$\kappa_m \to 0, \qquad \ell_d^*(S^{(d)}) \to \mu_\theta \quad \text{in } L^2, \qquad \mathrm{MSE}(d) \to \sigma_\theta^2.$$

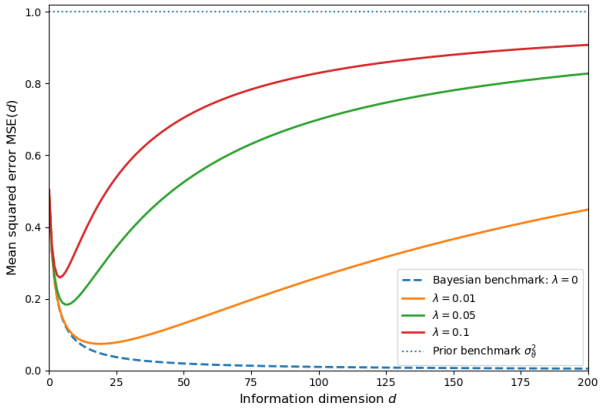


Figure 1: Information overload in the Gaussian–quadratic benchmark

When $\lambda = 0$, by contrast, $\kappa_m = \beta_m \to 1$, and averaging increasingly many independent signals permits arbitrarily accurate learning.

The preceding analysis compares statistical accuracy across exogenously given information dimensions. Suppose instead that the agent may choose d. Because the rule at each dimension already minimizes mean squared error plus processing cost conditional on d, the relevant outer objective is the minimized total loss

$$L(d) := \mathrm{MSE}(d) + \lambda m \,\mathrm{Var}\big(\ell_d^*(S^{(d)})\big)\,. \tag{11}$$

Substituting (8) and (10) gives

$$L(d) = \sigma_\theta^2 - \frac{\sigma_\theta^4 m}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)\,(1+\lambda m)}, \qquad m = d+1. \tag{12}$$

Thus,

$$L(d) - \mathrm{MSE}(d) = \lambda m \,\mathrm{Var}\big(\ell_d^*(S^{(d)})\big) \geq 0.$$

Information overload is defined by the statistical accuracy $\mathrm{MSE}(d)$. The preferred information dimension is instead determined by $L(d)$, which internalizes the resources used to implement the optimal estimate. The dimension minimizing MSE therefore need not coincide with the dimension minimizing total loss.

Proposition 7. *Suppose Assumption 6 and the processing cost in (5) hold.*

(i) *If $\lambda = 0$, then $L(d)$ is strictly decreasing in d.*

(ii) *If $\lambda > 0$, then, for $m = d + 1$,*

$$L(d+1) - L(d) = \frac{\sigma_\theta^4 \left[\lambda \sigma_\theta^2 m(m+1) - \sigma_\varepsilon^2\right]}{\left(\sigma_\varepsilon^2 + \sigma_\theta^2 m\right)\left(1 + \lambda m\right)\left(\sigma_\varepsilon^2 + \sigma_\theta^2 (m+1)\right)\left(1 + \lambda (m+1)\right)}. \tag{13}$$

Consequently, $L(d)$ is single-troughed.

Define

$$x_\lambda := \frac{\sqrt{1 + 4\sigma_\varepsilon^2/(\lambda \sigma_\theta^2)} - 1}{2}. \tag{14}$$

If x_λ is not a positive integer, the unique minimizer is[3]

$$d_L^* = \lceil x_\lambda \rceil - 1.$$

If $x_\lambda = n$ for some positive integer n, the two minimizers are

$$d = n - 1 \qquad \textit{and} \qquad d = n.$$

In particular, $d = 0$ is the unique minimizer if

$$\lambda > \frac{\sigma_\varepsilon^2}{2\sigma_\theta^2},$$

while $d = 0$ and $d = 1$ are tied if equality holds.

Proposition 7 gives the exact solution to the discrete dimension problem. Recall that $d = 0$ corresponds to one observed signal. When processing costs are sufficiently high, the agent therefore chooses the smallest feasible information environment. At lower processing costs, averaging additional observations is worthwhile until the marginal statistical benefit is offset by the higher cost of maintaining responsiveness.

For $\lambda > 0$, the continuous relaxation over the feasible domain $m \geq 1$ is minimized at

$$m_L = \max\left\{1, \sqrt{\frac{\sigma_\varepsilon^2}{\lambda \sigma_\theta^2}}\right\}. \tag{15}$$

A higher signal-noise variance σ_ε^2 weakly increases the preferred number of observations because averaging becomes more valuable. A larger processing-cost coefficient λ reduces the preferred dimension. A higher prior variance σ_θ^2 also reduces the preferred dimension in this benchmark because a responsive estimate then exhibits greater variation and is consequently more expensive under the variance-based processing technology.

[3] $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

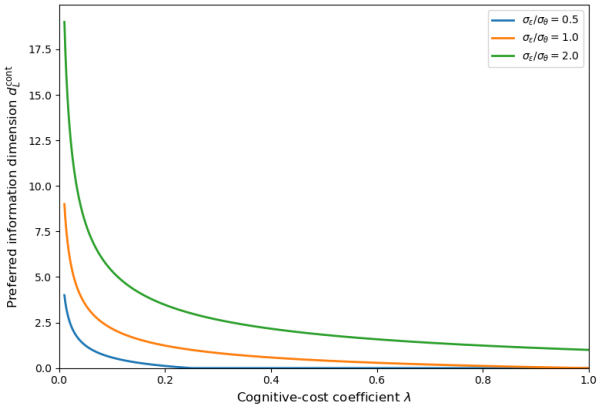


Figure 2: Preferred information dimension versus processing cost

5.3 Heterogeneous Signal Quality

The symmetric benchmark assigns equal weights to all signals because they have identical noise variances. We now allow signal quality to differ across sources. The extension shows that the processing cost governs the overall scale of responsiveness, while the relative allocation of weight across signals continues to reflect their statistical precision.

Assumption 7. *The state satisfies*

$$\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2),$$

and the signals satisfy

$$S_k = \theta + \varepsilon_k, \qquad k \geq 0,$$

where the noise terms are mutually independent, independent of θ, and satisfy

$$\varepsilon_k \sim \mathcal{N}(0, \sigma_{\varepsilon,k}^2), \qquad \sigma_{\varepsilon,k}^2 > 0.$$

For a fixed dimension d, define total signal precision by

$$P_d := \sum_{k=0}^{d} \frac{1}{\sigma_{\varepsilon,k}^2}.$$

Proposition 8. *Under Assumption 7 and the processing cost in (5), the unique optimal rule is*

$$\ell_d^*(S^{(d)}) = (\omega_d^*)^\top S^{(d)} + c_d^*, \qquad c_d^* = \left(1 - (\omega_d^*)^\top \mathbf{1}_m\right) \mu_\theta,$$

where

$$\omega^*_{d,k} = \underbrace{\frac{1}{1+\lambda m}}_{\text{cognitive deflator}} \underbrace{\frac{\sigma^2_\theta}{1+\sigma^2_\theta P_d}\frac{1}{\sigma^2_{\varepsilon,k}}}_{\text{Bayesian weight}}, \qquad k = 0, 1, \ldots, d. \tag{16}$$

Consequently, for any $j, k \in \{0, \ldots, d\}$,

$$\sigma^2_{\varepsilon,k} < \sigma^2_{\varepsilon,j} \quad \Longleftrightarrow \quad \omega^*_{d,k} > \omega^*_{d,j},$$

and

$$\frac{\omega^*_{d,k}}{\omega^*_{d,j}} = \frac{\sigma^2_{\varepsilon,j}}{\sigma^2_{\varepsilon,k}}.$$

As in the symmetric benchmark, the optimal estimate is a deflated version of the Bayesian posterior mean:

$$\ell^*_d(S^{(d)}) = \mu_\theta + \frac{1}{1+\lambda m}\left(\mathbb{E}[\theta \mid S^{(d)}] - \mu_\theta\right).$$

The common factor $(1+\lambda m)^{-1}$ reduces the overall response to the signal vector. It does not distort the relative use of the different signals. More precise signals receive greater weight, and the ratios of the optimal coefficients coincide with the corresponding ratios of Bayesian coefficients.

The exact dimension comparative statics in Propositions 6 and 7 rely on signals of equal quality. With heterogeneous noise variances, information dimension alone no longer summarizes the statistical gain from expanding the signal vector. Adding a highly precise signal can be substantially more valuable than adding a nearly uninformative one. The heterogeneous extension therefore preserves the common-deflator and relative-weighting conclusions, while the paths of MSE and total loss depend on the sequence $\{\sigma^2_{\varepsilon,k}\}_{k\geq 0}$.

6 Application: Correlation Neglect

This section applies the Gaussian–quadratic specialization to correlation neglect: treating information from correlated sources as independent evidence (Dasaratha and He, 2021; Enke and Zimmermann, 2019). We compare the rule chosen by an agent who treats positively correlated signal noises as independent with the optimal rule under the true covariance structure.

6.1 Correlated Signals and Inference Rules

Fix $d \geq 0$ and write $m := d + 1$. Let $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, where $\sigma_\theta^2 > 0$, and suppose $S_k = \theta + \varepsilon_k$ for $k \geq 0$. The noise sequence is jointly Gaussian, centered, and independent of θ, with

$$\mathrm{Var}(\varepsilon_k) = \sigma_\varepsilon^2, \qquad \mathrm{Cov}(\varepsilon_j, \varepsilon_k) = \rho\sigma_\varepsilon^2 \quad \text{for } j \neq k,$$

where $\sigma_\varepsilon^2 > 0$ and $\rho \in [0, 1)$.[4] Let $\overline{S}^{(d)} := m^{-1} \sum_{k=0}^{d} S_k$ and define

$$q_m := \rho + \frac{1 - \rho}{m}. \tag{17}$$

Then

$$\mathrm{Var}\Big(\overline{S}^{(d)} - \theta\Big) = \sigma_\varepsilon^2 q_m, \qquad \mathrm{Var}\Big(\overline{S}^{(d)}\Big) = \sigma_\theta^2 + \sigma_\varepsilon^2 q_m.$$

When $\rho > 0$, averaging eliminates idiosyncratic noise but leaves the common component, so the noise variance of the sample average converges to $\rho\sigma_\varepsilon^2$.

We maintain the quadratic processing cost $C(\ell; d) = \lambda m \,\mathrm{Var}(\ell(S^{(d)}))$, with $\lambda \geq 0$. A correlation-aware agent evaluates both prediction error and processing cost under the true distribution. The sample average is sufficient for θ, and the posterior mean is $\mu_\theta + \beta_m^A(\overline{S}^{(d)} - \mu_\theta)$, where $\beta_m^A := \sigma_\theta^2/(\sigma_\theta^2 + \sigma_\varepsilon^2 q_m)$. The same quadratic optimization as in Proposition 5 gives

$$\ell_d^A(S^{(d)}) = \mu_\theta + \kappa_m^A\big(\overline{S}^{(d)} - \mu_\theta\big), \qquad \kappa_m^A := \frac{1}{1 + \lambda m} \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2 q_m}. \tag{18}$$

A correlation-neglect agent instead sets $\rho = 0$ with a mis-specification when evaluating both mean squared error and the variance entering her processing cost. She uses the same functional form for costs but solves the subjective independent-noise problem. Her optimal rule is, by setting $\rho = 0$ from the above equation,

$$\ell_d^N(S^{(d)}) = \mu_\theta + \kappa_m^N\big(\overline{S}^{(d)} - \mu_\theta\big), \qquad \kappa_m^N := \frac{1}{1 + \lambda m} \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2/m}. \tag{19}$$

The corresponding coefficient vectors are

$$\omega_m^A = \frac{\kappa_m^A}{m} \mathbf{1}_m, \qquad \omega_m^N = \frac{\kappa_m^N}{m} \mathbf{1}_m.$$

Both agents' realized estimates and performance are evaluated below under the true correlated-signal distribution.

[4] For example, let $\varepsilon_k = \sigma_\varepsilon(\sqrt{\rho}\, U + \sqrt{1 - \rho}\, U_k)$, where $U, U_0, U_1, \ldots$ are independent standard normal random variables, independent of θ.

Define the response gap by

$$\Delta_m := \kappa_m^N - \kappa_m^A = \frac{1}{1+\lambda m} \frac{\sigma_\theta^2 \sigma_\varepsilon^2 \rho (m-1)/m}{(\sigma_\theta^2 + \sigma_\varepsilon^2/m)(\sigma_\theta^2 + \sigma_\varepsilon^2 q_m)}. \tag{20}$$

Thus, $\Delta_m \geq 0$, with strict inequality whenever $\rho > 0$ and $m > 1$. Correlation neglect produces excessive responsiveness to the sample average by treating partly redundant observations as independent evidence. Both rules nevertheless contain the common cognitive deflator $(1+\lambda m)^{-1}$, which attenuates the absolute response to the signals.

6.2 Behavioral and Payoff Consequences

Define true total loss by

$$\mathcal{J}_m(\ell) := \mathbb{E}\Big[\big(\ell(S^{(d)}) - \theta\big)^2\Big] + \lambda m \operatorname{Var}\big(\ell(S^{(d)})\big), \tag{21}$$

where expectations and variances are evaluated under the true distribution. The correlation-aware rule ℓ_d^A minimizes $\mathcal{J}_m$, whereas ℓ_d^N minimizes the corresponding subjective objective.

Proposition 9. *Suppose* $\rho \in (0,1)$.[5]

(i) *If* $\lambda > 0$, *then*

$$\ell_d^A(S^{(d)}) \to \mu_\theta \qquad \text{and} \qquad \ell_d^N(S^{(d)}) \to \mu_\theta \quad \text{in } L^2$$

as $d \to \infty$. *Moreover, writing* $m = d+1$,

$$\big\|\ell_d^N(S^{(d)}) - \ell_d^A(S^{(d)})\big\|_2 = O\left(\frac{1}{m}\right), \tag{22}$$

and

$$\big\|\omega_m^N - \omega_m^A\big\|_{\mathrm{E}} = O\left(\frac{1}{m^{3/2}}\right), \tag{23}$$

where $\|\cdot\|_{\mathrm{E}}$ *denotes the Euclidean norm on coefficient vectors. The difference in true total loss satisfies*

$$0 \leq \mathcal{J}_m(\ell_d^N) - \mathcal{J}_m(\ell_d^A) = (1+\lambda m)\left(\sigma_\theta^2 + \sigma_\varepsilon^2 q_m\right)\Delta_m^2 = O\left(\frac{1}{m}\right). \tag{24}$$

(ii) *If* $\lambda = 0$, *then*

$$\kappa_m^A \longrightarrow \frac{\sigma_\theta^2}{\sigma_\theta^2 + \rho\sigma_\varepsilon^2}, \qquad \kappa_m^N \longrightarrow 1.$$

[5]We use standard asymptotic notation: $x_m = O(y_m)$ means that there exist constants $K < \infty$ and M such that $|x_m| \leq K|y_m|$ for every $m \geq M$.

Consequently,

$$\lim_{d\to\infty} \mathbb{E}\Big[\big(\ell_d^N(S^{(d)}) - \ell_d^A(S^{(d)})\big)^2\Big] = \frac{\rho^2\sigma_\varepsilon^4}{\sigma_\theta^2 + \rho\sigma_\varepsilon^2} > 0, \tag{25}$$

and

$$\lim_{d\to\infty} \left[\mathcal{J}_m(\ell_d^N) - \mathcal{J}_m(\ell_d^A)\right] = \frac{\rho^2\sigma_\varepsilon^4}{\sigma_\theta^2 + \rho\sigma_\varepsilon^2} > 0.$$

Part (i) illustrates the prior-reversion mechanism under two different covariance models. The correlation-aware agent continues to discount common noise, while the correlation-neglect agent regards the sample average as increasingly precise. Positive processing costs nevertheless suppress both responses. Equation (22) establishes convergence of the induced estimates under the true distribution. The exact identity in (24) further shows that the additional true total loss from correlation neglect vanishes.

Part (ii) isolates the role of the processing friction. Without processing costs, common noise cannot be averaged away, and both the behavioral gap and the true loss penalty persist. With positive processing costs, correcting the covariance model remains weakly beneficial in terms of true total loss, but the value of that correction converges to zero. This is a consequence of learning collapse. The application therefore illustrates how weak responsiveness can mask persistent statistical misspecification in observed estimates.

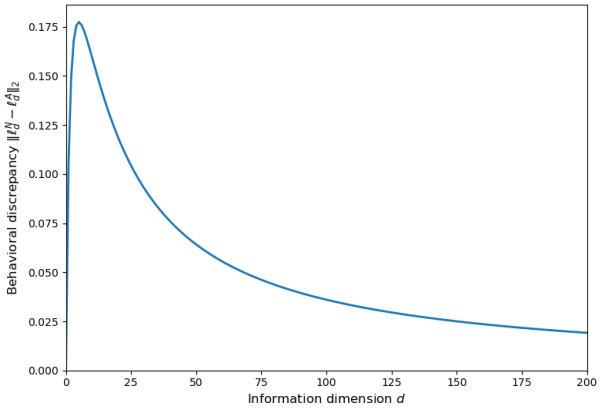


Figure 3: Behavioral gap between correlation-aware and correlation-neglect rules

7 Conclusion

This paper develops a rational theory of information overload. We develop a model in which information overload arises from rational learning under cognitive costs. In our framework, overload emerges endogenously from the trade-off between statistical accuracy and cognitive burden.

Our analysis delivers three main insights. First, we characterize conditions under which learning collapses in high-dimensional environments. When the marginal cost of responsiveness exceeds the marginal value of precision, agents attenuate responsiveness and revert toward prior-based rules. Second, we show that under variance-monotone costs, the optimal learning rule is affine and includes a cognitive deflator that uniformly shrinks responsiveness as dimensionality increases. Finally, this attenuation also changes the consequences of correlation neglect. With positive processing costs, the correctly specified and misspecified estimates converge to one another, and the additional true total loss from the covariance mistake vanishes. Their similarity reflects reduced responsiveness, not a correction of the underlying beliefs.

These results imply that the value of additional information depends not only on its statistical content, but also on the dimensionality of the information environment. Policies that expand information without reducing processing costs may therefore reduce decision quality. In fact, even well-intentioned policies like those aimed at increasing transparency can lower the quality of decisions. This suggests a role for information design that emphasizes simplification and dimensionality reduction, for example, by limiting options or structuring information in specific ways.

A Proofs

Proof of Proposition 1

Fix $d \in \mathbb{N}_0$ and work on the Hilbert space $\mathcal{H}_d$. By Assumption 3(i), there exists $X_0 \in \mathcal{H}_d$ with $C_d(X_0) < +\infty$. Since $X_0, \theta \in L^2$, this implies $J_d(X_0) < +\infty$, so J_d is proper.

Because $C_d \geq 0$,

$$J_d(X) \geq \mathbb{E}[(X-\theta)^2] = \|X-\theta\|_2^2 \geq \left(\|X\|_2 - \|\theta\|_2\right)^2.$$

Thus $J_d(X) \to +\infty$ as $\|X\|_2 \to +\infty$. Let $\{X_n\} \subseteq \mathcal{H}_d$ be a minimizing sequence. Coercivity implies that $\{X_n\}$ is bounded, so, because $\mathcal{H}_d$ is a Hilbert space, some subsequence converges weakly to an element $X^* \in \mathcal{H}_d$. The map $X \mapsto \|X-\theta\|_2^2$ is weakly lower semicontinuous. Moreover, a convex functional that is lower semicontinuous in the norm topology is weakly lower semicontinuous, so Assumption 3(ii)–(iii) implies that C_d is weakly lower semicontinuous. Therefore

$$J_d(X^*) \leq \liminf_{n\to\infty} J_d(X_n) = \inf_{X\in\mathcal{H}_d} J_d(X),$$

and X^* attains the minimum.

To prove uniqueness, suppose X_1 and X_2 are two minimizers that are not equal almost surely. Their objective values, and hence their costs, are finite. For every $\alpha \in (0,1)$,

$$\mathbb{E}\Big[\big(\alpha X_1 + (1-\alpha)X_2 - \theta\big)^2\Big] = \alpha\mathbb{E}[(X_1-\theta)^2] + (1-\alpha)\mathbb{E}[(X_2-\theta)^2] \\ - \alpha(1-\alpha)\|X_1 - X_2\|_2^2.$$

By convexity of C_d,

$$C_d\big(\alpha X_1 + (1-\alpha)X_2\big) \le \alpha C_d(X_1) + (1-\alpha)C_d(X_2).$$

Since $\|X_1 - X_2\|_2^2 > 0$, adding the two displays gives

$$J_d\big(\alpha X_1 + (1-\alpha)X_2\big) < \alpha J_d(X_1) + (1-\alpha)J_d(X_2),$$

contradicting optimality of X_1 and X_2. Hence the optimal induced estimate is unique up to almost-sure equality.

■

Proof of Proposition 2

Fix any selection of optimal rules and set $X_d := \ell_d^*(S^{(d)})$ and $v_d := \mathrm{Var}(X_d)$. By Assumption 1, the prior-mean rule is feasible at zero cost, so optimality gives

$$\mathbb{E}[(X_d-\theta)^2] + C(\ell_d^*; d) \le \sigma_\theta^2$$

for every $d \in \mathbb{N}_0$. In particular,

$$\mathbb{E}[(X_d-\theta)^2] \le \sigma_\theta^2 \qquad \text{and} \qquad C(\ell_d^*; d) \le \sigma_\theta^2.$$

We first prove that $v_d \to 0$. Suppose otherwise. Then there exist $\eta > 0$ and a strictly increasing sequence $\{d_k\}_{k\ge1}$ such that $v_{d_k} \ge \eta$ for every k. For $d \ge d_1$, let $k(d) := \max\{k : d_k \le d\}$ and define the lifted rule

$$\widehat{\ell}_d(s_0, \dots, s_d) := \ell^*_{d_{k(d)}}(s_0, \dots, s_{d_{k(d)}}).$$

For $d < d_1$, set $\widehat{\ell}_d \equiv \mu_\theta$. By nestedness,

$$\widehat{\ell}_d(S^{(d)}) = \ell^*_{d_{k(d)}}(S^{(d_{k(d)})})$$

almost surely for every $d \geq d_1$, and therefore

$$\mathrm{Var}\Big(\widehat{\ell}_d(S^{(d)})\Big) = v_{d_{k(d)}} \geq \eta.$$

Hence $\liminf_{d\to\infty} \mathrm{Var}(\widehat{\ell}_d(S^{(d)})) \geq \eta > 0$, so Assumption 2 requires $C(\widehat{\ell}_d; d) \to +\infty$. Along the subsequence $d = d_k$, however, $\widehat{\ell}_{d_k} = \ell^*_{d_k}$, and thus

$$C(\widehat{\ell}_{d_k}; d_k) = C(\ell^*_{d_k}; d_k) \leq \sigma_\theta^2$$

for every k, a contradiction. Therefore $v_d \to 0$.

Let $b_d := \mathbb{E}[X_d] - \mu_\theta$. Expanding the mean squared error and using the preceding upper bound gives

$$\sigma_\theta^2 + v_d + b_d^2 - 2\,\mathrm{Cov}(X_d, \theta) = \mathbb{E}[(X_d - \theta)^2] \leq \sigma_\theta^2.$$

Consequently,

$$0 \leq v_d + b_d^2 \leq 2\,\mathrm{Cov}(X_d, \theta) \leq 2\,|\mathrm{Cov}(X_d, \theta)| \leq 2\sigma_\theta\sqrt{v_d},$$

where the last inequality is Cauchy–Schwarz. Since $v_d \to 0$, it follows that $b_d \to 0$, and hence

$$\mathbb{E}[(X_d - \mu_\theta)^2] = v_d + b_d^2 \to 0.$$

This proves (2). The variance and mean conclusions follow from $v_d \to 0$ and $b_d \to 0$. Finally, the reverse triangle inequality gives

$$\Big|\sqrt{\mathrm{MSE}(d)} - \sigma_\theta\Big| = |\|X_d - \theta\|_2 - \|\mu_\theta - \theta\|_2| \leq \|X_d - \mu_\theta\|_2 \to 0,$$

so $\mathrm{MSE}(d) \to \sigma_\theta^2$.

■

Proof of Proposition 3

Necessity. Suppose information overload occurs at d_1. Then

$$\mathrm{MSE}(d_1) < \mathrm{MSE}(d_1 + 1) \leq \sigma_\theta^2,$$

where the second inequality follows from (1). Thus $\mathrm{MSE}(d_1) < \sigma_\theta^2$, so (4) holds with $d_0 = d_1$.

Sufficiency. Suppose there exists d_0 such that $\mathrm{MSE}(d_0) < \sigma_\theta^2$, and define

$$\delta := \frac{\sigma_\theta^2 - \mathrm{MSE}(d_0)}{2} > 0.$$

By Proposition 2, there exists an integer $D \geq d_0$ such that

$$\mathrm{MSE}(d) > \sigma_\theta^2 - \delta = \frac{\sigma_\theta^2 + \mathrm{MSE}(d_0)}{2} > \mathrm{MSE}(d_0)$$

for every $d > D$. Let d_1 be the largest minimizer of $\mathrm{MSE}(d)$ over the finite set $\{0, 1, \ldots, D\}$. If $d_1 < d \leq D$, maximality of d_1 implies $\mathrm{MSE}(d) > \mathrm{MSE}(d_1)$. If $d > D$, then

$$\mathrm{MSE}(d) > \mathrm{MSE}(d_0) \geq \mathrm{MSE}(d_1),$$

where the weak inequality follows because $d_0 \leq D$ and d_1 is a minimizer on $\{0, \ldots, D\}$. Therefore $\mathrm{MSE}(d) > \mathrm{MSE}(d_1)$ for every $d > d_1$, which is information overload.

■

Proof of Proposition 4

Fix $d \in \mathbb{N}_0$. Because $\mathcal{A}^{(d)}$ is a closed subspace of $\mathcal{H}_d$, the orthogonal projection $P_{\mathcal{A}^{(d)}} : \mathcal{H}_d \to \mathcal{A}^{(d)}$ is well defined. For any $X \in \mathcal{H}_d$, write

$$X = A + R, \qquad A := P_{\mathcal{A}^{(d)}} X, \qquad R := X - A \in (\mathcal{A}^{(d)})^\perp.$$

Let $M_d := \mathbb{E}[\theta \mid S^{(d)}]$ and $U_d := \theta - M_d$. Assumption 4 gives $M_d \in \mathcal{A}^{(d)}$, while $\mathbb{E}[U_d Y] = 0$ for every $Y \in \mathcal{H}_d$. Since $A - M_d \in \mathcal{A}^{(d)}$, $R \perp \mathcal{A}^{(d)}$, and $U_d \perp \mathcal{H}_d$, the Pythagorean identity yields

$$\mathbb{E}[(X - \theta)^2] = \|A - M_d\|_2^2 + \|R\|_2^2 + \|U_d\|_2^2 = \mathbb{E}[(A - \theta)^2] + \|R\|_2^2.$$

Thus replacing X by A weakly reduces mean squared error, and the reduction is strict whenever $R \neq 0$ in L^2.

Because $1 \in \mathcal{A}^{(d)}$, orthogonality gives $\mathbb{E}[R] = 0$. Moreover, $A - \mathbb{E}[A] \in \mathcal{A}^{(d)}$, so $\mathrm{Cov}(A, R) = 0$. Therefore

$$\mathrm{Var}(X) = \mathrm{Var}(A) + \mathrm{Var}(R) = \mathrm{Var}(A) + \|R\|_2^2 \geq \mathrm{Var}(A).$$

Assumption 5 then implies $C_d(X) \geq C_d(A)$.

Let X_d^* be the unique minimizer from Proposition 1. Its objective value and cost are finite.

If $X_d^* \notin \mathcal{A}^{(d)}$, its projection $A_d^* := P_{\mathcal{A}^{(d)}} X_d^*$ satisfies $C_d(A_d^*) \le C_d(X_d^*) < +\infty$ and

$$\mathbb{E}[(A_d^* - \theta)^2] < \mathbb{E}[(X_d^* - \theta)^2].$$

Hence $J_d(A_d^*) < J_d(X_d^*)$, contradicting optimality. Therefore $X_d^* \in \mathcal{A}^{(d)}$, so $X_d^* = c_d^* + (\omega_d^*)^\top S^{(d)}$ for some $c_d^* \in \mathbb{R}$ and $\omega_d^* \in \mathbb{R}^{d+1}$. The induced estimate is unique, although its coefficient representation need not be unique if $1, S_0, \dots, S_d$ are linearly dependent in L^2.

■

Proof of Corollary 1

Fix $d \in \mathbb{N}_0$ and define the linear map

$$T_d : \mathbb{R}^{d+1} \times \mathbb{R} \to \mathcal{H}_d, \qquad T_d(\omega, c) := c + \omega^\top S^{(d)}.$$

Its range is $\mathcal{A}^{(d)}$. By Proposition 4, the unique optimal induced estimate for problem (P_d) belongs to $\mathcal{A}^{(d)}$. Restricting the original problem to the range of T_d therefore does not change its minimum.

Because T_d is linear, the function

$$(\omega, c) \mapsto \mathbb{E}\Big[\big(T_d(\omega, c) - \theta\big)^2\Big]$$

is convex. Moreover,

$$\widetilde{C}_d(\omega, c) = C_d\big(T_d(\omega, c)\big)$$

is convex because C_d is convex and T_d is linear. Hence the coefficient problem is convex.

The optimal induced estimate has at least one coefficient representation because it belongs to $\mathcal{A}^{(d)}$, so the coefficient problem admits a minimizer. Every minimizing coefficient pair induces an optimal estimate for problem (P_d); uniqueness of the optimal induced estimate then implies that all minimizing coefficient pairs generate $\ell_d^*(S^{(d)})$ almost surely. ■

Proof of Proposition 5

Fix $d \in \mathbb{N}_0$ and write $m := d + 1$. Under Assumption 6,

$$S^{(d)} = \theta \mathbf{1}_m + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_m),$$

where ε is independent of θ. By Proposition 4, it is enough to consider estimates of the form $\ell(S^{(d)}) = \omega^\top S^{(d)} + c$.

For fixed ω, the variance term does not depend on c, and the first-order condition for the squared-error term is

$$\mathbb{E}[\omega^\top S^{(d)} + c - \theta] = 0.$$

Since $\mathbb{E}[S^{(d)}] = \mu_\theta \mathbf{1}_m$, the optimal intercept is

$$c = \big(1 - \omega^\top \mathbf{1}_m\big)\mu_\theta.$$

Let $\widetilde{\theta} := \theta - \mu_\theta$ and $\widetilde{S} := S^{(d)} - \mu_\theta \mathbf{1}_m$. With the optimal intercept,

$$\ell(S^{(d)}) - \mu_\theta = \omega^\top \widetilde{S}, \qquad \ell(S^{(d)}) - \theta = \omega^\top \widetilde{S} - \widetilde{\theta}.$$

Define

$$\Sigma := \mathrm{Var}(\widetilde{S}) = \sigma_\theta^2 \mathbf{1}_m \mathbf{1}_m^\top + \sigma_\varepsilon^2 I_m, \qquad \mathrm{Cov}(\widetilde{S}, \widetilde{\theta}) = \sigma_\theta^2 \mathbf{1}_m.$$

The objective, after optimizing over c, is

$$\sigma_\theta^2 - 2\sigma_\theta^2 \omega^\top \mathbf{1}_m + (1 + \lambda m)\omega^\top \Sigma \omega.$$

Because Σ is positive definite, this is a strictly convex quadratic in ω. Its unique minimizer satisfies

$$(1 + \lambda m)\Sigma\omega = \sigma_\theta^2 \mathbf{1}_m,$$

and therefore

$$\omega = \frac{\sigma_\theta^2}{1 + \lambda m} \Sigma^{-1} \mathbf{1}_m.$$

The Sherman–Morrison formula[6] gives

$$\Sigma^{-1} = \frac{1}{\sigma_\varepsilon^2} I_m - \frac{\sigma_\theta^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2 + \sigma_\theta^2 m)} \mathbf{1}_m \mathbf{1}_m^\top, \qquad \Sigma^{-1}\mathbf{1}_m = \frac{1}{\sigma_\varepsilon^2 + \sigma_\theta^2 m} \mathbf{1}_m.$$

Hence

$$\omega_d^* = \alpha_m \mathbf{1}_m, \qquad \alpha_m = \frac{1}{1 + \lambda m} \frac{\sigma_\theta^2}{\sigma_\varepsilon^2 + \sigma_\theta^2 m},$$

[6] If A is invertible and u, v are conformable vectors with $1 + v^\top A^{-1} u \neq 0$, then

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

Here, take $A = \sigma_\varepsilon^2 I_m$, $u = \sigma_\theta^2 \mathbf{1}_m$, and $v = \mathbf{1}_m$.

and

$$c_d^* = (1 - m\alpha_m)\mu_\theta.$$

Finally,

$$(\omega_d^*)^\top S^{(d)} + c_d^* = \mu_\theta + m\alpha_m\big(\overline{S}^{(d)} - \mu_\theta\big),$$

and

$$m\alpha_m = \frac{1}{1+\lambda m}\frac{\sigma_\theta^2 m}{\sigma_\varepsilon^2 + \sigma_\theta^2 m} = \kappa_m.$$

This proves both representations in the proposition.

■

Proof of Proposition 6

If $\lambda = 0$, equation (10) gives

$$\mathrm{MSE}(d) = \sigma_\theta^2 - \frac{\sigma_\theta^4 m}{\sigma_\varepsilon^2 + \sigma_\theta^2 m} = \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\theta^2 m}, \qquad m = d + 1.$$

This expression is strictly decreasing in m and converges to zero, proving part (i).

Now suppose $\lambda > 0$. Differentiating the continuous extension in the proposition yields

$$\mathcal{M}'(m) = \frac{\sigma_\theta^4\,(2\sigma_\theta^2\lambda^2 m^3 - 3\sigma_\varepsilon^2\lambda m - \sigma_\varepsilon^2)}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)^2\,(1+\lambda m)^3}.$$

The denominator is strictly positive. Define

$$h(m) := 2\sigma_\theta^2\lambda^2 m^3 - 3\sigma_\varepsilon^2\lambda m - \sigma_\varepsilon^2.$$

Then $h(0) = -\sigma_\varepsilon^2 < 0$, $h(m) \to +\infty$, and

$$h'(m) = 6\sigma_\theta^2\lambda^2 m^2 - 3\sigma_\varepsilon^2\lambda.$$

Thus h is strictly decreasing on $(0, m_0)$ and strictly increasing on (m_0, ∞), where

$$m_0 := \sqrt{\frac{\sigma_\varepsilon^2}{2\sigma_\theta^2\lambda}}.$$

Because $h(m_0) < h(0) < 0$, the function h crosses zero exactly once on $(0, \infty)$. Denote this root by $\widehat{m}_{\mathrm{MSE}}$. The sign of $\mathcal{M}'(m)$ is therefore negative for $m < \widehat{m}_{\mathrm{MSE}}$ and positive for $m > \widehat{m}_{\mathrm{MSE}}$, proving the asserted single-trough shape.

If $\widehat{m}_{\mathrm{MSE}} \leq 1$, then $\mathcal{M}$ is strictly increasing on $[1, \infty)$, so $m = 1$ is the unique integer minimizer. If $\widehat{m}_{\mathrm{MSE}} > 1$, strict monotonicity implies that every integer below $\lfloor \widehat{m}_{\mathrm{MSE}} \rfloor$ is dominated by $\lfloor \widehat{m}_{\mathrm{MSE}} \rfloor$, while every integer above $\lceil \widehat{m}_{\mathrm{MSE}} \rceil$ is dominated by $\lceil \widehat{m}_{\mathrm{MSE}} \rceil$. The integer minimizers therefore lie in the one- or two-element set

$$\{\lfloor \widehat{m}_{\mathrm{MSE}} \rfloor, \lceil \widehat{m}_{\mathrm{MSE}} \rceil\} \cap \mathbb{N},$$

so there is either one minimizing integer or two adjacent minimizing integers. Since $m = d+1$, the same conclusion holds for $d \mapsto \mathrm{MSE}(d)$.

For every finite $m > 0$,

$$\mathcal{M}(m) = \sigma_\theta^2 - \frac{\sigma_\theta^4 m(1 + 2\lambda m)}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)\,(1 + \lambda m)^2} < \sigma_\theta^2,$$

because the subtracted term is strictly positive. Moreover, this term is $O(m^{-1})$, so $\mathcal{M}(m) \to \sigma_\theta^2$. Hence $\mathrm{MSE}(d) < \sigma_\theta^2$ for every finite d and $\mathrm{MSE}(d) \to \sigma_\theta^2$. Information overload follows from Proposition 3.

■

Proof of Proposition 7

Write $m := d + 1$. By (12),

$$L(d) = \sigma_\theta^2 - \frac{\sigma_\theta^4 m}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)\,(1 + \lambda m)}.$$

Taking the difference between the expressions at $m+1$ and m, placing the two fractions over a common denominator, and simplifying gives

$$L(d+1) - L(d) = \frac{\sigma_\theta^4 \left[\lambda \sigma_\theta^2 m(m+1) - \sigma_\varepsilon^2\right]}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)\,(1 + \lambda m)\,(\sigma_\varepsilon^2 + \sigma_\theta^2 (m+1))\,(1 + \lambda(m+1))}.$$

Indeed, the numerator generated by the difference of the two fractions is

$$\begin{aligned} &m\left(\sigma_\varepsilon^2 + \sigma_\theta^2(m+1)\right)(1 + \lambda(m+1)) - (m+1)\left(\sigma_\varepsilon^2 + \sigma_\theta^2 m\right)(1 + \lambda m) \\ &\quad = \lambda \sigma_\theta^2 m(m+1) - \sigma_\varepsilon^2. \end{aligned}$$

If $\lambda = 0$, the denominator is positive and the numerator is $-\sigma_\varepsilon^2 < 0$, so $L(d+1) < L(d)$ for every d. This proves part (i).

Suppose $\lambda > 0$. The denominator in the increment formula is strictly positive, so the sign of $L(d+1) - L(d)$ is the sign of

$$m(m+1) - \frac{\sigma_\varepsilon^2}{\lambda \sigma_\theta^2}.$$

The map $x \mapsto x(x+1)$ is strictly increasing on $(0, \infty)$, and its unique positive solution to

$$x(x+1) = \frac{\sigma_\varepsilon^2}{\lambda \sigma_\theta^2}$$

is

$$x_\lambda = \frac{\sqrt{1 + 4\sigma_\varepsilon^2/(\lambda \sigma_\theta^2)} - 1}{2}.$$

Therefore the increments are negative when $m < x_\lambda$, zero when $m = x_\lambda$, and positive when $m > x_\lambda$. This proves that $L(d)$ is single-troughed.

If $x_\lambda \notin \mathbb{N}$, let $n := \lceil x_\lambda \rceil$. Then the increments are negative for every integer $m < n$ and positive for every integer $m \geq n$, so the unique minimizer is $d = n - 1 = \lceil x_\lambda \rceil - 1$. If $x_\lambda = n \in \mathbb{N}$, the increments are negative for $m < n$, zero for $m = n$, and positive for $m > n$. Hence $d = n - 1$ and $d = n$ are the two minimizers. Equivalently, these adjacent dimensions are tied exactly when

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\theta^2 n(n+1)}.$$

In particular, $d = 0$ is the unique minimizer when $x_\lambda < 1$, equivalently when $\lambda > \sigma_\varepsilon^2/(2\sigma_\theta^2)$, and $d = 0$ and $d = 1$ are tied when equality holds.

For completeness, the continuous extension

$$\widetilde{L}(m) := \sigma_\theta^2 - \frac{\sigma_\theta^4 m}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)(1 + \lambda m)}$$

satisfies

$$\widetilde{L}'(m) = -\frac{\sigma_\theta^4 (\sigma_\varepsilon^2 - \lambda \sigma_\theta^2 m^2)}{(\sigma_\varepsilon^2 + \sigma_\theta^2 m)^2 (1 + \lambda m)^2}.$$

Thus its unconstrained minimizer over $m > 0$ is $\sqrt{\sigma_\varepsilon^2/(\lambda \sigma_\theta^2)}$, while its minimizer over the feasible continuous domain $m \geq 1$ is

$$\max\left\{1, \sqrt{\frac{\sigma_\varepsilon^2}{\lambda \sigma_\theta^2}}\right\}.$$

■

Proof of Proposition 8

Fix $d \in \mathbb{N}_0$ and write $m := d+1$. Under Assumption 7, the state and signal vector are jointly Gaussian, so Assumption 4 holds. The quadratic processing cost satisfies Assumptions 3 and 5. Proposition 4 therefore allows us to restrict attention to estimates $\ell(S^{(d)}) = \omega^\top S^{(d)} + c$.

As in the proof of Proposition 5, the optimal intercept is $c = \big(1 - \omega^\top \mathbf{1}_m\big)\mu_\theta$. After centering, the objective becomes

$$\sigma_\theta^2 - 2\sigma_\theta^2 \omega^\top \mathbf{1}_m + (1 + \lambda m)\omega^\top \Sigma \omega,$$

where

$$\Sigma = \sigma_\theta^2 \mathbf{1}_m \mathbf{1}_m^\top + D, \qquad D := \operatorname{diag}\big(\sigma_{\varepsilon,0}^2, \ldots, \sigma_{\varepsilon,d}^2\big)\,.$$

The matrix Σ is positive definite, so the unique minimizer satisfies

$$\omega = \frac{\sigma_\theta^2}{1 + \lambda m} \Sigma^{-1} \mathbf{1}_m.$$

The Sherman–Morrison formula gives

$$\Sigma^{-1} = D^{-1} - \frac{\sigma_\theta^2 D^{-1} \mathbf{1}_m \mathbf{1}_m^\top D^{-1}}{1 + \sigma_\theta^2 \mathbf{1}_m^\top D^{-1} \mathbf{1}_m}.$$

Since $P_d = \mathbf{1}_m^\top D^{-1} \mathbf{1}_m$,

$$\Sigma^{-1} \mathbf{1}_m = D^{-1} \mathbf{1}_m \left(1 - \frac{P_d}{1/\sigma_\theta^2 + P_d}\right) = \frac{1}{1 + \sigma_\theta^2 P_d} D^{-1} \mathbf{1}_m.$$

Consequently,

$$\omega_{d,k}^* = \frac{1}{1 + \lambda m} \frac{\sigma_\theta^2}{1 + \sigma_\theta^2 P_d} \frac{1}{\sigma_{\varepsilon,k}^2}, \qquad k = 0, 1, \ldots, d,$$

and $c_d^* = (1 - (\omega_d^*)^\top \mathbf{1}_m)\mu_\theta$. The factor multiplying $1/\sigma_{\varepsilon,k}^2$ is positive and common to all coordinates. Hence

$$\sigma_{\varepsilon,k}^2 < \sigma_{\varepsilon,j}^2 \quad \Longleftrightarrow \quad \omega_{d,k}^* > \omega_{d,j}^*, \qquad \frac{\omega_{d,k}^*}{\omega_{d,j}^*} = \frac{1/\sigma_{\varepsilon,k}^2}{1/\sigma_{\varepsilon,j}^2} = \frac{\sigma_{\varepsilon,j}^2}{\sigma_{\varepsilon,k}^2}.$$

■

Proof of Proposition 9

Write $m := d+1$ and abbreviate the induced estimates $\ell_d^A(S^{(d)})$ and $\ell_d^N(S^{(d)})$ by ℓ_d^A and ℓ_d^N. Under the true correlated-Gaussian law,

$$\Sigma_m^A := \mathrm{Var}(S^{(d)}) = \sigma_\varepsilon^2(1-\rho)I_m + (\sigma_\theta^2 + \rho\sigma_\varepsilon^2)\mathbf{1}_m\mathbf{1}_m^\top.$$

The same first-order condition used in the proof of Proposition 5 gives

$$\omega_m^A = \frac{\sigma_\theta^2}{1+\lambda m}(\Sigma_m^A)^{-1}\mathbf{1}_m.$$

Since

$$\Sigma_m^A\mathbf{1}_m = m\left(\sigma_\theta^2 + \sigma_\varepsilon^2 q_m\right)\mathbf{1}_m,$$

we obtain

$$\omega_m^A = \frac{\kappa_m^A}{m}\mathbf{1}_m,$$

where κ_m^A is defined in (18). Under the subjective independent-noise law, the same calculation sets $\rho = 0$ and gives

$$\omega_m^N = \frac{\kappa_m^N}{m}\mathbf{1}_m,$$

where κ_m^N is defined in (19).

Now let

$$Z_m := \overline{S}^{(d)} - \mu_\theta, \qquad V_m := \mathrm{Var}(Z_m) = \sigma_\theta^2 + \sigma_\varepsilon^2 q_m.$$

Under the true law, $\mathrm{Cov}(Z_m, \theta) = \sigma_\theta^2$. For $\ell_\kappa := \mu_\theta + \kappa Z_m$, the true total loss is

$$\mathcal{J}_m(\ell_\kappa) = \sigma_\theta^2 - 2\sigma_\theta^2\kappa + (1+\lambda m)V_m\kappa^2,$$

whose minimizer is $\kappa_m^A = \sigma_\theta^2/((1+\lambda m)V_m)$. Completing the square yields

$$\mathcal{J}_m(\ell_\kappa) = \mathcal{J}_m(\ell_d^A) + (1+\lambda m)V_m(\kappa - \kappa_m^A)^2.$$

Setting $\kappa = \kappa_m^N$ and recalling $\Delta_m = \kappa_m^N - \kappa_m^A$ gives the exact loss identity

$$\mathcal{J}_m(\ell_d^N) - \mathcal{J}_m(\ell_d^A) = (1+\lambda m)V_m\Delta_m^2 \geq 0.$$

Moreover,

$$\ell_d^N - \ell_d^A = \Delta_m Z_m, \qquad \omega_m^N - \omega_m^A = \frac{\Delta_m}{m}\mathbf{1}_m,$$

so

$$\|\ell_d^N - \ell_d^A\|_2 = |\Delta_m|\sqrt{V_m}, \qquad \|\omega_m^N - \omega_m^A\|_{\mathrm{E}} = \frac{|\Delta_m|}{\sqrt{m}},$$

where $\|\cdot\|_{\mathrm{E}}$ denotes the Euclidean norm on coefficient vectors.

Suppose first that $\lambda > 0$. Since $q_m \to \rho$, the sequence V_m converges to $\sigma_\theta^2 + \rho\sigma_\varepsilon^2$ and is bounded above and away from zero. The formulas for the two coefficients imply

$$\kappa_m^A = O(m^{-1}), \qquad \kappa_m^N = O(m^{-1}), \qquad \Delta_m = O(m^{-1}).$$

Therefore

$$\|\ell_d^A - \mu_\theta\|_2 = |\kappa_m^A|\sqrt{V_m} \to 0, \qquad \|\ell_d^N - \mu_\theta\|_2 = |\kappa_m^N|\sqrt{V_m} \to 0.$$

The preceding exact identities also give

$$\|\ell_d^N - \ell_d^A\|_2 = O(m^{-1}), \qquad \|\omega_m^N - \omega_m^A\|_{\mathrm{E}} = O(m^{-3/2}),$$

and

$$0 \le \mathcal{J}_m(\ell_d^N) - \mathcal{J}_m(\ell_d^A) = (1 + \lambda m)V_m\Delta_m^2 = O(m^{-1}).$$

This proves part (i).

Now suppose $\lambda = 0$. Since $q_m \to \rho$,

$$\kappa_m^A \to \frac{\sigma_\theta^2}{\sigma_\theta^2 + \rho\sigma_\varepsilon^2}, \qquad \kappa_m^N \to 1, \qquad \Delta_m \to \frac{\rho\sigma_\varepsilon^2}{\sigma_\theta^2 + \rho\sigma_\varepsilon^2}.$$

Because $V_m \to \sigma_\theta^2 + \rho\sigma_\varepsilon^2$,

$$\mathbb{E}\Big[\big(\ell_d^N - \ell_d^A\big)^2\Big] = V_m\Delta_m^2 \to \frac{\rho^2\sigma_\varepsilon^4}{\sigma_\theta^2 + \rho\sigma_\varepsilon^2}.$$

When $\lambda = 0$, the exact loss identity reduces to

$$\mathcal{J}_m(\ell_d^N) - \mathcal{J}_m(\ell_d^A) = V_m\Delta_m^2,$$

so the true total-loss gap has the same strictly positive limit. This proves part (ii).

■

References

Avoyan, A. and A. Schotter (2020). "Attention in Games: An Experimental Study". *European Economic Review* 124, 103410.

Blackwell, D. (1953). "Equivalent comparisons of experiments". *The Annals of Mathematical Statistics* 24.2, pp. 265–272.

Blankespoor, E., E. deHaan, and I. Marinovic (2020). "Disclosure Processing Costs, Investors' Information Choice, and Equity Market Outcomes: A Review". *Journal of Accounting and Economics* 70.2–3, p. 101344.

Branco, F., M. Sun, and J. M. Villas-Boas (2016). "Too Much Information? Information Provision and Search Costs". *Marketing Science* 35.4, pp. 605–618.

Chadd, I., E. Filiz-Ozbay, and E. Y. Ozbay (2021). "The Relevance of Irrelevant Information". *Experimental Economics* 24.3, pp. 985–1018.

Chewning Jr, E. G. and A. M. Harrell (1990). "The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task". *Accounting, organizations and society* 15.6, pp. 527–542.

Dasaratha, K. and K. He (2021). "An experiment on network density and sequential learning". *Games and Economic Behavior* 128, pp. 182–192.

Dummel, S., J. Rummel, and A. Voss (2016). "Additional information is not ignored: New evidence for information integration and inhibition in take-the-best decisions". *Acta psychologica* 163, pp. 167–184.

Enke, B. and F. Zimmermann (2019). "Correlation neglect in belief formation". *The review of economic studies* 86.1, pp. 313–332.

Eppler, M. J. and J. Mengis (2004). "The concept of information overload: A review of literature from organization Science, accounting, marketing, MIS, and related disciplines". *The Information Society* 20.5, pp. 325–344.

Gabaix, X. (2014). "A sparsity-based model of bounded rationality". *The Quarterly Journal of Economics* 129.4, pp. 1661–1710.

Iyengar, S. S. and M. R. Lepper (2000). "When choice is demotivating: Can one desire too much of a good thing?" *Journal of Personality and Social Psychology* 79.6, pp. 995–1006.

Jacoby, J., D. E. Speller, and C. A. Kohn (1974). "Brand choice behavior as a function of information load". *Journal of marketing research* 11.1, pp. 63–69.

Lu, J. (2022). "Limited Attention: Implications for Financial Reporting". *Journal of Accounting Research* 60.5, pp. 1991–2027.

Maćkowiak, B., F. Matějka, and M. Wiederholt (2023). "Rational Inattention: A Review". *Journal of Economic Literature* 61.1, pp. 226–273.

Malhotra, N. K. (1982). "Information load and consumer decision making". *Journal of Consumer Research* 8.4, pp. 419–430.

Matějka, F. and A. McKay (2015). "Rational inattention to discrete choices: A new foundation for the multinomial logit model". *American Economic Review* 105.1, pp. 272–298.

Oprea, R. (2020). "What Makes a Rule Complex?" *American Economic Review* 110.12, pp. 3913–3951.

Payne, J. W. (1976). "Task complexity and contingent processing in decision making: An information search and protocol analysis". *Organizational behavior and human performance* 16.2, pp. 366–387.

Roetzel, P. G. (2019). "Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development". *Business research* 12.2, pp. 479–522.

Saint-Paul, G. (2017). "A "Quantized" Approach to Rational Inattention". *European Economic Review* 100, pp. 50–71.

Sims, C. A. (2003). "Implications of rational inattention". *Journal of monetary Economics* 50.3, pp. 665–690.