Constrained Data-Fitters*

Larry Samuelson
Yale University

Jakub Steiner
University of Zurich, CERGE-EI, and CTS

June 21, 2024

Abstract

We study the maximum-likelihood estimation and updating of an agent subject to frictions that may reflect computational limitations, cognitive constraints, or behavioral biases. We jointly characterize the agent's constrained estimation and updating in a framework reminiscent of a machine learning algorithm. In the absence of frictions, this framework simplifies to standard maximum-likelihood estimation and Bayesian updating. We demonstrate that, under certain intuitive cognitive constraints, simple models yield the most effective constrained fit to the actual data-generating process—more complex models offer a superior fit, but the agent may lack the capability to assess this fit accurately. With some additional structure, the agent's problem is isomorphic to a familiar rational inattention problem.

1 Introduction

We study an agent who fits a statistical model to observed data and uses the model to guide her beliefs about unobserved variables. The agent faces friction in her evaluation of the model's fit to data and in her ability to update her beliefs about their unobserved counterparts. What statistical model should the agent adopt, when she recognizes her own limitations in applying this model?

For example, the agent may be a human resources manager, evaluating job candidates. The observed variable x specifies a candidate's educational attainment, demographic information, references, and other information in the can-

*We thank Sandro Ambuehl, Mira Frick, Heidi Thysen, Ryota Iijima, Rava da Silveira, Ran Spiegler and various seminar and workshop audiences for comments. Pavel Kocourek provided excellent research assistance. Steiner has benefited from grant GACR 24-10145S.

didate's file. The unobserved variable z identifies the characteristics of the candidate (such as intelligence, socioeconomic background, talent, creativity, or reliability) that are potentially relevant to the employer. The manager observes a sample $(x_i)_{i=1}^n$ of job applications independently drawn from the distribution $q_0(x)$, but does not observe the corresponding applicants' latent characteristics $(z_i)_{i=1}^n$. The agent has preconceived partial knowledge of the joint distribution of latent characteristics and observables in the general population of the current hiring season, and forms beliefs about the characteristics of the job candidates in her sample. As in the asymptotic estimation literature, we focus on large samples by letting $n \to \infty$.

The agent's preconceived partial knowledge is represented by a set $\mathcal{P}$ of prospective statistical models. For each model—a probability distribution $p(x, z) \in \mathcal{P}$ of latent and observable variables in the general population—the agent evaluates the likelihood of the observed sample and then chooses the model with the best fit, as in maximum-likelihood estimation.

The existing literature assumes that the agent flawlessly computes the fit of each feasible model. We depart from this literature by assuming that the agent encounters computational and other frictions in calculating the fits.

The departure from the standard framework occurs because our agent, in addition to considering the data points x_i, also reasons about the corresponding latent variables z_i. This reasoning helps her evaluate the likelihood of a candidate model $p(x, z)$. Accordingly, we refer to $(x_i, z_i)_{i=1}^n$ as the extended sample and let the agent form a belief $q(x, z)$ about its empirical frequencies. The agent then computes the fit of the model $p(x, z)$ to the hypothesized extended sample, accounting for the number of extended samples that are consistent with her belief $q(x, z)$.

We consider three combinations of frictions the agent may face. First, consider an agent who faces no frictions. This agent knows the correct $p(x, z)$ and flawlessly updates her belief about the extended sample. For large samples, the law of large numbers implies that the empirical frequencies of the extended sample satisfy $q(x, z) = p(x, z)$. The agent has the correct belief $p(x)$ about the distribution of observables, and her belief $p(z_i \mid x_i)$ about the characteristics of candidate i is given by Bayes' rule applied to the model p.

Second, the agent may not know the true model that describes the general population. For example, our manager may believe that educational attainment is a function of innate intelligence and socioeconomic background, but may be unsure of the functional form of the relationship. Hence, the agent chooses,

from a set $\mathcal{P}$ of considered models, the model $p(x, z)$ that maximizes the likelihood of the observed data. Suppose this agent faces no updating friction. Her reasoning about the extended sample is then straightforward. The sample specifies the marginal frequencies $q_0(x)$ of x, and for each x, the agent again forms Bayesian updates about the latent counterpart of x. Thus, for a given model p, the agent concludes that the empirical frequencies of the extended sample are $q(x, z) = q_0(x)p(z \mid x)$. Using familiar combinatorial asymptotic approximations, we find that the agent's evaluation of the fit of each model $p(x, z)$ approximates the likelihood $\prod_{i=1}^{n} p(x_i)$. The fit, appropriately transformed, then equals $-\operatorname{KL}\big(q_0(x) \parallel p(x)\big)$, and the agent thus chooses the "least wrong" model—the model $p(x, z)$ that minimizes the Kullback-Leibler divergence between the true process $q_0(x)$ and the model's margin $p(x)$, as in the standard results of White (1982) and Berk (1966). In particular, when the agent is well specified, in that the correct model p is included in $\mathcal{P}$ (in our example, if socioeconomic background and innate intelligence are the *only* factors that affect educational attainment, and do so in a functional form considered by the manager), then we replicate Wald (1949): the agent learns the true model.

In this paper, we focus on the third case, in which the agent faces an updating friction that precludes the exact evaluation of models' likelihoods. For example, the human resources manager may fail to account for the confounding effect of socioeconomic background on the relationship between educational attainment and innate intelligence, or may be subject to any of the foibles known to disrupt Bayesian updating. Again, to evaluate the fit of a model $p(x, z)$, the agent forms the belief $q(x, z)$ about the frequencies of the extended sample and computes the fit of the model p as the total p-likelihood of all extended samples with frequencies $q(x, z)$. In this case, however, the agent is constrained to draw $q(x, z)$ from a set $\mathcal{Q}$ that specifies the joint distributions she considers while analyzing data. This set may consist of all distributions from a simple parametric family, or the considered distributions may need to satisfy certain causal relationships, and so on. We assume that for a given model $p(x, z)$, the agent chooses the frequencies $q(x, z) \in \mathcal{Q}$ of the extended sample that generate the best fit of the model p, and call this fit the *constrained likelihood* of the model p. Finally, the agent chooses a model $p(x, z)$ with the highest constrained likelihood. We show that this procedure leads to the adoption of the model $p(x, z)$ and of the belief $q(x, z)$ about the frequencies of the extended sample that jointly minimize $\operatorname{KL}\big(q(x, z) \parallel p(x, z)\big)$, subject to both models being from the considered sets of the distributions.

An agent in our framework thus chooses a pair of models: $p(x, z) \in \mathcal{P}$ and $q(x, z) \in \mathcal{Q}$. The set $\mathcal{P}$ encompasses the set of models that the agent considers as possible data-generating processes. The updating frictions captured by the set $\mathcal{Q}$ reflect the fact that Bayesian updating and hence likelihood evaluation may be constrained by conceptual or computational considerations. By appropriately specifying the set $\mathcal{Q}$, we construct a framework that relaxes Bayes' rationality, captures many ideas from behavioral economics, and is open to the analytical methods of information design.

Importantly, the sets $\mathcal{P}$ and $\mathcal{Q}$ may differ. When contemplating the pool of potential candidates, our human resources manager may view the latent variables as causes of the observable variables. It is then natural to organize her model in terms of a marginal distribution $p(z)$ of the latent variables and conditional distributions $p(x \mid z)$ describing how the latent variables determine the observables. When assessing candidates, the human resources manager observes the empirical distribution $q(x)$ of the applications and forms updates $q(z \mid x)$ about candidates' latent characteristics. An ideal reasoner recognizes these two processes as different views of a single underlying relationship, but an ordinary reasoner, constrained in her choice of $p(z)$, $p(x \mid z)$, and $q(z \mid x)$, may well approach these processes differently.

Section 2 presents the framework. Working backwards, we begin with the updating process. We let the agent hold a candidate model $p(x, z)$ and form a belief $q(x, z)$ about the empirical frequencies of the extended sample. Initially, we examine a reduced-form formulation that generalizes Bayes' rule, drawing from the variational inference literature. We interpret the feasible set $\mathcal{Q}$ as capturing a variety of cognitive constraints from behavioral economics. The analogy-based reasoning of Jehiel (2005, 2022), the correlation neglect of Eyster and Rabin (2005), the causality modeled by the directed acyclic graphs of Spiegler (2016), and the posterior separability reminiscent of rational inattention (Caplin and Dean, 2013) all appear as versions of the feasible set $\mathcal{Q}$.

The constrained likelihood maximization that guides the agent's choice of the model $q(x, z)$ is motivated in the variational inference literature as a computationally tractable approximation to Bayesian updating. We provide a precise foundation for this objective, showing that it emerges naturally from our maximum-likelihood estimation sketched above.

We then let the agent *jointly* choose a *generative* model $p(x, z)$ of the process generating the data and a *recognition* model $q(x, z)$ specifying the agent's updating. This leads to an optimization problem known as a variational au-

toencoder, introduced by Kingma and Welling (2013) in the machine learning literature as a feasible approximation of computationally intractable maximum-likelihood estimation. Our interpretation of the problem is more in the spirit of Luce and Raiffa (1957), who conceptualize prior and updated beliefs as being jointly determined in order to ensure their consistency.

Section 3 presents our first set of applications. We assume that the set $\mathcal{P}$ of data-generating models considered by the agent has a permissive structure—the choice of the marginal distribution of latent variables $p(z)$ is unconstrained and separable from the choice of conditional distributions $p(x \mid z)$. We develop two results. First, the agent will choose models that posit simple deterministic relationships between some of the latent variables, thus avoiding the complexity of stochastic relationships. Second, the agent will exhibit rational expectations, in the sense that her prior belief $p(z)$ equals the expected value of the updated belief $q(z \mid x)$. For a Bayesian agent, this relationship is an identity implied by Bayes' rule. In our case, the relationship can fail, but holds for an optimally selected model, even when the agent is misspecified and constraints prevent her from forming Bayesian updates.

Section 4 elucidates the relationship between our framework and the established asymptotic results on misspecified estimation. When the updating constraint is relaxed, the agent accurately evaluates the fit of models, thus reducing our problem to the canonical results of Berk (1966) and White (1982). Conversely, if the agent's ability to update is restricted, she may favor a simpler model, whose fit she can efficiently evaluate, over a correct yet more complex model. We illustrate this preference for simplicity with an example where the agent exhibits correlation neglect, opting for a simpler model that overlooks correlation but provides a higher fit under constraints than the accurate model.

Section 5 shows that when the constraints imposed on our model-fitting agent are sufficiently relaxed, her problem becomes isomorphic to the rational inattention problem with entropic information costs. Consequently, insights from this literature translate to our setting. By exploiting the local invariance of the rational inattention solution to the prior distribution, we demonstrate that certain aspects of the agent's constrained optimal models remain invariant to local alterations in the underlying true data-generating process. This phenomenon leads to an effect reminiscent of base rate neglect. We again derive a simplicity result, showing that if the set of feasible posteriors is convex, then the agent attaches positive probability to only a limited number of values of the latent variable. Finally, Section 6 places our work in the literature.

2 The Model-Fitting Problem

Section 2.1 presents the description of the agent. Working backward, Section 2.2 endows the agent with a fixed model p of the data-generating process and examines the agent's updating, presenting a framework that encompasses Bayesian updating as a special case. As we explain, this framework is motivated in the Bayesian statistics literature as a tractable approximation of Bayesian updating. Section 2.3 shows that this approach can instead be rationalized as a description of the agent's deductive process when dealing with extensive samples and their unobserved counterparts. Section 2.4 then addresses the complete model-fitting problem, examining an agent who simultaneously chooses a model p and the attendant updates q.

2.1 Generative and Recognition Models

An agent considers a model $p(x, z) \in \Delta(X \times Z)$ for a pair of random variables x and z that take values in finite sets X and Z. She observes the realization x but not the realization z, and reasons about the likely value of z. We refer to x as the *observable* variable, to z as the *latent* variable, and to the joint distribution p as the *generative model*. We dub the marginal distribution $p(x)$ as the *belief process*, indicating that the agent believes this process generates the observable variable.[1]

The observable variable x is drawn from an objective distribution $q_0(x)$, referred to as the *true process*. In keeping with the misspecification literature, $q_0(x)$ is allowed to differ from the belief process $p(x)$. Upon observing a realization of x, the agent updates her belief about the latent variable z to a distribution $q(z \mid x) \in \Delta(Z)$, referred to as her *update*. The *recognition model* is the joint distribution

$$q(x, z) = q_0(x)q(z \mid x) \in \Delta(X \times Z) \tag{1}$$

which specifies both the true process and (possibly non-Bayesian) updates. This bundling of the true process and the updates into the same object facilitates formulation of the results below.

For an ideal reasoner, the generative and recognition models $p(x, z)$ and

[1] We use the same symbol to denote a joint distribution, say $p(x, z) \in \Delta(X \times Z)$, and the associated marginal distribution, say $p(x) \in \Delta(X)$, or conditional distributions, say $\big(p(z \mid x)\big)_x \in \Delta(Z)^X$, and rely on the arguments for the distinction.

x	observable variable
z	latent variable
$p(x, z)$	generative model
$p(x)$	belief process
$q(x, z)$	recognition model
$q_0(x)$	true process
$q(z \mid x)$	update

Table 1: Notation and terminology.

$q(x, z)$ would be identical. Our point of departure is that these models may differ, reflecting distinct reasoning processes applied to the data-generating process and to the sample. Returning to the human resources manager of the introduction, the model $p(x, z)$ captures her view of the population from which the candidates are drawn. Here, the agent may view the latent variable z (innate intelligence, socioeconomic background, and so on) as causing the observable variable x (educational attainments and so on). The agent may thus first reason about the distribution of z in the population and then about the causal relationship between z and x. The agent may restrict her causal reasoning to linear models and to subsets of variables, and may impose other simplifications.

In contrast, the model $q(x, z)$ describes how the human resources manager reasons about her sample of candidates. Here, the distribution of x is given by the sample, and the manager's task is to update her beliefs about the latent variable z. She may simplify the updating task by assuming causal relationships, originating with some of the observed variables x and explaining the latent variables z via a network of causal relationships, that may not be compatible with her reasoning about the data-generating process captured by her model p. Below, we let the agent choose the two models from distinct constraint sets $\mathcal{P}$ and $\mathcal{Q}$ that reflect the distinct reasoning procedures the agent applies to the data-generating process and to the sample, respectively.

2.2 Constrained Updating and Likelihood Evaluation

Following the literature on variational Bayesian methods, we first fix the generative model $p(x, z)$ and focus on the agent's updating and evaluation of the fit of the generative model to the true process. This can be viewed as the first step in examining the agent's estimation problem, or as capturing an agent who is confident that her generative model is the "right" prior.

2.2.1 The Constrained-Updating Problem

Following Jordan et al. (1999), the agent is endowed with a compact set $\mathcal{Q} \subseteq \Delta(X \times Z)$ of recognition models $\tilde{q}(x, z)$ she considers and adopts the recognition model $q(x, z)$ that solves the *constrained-updating problem*[2]

$$\begin{aligned} \max_{\tilde{q}(x,z)} \quad & \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x, z)\right) \\ \text{s.t.} \quad & \tilde{q}(x, z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x), \end{aligned} \tag{2}$$

where H stands for Shannon entropy.[3] Since the marginal distribution $\tilde{q}(x) = q_0(x)$ is fixed, the agent controls only the updates $\tilde{q}(z \mid x)$. We refer to the value of the constrained-updating problem, $\mathrm{E}_{q(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(q(x, z)\right)$, as the *constrained likelihood*.[4]

Before providing our own microfoundation in Section 2.3, we review the standard motivation for this problem. The first, "reconstruction term" of the objective is the expected p-log-likelihood induced by the chosen recognition model $\tilde{q}$. This term specifies how well pairs (x, z) drawn from the recognition model $\tilde{q}$ fit the generative model p. The second, "regularization" term, equal to the entropy of the generative model, is justified in the literature as preventing over-fitting, since it favors generative models that exhibit uncertainty. The objective is commonly motivated as a lower bound on the likelihood, or the "evidence," that can be obtained by selecting from the set of feasible recognition models.[5]

[2] To keep the notation simple, we do not distinguish between a random variable and its realization, except in the case of expectation, where we indicate the random variable over which the expectation is taken with a hat. For example, $\mathrm{E}_{p(x)} f(\hat{x}, z)$ is an expectation with respect to the random variable $\hat{x}$ drawn from the distribution $p(x)$, with the realized value z treated as a parameter.

[3] The entropy of a distribution $q(y)$ is $-\sum_y q(y) \ln q(y)$. We apply the convention $0 \ln 0 = 0$ throughout the paper.

[4] We assume that $\mathcal{Q}$ contains at least one distribution $q(x, z)$ such that $q(x) = q_0(x)$ and whose support is a subset of the support of $p(x, z)$. The existence of an optimizer is then ensured. This distribution achieves finite value and the set of feasible distributions that achieves at least this value is compact. Since the objective is continuous on this set, the solution exists. Note that $\mathrm{supp}(q_0) \subseteq \mathrm{supp}(p(x))$ implies that the agent cannot refute the model p using data drawn from q_0.

[5] When the updating constraint is separable across x, Problem (2) can be separated across values x as a maximization of $\mathrm{E}_{\tilde{q}(z|x)} \ln p(x, z) + \mathrm{H}\left(\tilde{q}(z \mid x)\right)$ over $\tilde{q}(z \mid x)$ from some set $\mathcal{Q}'$. The objective of this problem is called the *evidence lower bound* (ELBO). This term is justified by the fact that this objective can be rewritten as $\ln p(x) - \mathrm{KL}\left(\tilde{q}(z \mid x) \parallel p(z \mid x)\right)$, where KL is the Kullback-Leibler divergence. Since $\ln p(x)$ is called *evidence* in Bayesian statistics and the KL-divergence is non-negative, the ELBO is indeed a lower bound on the

We refer to the first and second constraints as the *updating constraint* and the *empirical constraint*, respectively. In Section 2.3 we interpret the marginal recognition model $\tilde{q}(x)$ as the empirical distribution of the observed data. In the limiting case of a diverging sample size, the distribution of observed data will match the true process $q_0(x)$, giving our empirical constraint. We interpret this constraint as ensuring that the agent cannot fabricate data, while noting that it does not imply that the agent's generative model p matches the data. We illustrate the frictions captured by the updating constraint in Section 2.2.2.

The constrained-updating problem (2) can be rewritten as

$$\min_{\tilde{q}(x,z)} \mathrm{KL}\left(\tilde{q}(x,z) \parallel p(x,z)\right), \tag{3}$$

subject to the same constraints as in (2). Here, KL, representing the Kullback-Leibler divergence, is often interpreted as a pseudo-distance between two distributions.[6] This reformulation of the objective emphasizes that the agent attempts to form the recognition model that is consistent with her generative model.

We interpret the updating problem (2) as a reduced-form representation of the reasoning of an agent who faces a rich sample of observations $(x_i)_{i=1}^n$. In Section 2.3 we explicitly study the agent's analysis of the large samples and derive Problem (2) in the limit of an arbitrarily large sample size.

2.2.2 The Updating Constraint

The updating constraint captures the agent's reasoning about the relationships between the observed data x and their latent counterparts z. We review some illustrative examples.

Unconstrained Updating. Reassuringly, Bayesian updating and unconstrained likelihood evaluation emerge as the solution to the constrained-updating problem when the updating constraint is lifted.

Proposition 1. *If updating is unconstrained, $\mathcal{Q} = \Delta(X \times Z)$, then the agent forms Bayesian updates, $q(z \mid x) = p(z \mid x)$ for all x in the support of $q_0(x)$, and achieves a value of the updating problem equal to $\mathrm{E}_{q_0(x)} \ln p(\hat{x}) + C$, where C is constant across recognition models that satisfy the empirical constraint.*

evidence.

[6] The KL-divergence, also referred to as the relative entropy of distributions $q(y)$ and $p(y)$, is $\sum_y q(y) \ln \frac{q(y)}{p(y)}$.

Proof. Using the chain rule for KL-divergence, rewrite the objective in (3) as

$$\mathrm{KL}\left(\tilde{q}(x,z) \parallel p(x,z)\right) = \mathrm{KL}\left(q_0(x) \parallel p(x)\right) + \mathrm{E}_{q_0(x)} \mathrm{KL}\left(\tilde{q}(z \mid \hat{x}) \parallel p(z \mid \hat{x})\right),$$

and observe that the agent does not control the first term. When $\left(\tilde{q}(z \mid x)\right)_x$ is unconstrained, then the minimizer satisfies $q(z \mid x) = p(z \mid x)$ for all x in the support of $q_0(x)$ because the KL-divergence is minimized with value 0 when its two arguments coincide. The optimal recognition model for Problem (2) thus achieves the value $-\mathrm{KL}\left(q_0(x) \parallel p(x)\right) = \mathrm{E}_{q_0(x)} \ln p(\hat{x}) + C$, where $C = \mathrm{H}\left(q_0(x)\right)$ is constant across recognition models that satisfy the empirical constraint. □

Analogy-Based Constraint. Following Jehiel (2005, 2022), the agent's updates $\left(q(z \mid x)\right)_x$ may be constrained to be measurable with respect to a partition of X, where each element of the partition is a set of observable values that the agent considers analogous at the updating stage.

Causality. An agent's (mis)perception of the correlation structure of the various variables can be represented by a directed acyclic graph (DAG), as in Spiegler (2016). Let the latent and observable variables be multidimensional, $z = (z_1, \ldots, z_k)$ and $x = (x_1, \ldots, x_l)$, and let $y = (x, z)$ be the tuple of all variables, with the nodes in the graph given by $(y_i)_{i=1}^{l+k}$. As the name suggests, the graph has directed edges that induce no cycles. The interpretation is that a variable y_i in the graph is determined only by the variables sitting at the origins of the edges terminating at y_i. The DAG is said to capture the causal structure of the variables (Pearl, 2009). Sloman (2005) explains how DAGs are used in the psychology literature to model boundedly rational reasoning (see also Sloman and Lagnado (2015)).

More precisely, the DAG restricts correlations among the variables y. Given a node i, let $R(i)$ denote the set of its immediate predecessors (which may be empty), and let $y_{R(i)}$ be the set of corresponding variables. Then the recognition model consistent with the DAG must factorize as

$$q(y) = \prod_{i=1}^{l+k} q(y_i \mid y_{R(i)}), \tag{4}$$

implying the updating constraint. For illustration, the DAG $z_1 \leftarrow x \rightarrow z_2$ factorizes as $q(x,z) = q(x)q(z_1 \mid x)q(z_2 \mid x)$, capturing an agent who restricts her recognition model to exhibit conditional independence over the two latent

variables.

2.3 Microfoundations

The Bayesian statistics literature motivates the constrained-updating problem (2) as a computationally tractable approximation to Bayes' rule. We provide a microfoundation based on the agent's analysis of data. Readers interested in applications rather than foundations can skip this subsection.

We endow the agent with a sample of draws of the observable variable, with each draw accompanied by an unobserved draw of the latent variable. The agent estimates the joint frequencies $q(x, z)$ of the observable-latent variable pairs in this extended sample among joint distributions from $\mathcal{Q}$ that are consistent with the observed sample, $q(x) = q_0(x)$.

As the sample becomes large, the agent's estimate is characterized by Sanov's theorem. This theorem implies that asymptotically, the agent's belief over the extended samples generated by drawing from $p(x, z)$ concentrates on those extended samples whose joint frequencies $q(x, z)$ minimize $\mathrm{KL}\left(\tilde{q}(x, z) \parallel p(x, z)\right)$, subject to the two constraints from (2).

Instead of applying Sanov's theorem directly, we find it instructive to derive the constrained-updating problem from first principles.[7] To see how the estimation procedure leads to the objective from (2), we note that two factors enter into the likelihood that an extended sample with frequencies matching $q(x, z)$ was produced by drawing from $p(x, z)$. For any given extended sample with frequencies $q(x, z)$, the log-likelihood of drawing that sample from p corresponds to the first term in (2). In addition, the larger the entropy of $q(x, z)$, the larger the number of distinct samples with frequencies matching $q(x, z)$, and hence the higher the likelihood that draws from $p(x, z)$ yield such frequencies. The number of distinct permutations of a sample increases exponentially with its length, at a rate equal to the entropy of the sample's distribution, leading to the entropy term in the objective from (2).

More precisely, we consider a series of settings indexed by $n \in \mathbb{N}$. In each setting n, the agent observes a sample $x^n = (x_1, \ldots, x_n)$ with the empirical distribution $q_0^n(x)$. For each setting $n \in \mathbb{N}$, the agent is endowed with a set $\mathcal{Q}^n \subseteq \Delta(X \times Z)$ of the joint distributions she considers. Each distribution $\tilde{q}(x, z)$ from this set satisfies the integer constraint (i.e., $\tilde{q}(x, z)n \in \mathbb{N}$), the empirical

[7] Samuelson and Steiner (2023) use similar arguments as a foundation for examining growth processes.

constraint $\tilde{q}(x) = q_0^n(x)$, and a cognitive constraint that further restricts the feasible distributions.

We consider the limit of a large sample, $n \to \infty$, and let the set $\mathcal{Q}^n$ approximate the feasible set from the constrained-updating problem (2). For this, we introduce the parameter $\theta \in [0,1]$, let $\mathcal{Q}(0) = \mathcal{Q} \cap \left\{\tilde{q}(x,z) : \tilde{q}(x) = q_0(x)\right\}$ be the feasible set from Problem (2), and $\mathcal{Q}(\theta) = \mathcal{Q}^{\lfloor \frac{1}{\theta} \rfloor}$ be the feasible set from the setting with $n = \lfloor \frac{1}{\theta} \rfloor$. We assume that the correspondence $\mathcal{Q}(\theta)$ is continuous (i.e., both upper and lower hemicontinuous) at $\theta = 0$.

The agent forms an estimate $q^n(x,z)$ of the joint empirical distribution $\sum_{i=1}^n \mathbb{1}_{(x_i,z_i)=(x,z)}/n$ for the extended sample $(x_i,z_i)_{i=1}^n$ from setting n as follows. The p-likelihood of any single extended sample with empirical distribution $\tilde{q}(x,z)$ is

$$\prod_{i=1}^n p(x_i,z_i) = \prod_{x,z} p(x,z)^{\tilde{q}(x,z)n}.$$

Accounting for the number of such samples, the p-likelihood of the distribution $\tilde{q}(x,z)$ is

$$\ell^n(\tilde{q}) := \mathcal{N}_n(\tilde{q}) \prod_{x,z} p(x,z)^{\tilde{q}(x,z)n}, \tag{5}$$

where $\mathcal{N}_n(\tilde{q})$ stands for the number of distinct extended samples $(x_i,z_i)_{i=1}^n$ that have the empirical distribution $\tilde{q}(x,z)$ and match the observed sample x^n on the margin. The agent's estimate in setting n,

$$q^n(x,z) \in \arg\max_{\tilde{q} \in \mathcal{Q}^n} \ell^n(\tilde{q}), \tag{6}$$

maximizes this p-likelihood.[8]

To simplify the statement of the result, we assume that the constrained-updating problem (2) has a unique solution $q(x,z)$ and let $\ell = \mathrm{E}_{q(x,z)} \ln p(\hat{x},\hat{z}) + \mathrm{H}\left(q(x,z)\right)$ denote the value this solution achieves.

Proposition 2. *As $n \to \infty$, the estimate converges to the solution of the constrained-updating problem and the rescaled log-likelihood to the value of this problem:*

$$\begin{aligned} q^n(x,z) &\to q(x,z), \text{ and} \\ \frac{1}{n} \ln \ell^n(q^n) &\to \ell - \mathrm{H}\left(q_0(x)\right). \end{aligned}$$

[8] That is, the agent forms the maximum a posteriori estimate of the empirical joint distribution constrained to the set $\mathcal{Q}^n$.

The first result indicates that the solution to the constrained-updating problem approximates the estimate from the discrete setting, with the approximation becoming arbitrarily sharp as n gets large. We prove the result in the Appendix by showing that the objective from Problem (2) approximates the rescaled log-likelihood from the discrete setting, and then appealing to the Maximum Theorem. The intuition can be gleaned from the expression for the likelihood in (5). The number $\mathcal{N}_n(\tilde{q})$ of samples with the empirical distribution $\tilde{q}(x,z)$ grows exponentially at the rate $\mathrm{H}(\tilde{q})$ (modulo constant), giving rise to the second term in (2), while the likelihood of each such sample corresponds to the first term in (2).

The second result indicates that the value of Problem (2) approximates the rescaled log-likelihood achieved in the discrete setting. Thus, the solution of Problem (2) is informative not only about the updates of the boundedly rational agent but also about her constrained evaluation of the likelihood. We build on this approximation in the next section, where we let the agent jointly optimize over both her models to maximize this subjective fit.

2.4 Model Fitting

We now let the agent form her generative model with the aim of fitting the observed data. The novelty in the proposed approach is that our agent accounts for her limitations in evaluating the fit of her own generative model. Her subjectively evaluated fit is jointly determined by the generative and recognition models, with the latter defining the criterion by which the former is evaluated, rather than by the generative model alone.

Accordingly, the agent selects a pair of models that jointly solves[9]

$$\begin{aligned} \min_{\tilde{p}(x,z),\tilde{q}(x,z)} \quad & \mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) \\ \text{s.t.} \quad & \tilde{p}(x,z) \in \mathcal{P} \\ & \tilde{q}(x,z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x). \end{aligned} \tag{7}$$

We refer to (7) as the *model-fitting problem*. Compared to the standard mis-

[9] For existence, we again assume that $\mathcal{P}$ and $\mathcal{Q}$ contain at least one pair of p and q such that $q(x) = q_0(x)$ and the support of $q(x,z)$ is a subset of the support of $p(x,z)$. This pair then achieves a finite value and the set of model pairs that achieve at most this value is compact. Continuity of the objective then ensures that a solution exists.

specification framework from Berk (1966), the problem at hand involves not only reasoning about the generating process (captured by control of p) but also reasoning about the latent components of data (captured by control of q).

The joint determination of the generative and recognition models is familiar in the machine learning literature on variational autoencoders (e.g., Kingma and Welling (2013)), capturing the idea that limitations in evaluating the fit of a model play a role in the choice of the model. We explain in Section 6 that this perspective also has familiar antecedents in economics.

The compact sets $\mathcal{P}$ and $\mathcal{Q}$ specify the distributions that the agent considers at two distinct stages of her reasoning. The choice of the generative model $p(x, z)$ pertains to the agent's reasoning about the *process* that generates the data, whereas the choice of the recognition model $q(x, z)$ pertains to the agent's reasoning about the *data* and their latent counterparts. Consequently, the sets $\mathcal{P}$ and $\mathcal{Q}$ will in general be distinct.

When all three constraints of the model-fitting problem overlap, any pair $p = q$ from their intersection constitutes a solution, corresponding to a perfect fit $p(x) = q_0(x)$ accompanied by Bayesian updates $q(z \mid x) = p(z \mid x)$. We are primarily interested in the case of non-overlapping constraint sets, in which case the agent suffers from at least one of two frictions. The misspecification friction is familiar. The set $\mathcal{P}$ will in general exclude the true data-generating process. The best the agent can then hope to do is select the "least wrong" model.

The updating friction, captured by $\mathcal{Q}$, involves difficulties in the evaluation of the model's fit. The agent does not simply calculate the precise p-likelihood of the observed sample. In the machine learning literature, this calculation is well acknowledged to be computationally infeasible when the latent space is too large to allow for numerically practical marginalization $\sum_z p(x, z)$. The variational inference literature similarly considers Bayesian updating to be infeasible. Alternatively, as we have seen, the updating friction may reflect a variety of preconceived notions that restrict the agent's reasoning about data.

3 Optimal Simplicity

Statistical models that best fit generic data-generating processes tend to be complex. However, as we now show, once we account for plausible constraints in likelihood evaluation, the optimal models solving the model-fitting problem tend to be simple. The simplicity notion, made precise in Definition 2, captures the

idea that deterministic relationships are simpler than stochastic relationships. This finding provides a new perspective on the preference for simple models, commonly attributed to William of Ockham but with antecedents, that pervades scientific reasoning.

To state the optimal simplicity result, we impose assumptions on both feasible sets $\mathcal{P}$ and $\mathcal{Q}$. For $\mathcal{P}$, we assume that the agent is fully flexible when reasoning about the latent variables at the generative stage, as captured by Definition 1. For $\mathcal{Q}$, we constrain the agent's reasoning about data by a causal network represented by a DAG.

Definition 1. *The set* $\mathcal{P}$ has unconstrained margin *if* $p(x,z) \in \mathcal{P}$ *if and only if* $\big(p(x \mid z)\big)_z \in \tilde{\mathcal{P}}$ *for some set* $\tilde{\mathcal{P}} \subseteq \Delta(X)^Z$.[10]

A set $\mathcal{P}$ with unconstrained margin captures an agent who has some preconceived knowledge about the statistical implications of each latent value z for the observable variable x but does not know the distribution of the latent values. That is, when $\mathcal{P}$ has unconstrained margin, then (i) all marginal distributions $p(z)$ are feasible and (ii) a tuple of conditional distributions $\big(p(x \mid z)\big)_z$ is feasible for one specification of $p(z)$ if and only if it is also feasible for any other specification of $p(z)$.[11] We impose this assumption on $\mathcal{P}$ throughout the rest of the paper.

We first illustrate the agent's preference for simple deterministic models with an example, and then present a general result.

Example 1 (causal chain)**.** Suppose the set $\mathcal{P}$ has unconstrained margin, the latent variable $z = (z_1, z_2)$ is two dimensional, and the agent restricts her recognition model to comply with the DAG $x \to z_1 \to z_2$, referred to as a chain.

For example, the variable x may be a cv examined by our human resource manager, while z_1 and z_2 may measure the aptitude and grit of the applicant. When thinking about the population of potential job candidates, the manager may recognize that a cv is the joint product of aptitude and grit, and that aptitude and grit are imperfectly correlated. However, when reviewing a cv and drawing inferences about a particular candidate, the manager may simplify her reasoning by first forming an assessment of the candidate's aptitude, and then turning to grit, using only information she gleaned while assessing aptitude, without checking the cv again. Such a succession of one-variable updates may be more tractable than jointly considering aptitude and grit, and the manager

[10] When $p(z) = 0$, then $p(x \mid z)$ can be chosen arbitrarily.

[11] Condition (ii) is analogous to the rectangularity condition of Epstein and Schneider (2003).

may opt for such a simplification either by mistake or to conserve reasoning effort. The following proposition is a special case of Proposition 4 below.

Proposition 3 (deterministic collapse for chain)**.** *The agent from this example believes that z_1 deterministically causes z_2: there exists a deterministic function $d(z_1)$ and a solution to the model-fitting problem such that $z_2 = d(z_1)$ almost surely under both models p and q.*

An agent who frictionlessly maximizes the likelihood of the observed data generically chooses a generative model p that exhibits non-degenerate conditional distributions $p(z_2 \mid z_1)$. In contrast, the stochasticity of $z_2 \mid z_1$ is of no help in improving the constrained fit evaluated by our agent, whose evaluation of likelihood is restricted by her DAG. Even though the agent is able to comprehend a stochastic relation between z_1 and z_2, both when forming her generative model and in the updating stage, the agent chooses to model this relationship in a simple deterministic manner. Given our human resources manager's two-step consideration of aptitude and grit, she finds that recognition models with a deterministic relationship between aptitude and grit improve the fit she can achieve when evaluating her extended sample. Realizing that her reasoning at the recognition stage will take this form, she then restricts her attention to such models at the generative stage. ▲

To extend the chain example, we use the concept of Markov boundary introduced by Pearl (1988). For a DAG over random variables $(x, z_1, \ldots, z_K)$, the *Markov boundary* z^B of a variable x is the smallest subset of the variables $z_1, \ldots, z_K$ that contains all the information about x and hence, once conditioned on the values of the variables from this subset, x is independent of all the other variables. For the chain $x \to z_1 \to z_2$ from Example 1, the Markov boundary of x includes z_1 but not z_2.

Pearl shows that in general, the Markov boundary of x consists of x's immediate predecessors (parents), immediate successors (children), and any immediate predecessor of an immediate successor of x (partners). Let z^{-B} be the complementary tuple of the latent variables that are not in z^B. Our simplicity notion is then:

Definition 2. *The generative model $q'(x, z)$ is* simpler *than $q(x, z)$ if $q'\big(x, z^B\big) = q\big(x, z^B\big)$ and there exists a deterministic function $d\big(z^B\big)$ such that $z^{-B} = d\big(z^B\big)$ almost surely under q'.*

That is, q' is simpler than q if the two distributions coincide when restricted to x and its Markov boundary, and latent variables $z^{-B} \mid z^B$ outside the Markov boundary are deterministic under the simpler distribution.

Suppose $\mathcal{P}$ has unconstrained margin. Consider a DAG over $(x, z_1, \ldots, z_K)$ and let the feasible set $\mathcal{Q}$ consist of all the joint distributions $q(x, z_1, \ldots, z_K)$ consistent with this DAG. Additionally, for each q consistent with the DAG, $\mathcal{Q}$ contains all q' simpler than q. Then the generalization of Example 1 is that variables outside the Markov boundary of x are viewed as deterministic:

Proposition 4. *[deterministic collapse for general DAGs] A solution to the model-fitting problem exists under which the agent believes that the latent variables z^{-B} outside of the Markov boundary of x are a deterministic function $d\big(z^B\big)$ of the latent variables z^B within the Markov boundary of x. That is, $z^{-B} = d\big(z^B\big)$ almost surely under both models p and q.*

Since the Markov boundary of a node in a DAG is typically a small subset of all its nodes, the proposition implies that optimized models treat latent variables as largely deterministic. Remark 1 below explains that we can generally expect solutions in which $z^{-B} = d\big(z^B\big)$ (almost surely) under both models p and q to be uniquely (instead of only weakly, as established in Proposition 4) optimal.

In the next two sections, we develop two intermediate implications of the assumption that $\mathcal{P}$ has unconstrained margin. Section 3.3 uses these to provide intuition for Example 1 and then provides the proof for Proposition 4.

3.1 Rational Expectations

When the set $\mathcal{P}$ has unconstrained margin, the agent who solves the model-fitting problem forms rational expectations. This implication is both of independent interest and useful for the proof of Proposition 4.

We say that the agent has *rational expectations* if

$$p(z) = q(z) \equiv \mathrm{E}_{q_0(x)}\, q(z \mid \hat{x}). \tag{8}$$

The latter identity in (8) is the familiar Bayes' plausibility condition applied to $q(x, z)$. The substantial condition is the first equality. It states that for an agent with rational expectations, there is no inconsistency between the agent's prior $p(z)$ and the updates $q(z \mid x)$ averaged across many draws from the true process $q_0(x)$.

Since our agent is neither Bayes-rational nor has access to the correct model of the true process, she fails to form rational expectations in general settings. Unlike in the extensive rational-expectations literature inspired by Muth (1961) and Lucas (1972), our agent may be systematically fooled. Yet, under the relatively permissive assumption that $\mathcal{P}$ has unconstrained margin, our agent forms rational expectations.

Proposition 5 (rational expectations)**.** *If the set $\mathcal{P}$ of the feasible generative models has unconstrained margin, then the agent has rational expectations.*

Proof. Fix the recognition model $\tilde{q}(x, z)$ and choose the generative model $p(x, z)$ that minimizes their KL-divergence. Using the chain rule, rewrite this objective as

$$\mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) = \mathrm{KL}\left(\tilde{q}(z) \parallel \tilde{p}(z)\right) + \sum_z \tilde{q}(z)\,\mathrm{KL}\left(\tilde{q}(x \mid z) \parallel \tilde{p}(x \mid z)\right).$$

Since the set of the feasible tuples $\left(\tilde{p}(x \mid z)\right)_z$ is independent of $\tilde{p}(z)$, the two terms on the right can be minimized separately. Unconstrained minimization of $\mathrm{KL}\left(\tilde{q}(z) \parallel \tilde{p}(z)\right)$ with respect to $\tilde{p}(z)$ implies $p(z) = \tilde{q}(z)$ for the best response p to the fixed $\tilde{q}$. Since this holds for any recognition model $\tilde{q}$, the optimal pair of generative and recognition models satisfies $p(z) = q(z)$. □

The standard notion of Bayesian plausibility similarly requires that the average updated belief must equal the prior belief. Bayesian plausibility is an identity forced by the mechanics of Bayesian updating, apart from any optimality considerations, and holds for any fixed joint distribution. The rational-expectation condition from the proposition is not an identity and indeed can fail, but holds at the optimum of the model-fitting problem.

A popular intuition in support of rational expectations states that an agent who is systematically surprised should eliminate the surprise by adjusting her prior. While this intuition has no place in the standard Bayesian framework with a fixed prior, it fits well within our framework. When the agent is fully flexible in her choice of $p(z)$, she maximizes the constrained likelihood by matching $p(z)$ to the empirical average of the updates.

The simple proof also illuminates when rational expectations fail. A constrained version of the rational expectations result arises when the agent is restricted in her choice of $p(z)$ to a set $\mathcal{P}_Z \subset \Delta(Z)$ (again, with the feasible set of $\left(p(x \mid z)\right)_z$ independent of $p(z)$). Then, the optimal $p(z)$ is the moment pro-

jection of $q(z)$ on $\mathcal{P}_Z$, i.e., $p(z) \in \arg\min_{\tilde{p}\in\mathcal{P}_Z} \mathrm{KL}\left(q(z) \parallel \tilde{p}(z)\right)$, approximating the empirical average $q(z)$ of her updates as closely as the constraint $\mathcal{P}_Z$ allows. Rational expectations may also fail when $p(z)$ is unrestricted but the feasible set $\tilde{\mathcal{P}}$ of $\left(p(x \mid z)\right)_z$ depends on choice of $p(z)$, leading to a trade-off between the consistency of $p(z)$ with $q(z)$ and the consistency of $p(z \mid x)$ with $q(z \mid x)$.

Spiegler (2020b) provides sufficient conditions for rational expectations in a related but non-nested bounded-rationality setting. His agent reasons about $x = (x_0, \ldots, x_n)$ drawn from $p(x)$. She sets her belief equal to the moment projection $p_R(x) = \prod_{i=0}^{n} p\left(x_i \mid x_{R(i)}\right)$ of the true process on the DAG R. The agent observes a realization of x_0 and forms a posterior belief about variable x_i. As under our definition, the agent is said to have rational expectations if the true average of the agent's subjective posterior beliefs matches her subjective prior. In Spiegler, rational expectations arise when the agent's subjective marginal beliefs match the true marginal distributions. In our framework, rational expectations emerge even when $p(x) \neq q_0(x)$, and the result does not require the DAG structure.

3.2 Posterior Approach

Beyond its direct economic significance, the rational-expectation result from the previous section implies a structure to the model-fitting problem that makes it compatible with the 'posterior approach' familiar from the information-design literature.

To state the analogy to the posterior approach, assume that $\mathcal{P}$ has unconstrained margin, and hence the agent has rational expectations. Then, a triple of $q(z)$, $\left(q(x \mid z)\right)_z$ and $\left(p(x \mid z)\right)_z$ specifies the pair $p(x, z)$ and $q(x, z)$ of the generative and recognition models, because $p(z) = q(z)$ by rational expectations. We refer to the triple as the *posterior representation*, to the conditional distributions $q(x \mid z) \in \Delta(X)$ as *recognition posteriors*, and analogously to $p(x \mid z)$ as *generative posteriors*.[12]

An advantage of the posterior representation is that the objective of the model-fitting problem is separable across latent values.

Lemma 1 (posterior-separable objective)**.** *Suppose $\mathcal{P}$ has unconstrained margin. Then $p(x, z)$ and $q(x, z)$ solve the model-fitting problem if and only if the*

[12] We face a terminological tension here. A natural choice would be to refer to the conditional distributions $q(z \mid x)$ and $p(z \mid x)$ as 'posteriors'. Instead, we attribute this term to $q(x \mid z)$ and $p(x \mid z)$, because this terminological choice facilitates connection to the 'posterior approach' from information design.

posterior representation $q(z)$, $\big(q(x \mid z)\big)_z$ *and* $\big(p(x \mid z)\big)_z$ *solves the equivalent problem:*

$$\begin{aligned} \max_{\tilde{q}(z),(\tilde{q}(x|z))_z,(\tilde{p}(x|z))_z} \quad & \mathrm{E}_{\tilde{q}(z)}\left[\mathrm{E}_{\tilde{q}(x|\hat{z})} \ln \tilde{p}(\hat{x} \mid \hat{z}) + \mathrm{H}\left(\tilde{q}(x \mid \hat{z})\right)\right] \qquad (9) \\ \text{s.t.} \quad & \big(\tilde{p}(x \mid z)\big)_z \in \tilde{\mathcal{P}} \\ & \tilde{q}(z)\tilde{q}(x \mid z) \equiv \tilde{q}(x, z) \in \mathcal{Q} \\ & \mathrm{E}_{\tilde{q}(z)}\, \tilde{q}(x \mid \hat{z}) = q_0(x). \end{aligned}$$

It is the rational expectations implied by the unconstrained margin that do the work in the proof; the unconstrained margin property is needed only to ensure rational expectations. The proof proceeds by showing that the objective in (7) is equivalent to the objective in (9) plus the divergence between the marginal distributions $q(z)$ and $p(z)$, which rational expectations ensure equals zero.

3.3 Proof of Proposition 4

To form some intuition for Proposition 4, we return to Example 1, in which the generative model is restricted to comply with the chain $x \to z_1 \to z_2$. This restriction is equivalent to the requirement $q(x \mid z_1, z_2) = q(x \mid z_1)$; that is, z_2 must be uninformative about x under the recognition model once the agent controls for z_1.[13]

Now consider any candidate solution p and q. Since (i) the objective (9) of the model-fitting problem is posterior separable and (ii) each posterior $q(x \mid z_1, z_2)$ depends only on z_1 but not z_2, we can, for each realization of z_1, modify $q(z_2 \mid z_1) = p(z_2 \mid z_1)$ to be a degenerate distribution that assigns all the probability to

$$d(z_1) \in \arg\max_{\tilde{z}_2} \mathrm{E}_{q(x|z_1)} \ln p\big(\hat{x} \mid z_1, \tilde{z}_2\big),$$

which is the realization of z_2 that maximizes the fit of $p(x \mid z_1, z_2)$ to the given posterior $q(x \mid z_1)$. Since this modification weakly improves the objective at each posterior and is feasible, it must constitute the solution.

The following proof extends this argument to general DAGs.

[13] The factorization constraint for the chain is $q(x, z_1, z_2) = q(z_2 \mid z_1)q(z_1 \mid x)q(x)$. Simple algebra then implies that $q(x \mid z_1, z_2) = q(x \mid z_1)$.

Proof of Proposition 4. Consider a pair of models p and q that solve the model-fitting problem. Since $\mathcal{P}$ has unconstrained margin, $q(z) = p(z)$. If $z^{-B} \mid z^B$ is deterministic under q and p, then the proposition holds. Accordingly, assume that it is not deterministic, in which case q must be consistent with the DAG. Starting from p and q, we construct an alternative pair of feasible models p' and q' such that q' is simpler than q and that, jointly, they achieve at least as high a value in the model-fitting problem as do p and q. Hence p' and q' constitute a solution.

We construct p' and q' from p and q as follows. For each realization of z^B, we replace the conditional distributions $q\big(z^{-B} \mid z^B\big) = p\big(z^{-B} \mid z^B\big)$ with degenerate distributions that assign all probability to z^{-B} equal to the deterministic value

$$d\big(z^B\big) \in \arg\max_{\tilde{z}^{-B}} \mathrm{E}_{q(x|z^B)} \ln p\big(\hat{x} \mid z^B, \tilde{z}^{-B}\big),$$

while keeping $q'\big(z^B\big) = p'\big(z^B\big) = q\big(z^B\big) = p\big(z^B\big)$, $\big(q'(x \mid z)\big)_z = \big(q(x \mid z)\big)_z$, and $\big(p'(x \mid z)\big)_z = \big(p(x \mid z)\big)_z$ unmodified.

By Lemma 1, p and q achieve the value

$$\begin{aligned}
&\mathrm{E}_{q(z)} \left[\mathrm{E}_{q(x|\hat{z})} \ln p(\hat{x} \mid \hat{z}) + \mathrm{H}\left(q(x \mid \hat{z})\right)\right] = \\
&\mathrm{E}_{q(z^B)} \left[\mathrm{E}_{q(x,z^{-B}|\hat{z}^B)} \ln p\big(\hat{x} \mid \hat{z}^B, \hat{z}^{-B}\big) + \mathrm{H}\left(q\big(x \mid \hat{z}^B\big)\right)\right],
\end{aligned}$$

where the equality follows from the independence of the posterior $q(x \mid z)$ from all latent variables outside the Markov boundary of x. This value is at most as high as

$$\begin{aligned}
&\mathrm{E}_{q(z^B)} \left[\max_{\tilde{z}^{-B}} \mathrm{E}_{q(x|\hat{z}^B)} \ln p\big(\hat{x} \mid \hat{z}^B, \tilde{z}^{-B}\big) + \mathrm{H}\left(q\big(x \mid \hat{z}^B\big)\right)\right] = \\
&\mathrm{E}_{q'(z)} \left[\mathrm{E}_{q'(x|\hat{z})} \ln p'(\hat{x} \mid \hat{z}) + \mathrm{H}\left(q'(x \mid \hat{z})\right)\right],
\end{aligned}$$

which is the value achieved by p' and q'.

Additionally, the modified models p' and q' are feasible: (i) the generative model p' is feasible since $\mathcal{P}$ has unconstrained margin. Hence any $p'(z)$ is feasible and $(p'(x \mid z))_z = (p(x \mid z))_z$ is feasible. (ii) $q' \in \mathcal{Q}$ since it is simpler than q and q is consistent with the DAG. (iii) The model q' satisfies the empirical constraint because

$$\mathrm{E}_{q'(z)}\, q'(x \mid \hat{z}) = \mathrm{E}_{q'(z^B)}\, q'\big(x \mid \hat{z}^B\big) = \mathrm{E}_{q(z^B)}\, q\big(x \mid \hat{z}^B\big) = \mathrm{E}_{q(z)}\, q(x \mid \hat{z}) = q_0(x).$$

Thus, p' and q' constitute a solution. □

Remark 1. We can typically expect simple models, in which variables outside the Markov boundary of x are deterministic functions of variables inside the Markov boundary, to be the *only* solutions to the model-fitting problem. When all feasible generative models $p \in \mathcal{P}$ have conditional distributions $p(x \mid z^B, z^{-B})$ that vary with z^{-B}, then $\arg\max_{\tilde{z}^{-B}} \mathrm{E}_{q(x|z^B)} \ln p(\hat{x} \mid z^B, \tilde{z}^{-B})$ is generically unique, and hence $p(z^{-B} \mid z^B) = q(z^{-B} \mid z^B)$ *must* be deterministic at the optimum. In this case the agent has a preconceived view, at the generative stage, of how z^{-B} affects x when controlling for z^B. However, at the recognition stage, she restricts attention to simple recognition models that deem z^{-B} uninformative about x (controlling for z^B). In Example 1, this restriction of the recognition reasoning effectively reduces her optimal generative modeling to a class of likelihood functions $p'(x \mid z_1) = p(x \mid z_1, z_2 = d(z_1))$ that employ only z_1 but not z_2 as the explanatory variable of x. ■

4 Misspecification and Beyond

We first clarify that the agent's constrained reasoning processes about the data-generating process and the data are represented by distinct projections on sets of feasible models, and we then discuss how the two constraints interact.

It is useful to distinguish two related approximations. The first, known as the *moment* projection, approximates $q(y)$ with

$$p(y) \in \arg\min_{\tilde{p} \in \tilde{\mathcal{P}}} \mathrm{KL}\left(q(y) \parallel \tilde{p}(y)\right),$$

where $\tilde{\mathcal{P}}$ is the set of the feasible models. The moment projection characterizes the asymptotic estimate of an agent who selects a model p to fit a large sample with frequencies q.

The second approximation, known as the *information* projection, approximates $p(y)$ by

$$q(y) \in \arg\min_{\tilde{q} \in \mathcal{Q}} \mathrm{KL}\left(\tilde{q}(y) \parallel p(y)\right),$$

where $\mathcal{Q}$ specifies the feasible set. By Sanov's theorem, the information projection arises when the agent is provided with a model $p(y)$ and then forms a belief about an empirical distribution $q(y)$ of a large sample drawn from the model conditional on the event $q(y) \in \mathcal{Q}$.

The model-fitting problem combines both projections—the optimal generative model p is the moment projection of the optimal recognition model q, and vice versa, q is the information projection of p. The two projections are distinct due to the asymmetry of the KL-divergence.

The following example builds on an influential framework for modeling coarse reasoning, and illustrates how distinguishing between the two projections informs an analyst about an appropriate specification of coarse beliefs.

Example 2 (analogy-based constraint)**.** As in Jehiel (2005), the agent's conditional distributions $f(z \mid x)$ must be measurable with respect to a partition $\{X_1, \ldots, X_K\}$ of the set X of the observable values. Let $\mathcal{F}$ be the set of joint distributions f such that $\big(f(z \mid x)\big)_x$ satisfy this measurability restriction. Let $X_k(x)$ be the set of values deemed analogous to x under the agent's partition.

To contrast misspecified learning and constrained updating, we compare two agents. The first agent observes both x and z jointly drawn from $q_0(x, z)$. The agent's asymptotic estimate of this true process converges to the *moment* projection of q_0 onto $\mathcal{F}$. Routine computation reveals that this agent forms conditional distributions $p(z \mid x)$ equal to the *arithmetic* mean $\mathrm{E}_{q_0(x)}\big[q_0(z \mid \hat{x}) \mid \hat{x} \in X_k(x)\big]$ of the Bayesian updates across the values $\tilde{x}$ that the agent deems analogous to x, as assumed in Jehiel (2005).

The second agent is endowed with a generative model $p(x, z)$, observes a large sample of draws of x and forms updates $q(z|x)$ about conditional frequencies in the extended sample as in Section 2.3. For comparison we assume that the marginal distribution of the model $p(x)$ coincides with the true process $q_0(x)$ and focus on the updating friction. In this case, the agent's estimate of frequencies in the extended sample converges to the *information* projection of $p(x, z)$ on $\mathcal{F}$. Another routine computation shows that this agent forms updates given by the *geometric* mean of the Bayesian updates $p(z \mid \tilde{x})$ across $\tilde{x}$ deemed analogous to x (up to renormalization). Thus, relative to the first agent, the coarse belief $q(z \mid x)$ of this second agent is sensitive to variations of small probabilities $p(z \mid \tilde{x})$, $\tilde{x} \in X_{k(x)}$. ▲

We now compare our model-fitting problem to the standard results on misspecified learning. We focus here on the agent's generative model of the observable variable—the belief process $p(x)$. Accordingly, fixing $\mathcal{P}$, denote the feasible set of the belief processes as $\mathcal{P}' = \big\{p'(x) : p'(x) = \tilde{p}(x) \text{ for some } \tilde{p}(x, z) \in \mathcal{P}\big\}$. The next result demonstrates that the model-fitting problem nests White's and Berk's standard results on asymptotic misspecified learning. When the updat-

ing constraint is removed, our optimal belief process $p(x)$ coincides with their standard prediction (coupled with Bayesian update about the latent variable).

Proposition 6. *If updating is unconstrained, $\mathcal{Q} = \Delta(X, Z)$, then the optimal belief process $p(x)$ is the moment projection of q_0 onto $\mathcal{P}'$:*

$$p(x) \in \operatorname*{arg\,min}_{\tilde{p}(x)\in\mathcal{P}'} \mathrm{KL}\left(q_0(x) \parallel \tilde{p}(x)\right). \tag{10}$$

Proof. Using the chain rule, rewrite the objective from (7) as

$$\mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) = \mathrm{KL}\left(q_0(x) \parallel \tilde{p}(x)\right) + \mathrm{E}_{q_0(x)} \mathrm{KL}\left(\tilde{q}(z \mid \hat{x}) \parallel \tilde{p}(z \mid \hat{x})\right).$$

Once the updates are optimized against a given $\tilde{p}$ in the absence of the updating constraint, $\tilde{q}(z \mid x) = \tilde{p}(z \mid x)$, so that the second term on the right vanishes. Thus, $\tilde{p}(x)$ achieves the value given by the objective from (10). □

However, when the updating constraint is nontrivial, it indirectly affects the generative model. The agent no longer selects the model that best fits the true process. Instead, she maximizes the constrained likelihood she is able to compute under her updating constraint. In the next example, the agent chooses to neglect observable correlation. This choice of the simple models arises even though it is feasible for the agent to model the correlation both at the generative and updating stage, because neglecting it facilitates her constrained likelihood evaluation.

Example 3 (correlation neglect)**.** The variables $x = (x_1, x_2)$ and $z = (z_1, z_2)$ are two-dimensional and the true process $q_0(x_1, x_2)$ exhibits correlation. For simplicity, we restrict the generative and recognition models to factorize as

$$p(x_1, x_2, z_1, z_2) = p(z_1, z_2)p(x_1 \mid z_1)p(x_2 \mid z_2) \tag{11}$$

$$q(x_1, x_2, z_1, z_2) = q(z_1)q(z_2)q(x_1, x_2 \mid z_1, z_2). \tag{12}$$

Both these constraints allow for arbitrary correlation between x_1 and x_2. In particular, the belief process $p(x)$ is unconstrained, $\mathcal{P}' = \Delta(X)$, and thus $\mathcal{P}'$ contains the true process $q_0(x)$; the agent is trivially well-specified. Therefore, if the agent were to choose $p(x)$ in the maximum-likelihood estimation without frictions in the evaluation of the likelihood, her asymptotic estimate would equal the true process, $p(x) = q_0(x)$, in line with Wald (1949).

Our agent, however, faces friction in the evaluation of likelihood, as expressed by the updating constraint (12). The next result states that the agent chooses a simple generative model with less than perfect fit, $p(x) \neq q_0(x)$. For the given updating constraint, this simple model, although it differs from the true process, allows the agent to achieve a higher constrained likelihood.

Proposition 7. *When the generative and recognition models are constrained by* (11) *and* (12)*, respectively, then the optimal generative model exhibits independence of* x_1 *and* x_2*.*

Proof. The set $\mathcal{P}$ of the feasible generative models that satisfy (11) has unconstrained margin. Hence Proposition 5 applies and therefore $p(z_1, z_2) = q(z_1, z_2)$. The recognition model is restricted by (12) to independence of z_1 and z_2. Thus, z_1 and z_2 are independent under the generative model: $p(z_1, z_2) = q(z_1, z_2) = q(z_1)q(z_2) = p(z_1)p(z_2)$. Consequently, x_1 and x_2 are independent under the generative model p due to the factorization restriction in (11). □

For illustration, imagine x_1 and x_2 as measuring a job candidate's education and performance on a skill or intelligence test, of the type for which some tech companies are legendary. The variables z_1 and z_2 measure a job candidate's intelligence and grit. When reasoning about the population of job candidates, our human resource manager views performance on the skill test as being determined primarily by innate intelligence and educational attainment as determined primarily by grit, while allowing arbitrary correlation between intelligence and grit. Her generative modeling is thus constrained by (11).

When reasoning about the sample of job candidates, the human resources manager employs a distinct procedure, captured by the constraint (12). Upon observing a collection of pairs (x_1, x_2) with distribution $q_0(x)$, the manager arranges these observations into the various bins, and then assigns to each bin a value (z_1, z_2), such as "high intelligence and ordinary grit", "average intelligence and exemplary grit", and so on. The manager thus arranges her observations into subsets and attributes a cause, in the form of realizations of the latent variables, to each subset. The manager controls how many and which observations of x she attributes to each z bin, and hence controls the distributions $q(z)$ and $q(x|z)$ subject to $\mathrm{E}_{q(z)}\, q(x \mid \hat{z}) = q^0(x)$. The manager mistakenly restricts her analysis to distributions $q(z_1, z_2)$ that exhibit independence. The example shows that this correlation neglect, imposed on the reasoning about the latent variables at the recognition stage, forces correlation neglect on the observables at the generative stage. ▲

5 Connection to Rational Inattention

We now impose additional assumptions on the constraints that allow for the application of techniques from the rational inattention literature.

5.1 Posterior Separable Constraints

We consider here feasibility sets $\mathcal{P}$ and $\mathcal{Q}$ that are *posterior separable*. That is, we assume that there exist compact sets $\bar{\mathcal{P}}, \bar{\mathcal{Q}} \subseteq \Delta(X)$ such that a generative model $p(x, z)$ is feasible if $p(x \mid z) \in \bar{\mathcal{P}}$ for each z in the support of $p(z)$ and a recognition model $q(x, z)$ is feasible if $q(x \mid z) \in \bar{\mathcal{Q}}$, again for each z in the support of $q(z)$. The marginal distributions of the latent variable, $p(z)$ and $q(z)$, are unconstrained.[14]

In context of machine learning and Bayesian statistics, posterior separability of $\mathcal{P}$ naturally applies to an agent who is endowed with a set $\bar{\mathcal{P}}$ of primitive distributions of the observable variable x, can construct mixture distributions from the convex hull of this set, and uses the latent variable z to label the primitive distributions.

Posterior separability of $\mathcal{Q}$ has a natural interpretation in terms of the agent's ability to organize data. As in the microfoundations from Section 2.3, consider an agent who observes a sample $x^n = (x_i)_{i=1}^n$. To compute the p-likelihood of the observed sample, the agent considers extended samples $(x_i, z_i)_i$ with a limited capacity to conceptualize them. Specifically, an agent constrained by posterior-separable $\mathcal{Q}$ only considers extended samples formed by dividing the observed sample x^n into at most $|Z|$ distinct subsamples, as in our interpretation of Example 3, and each subsample, labeled by a value $z \in Z$, must have an empirical distribution that belongs to $\bar{\mathcal{Q}}$.

When $\mathcal{P}$ and $\mathcal{Q}$ are posterior separable, and the latent space is large enough, $|Z| \geq |X|$, then the model-fitting problem is equivalent to a rational inattention problem. To highlight this equivalence, we attach to each primitive distribution an index $a \in A$ so that $\bar{\mathcal{P}} = \{p_a(x)\}_{a \in A}$ for some compact set A. The agent chooses any distribution $p(z) \in \Delta(Z)$ of the latent variable and an assignment $\phi : Z \to A$ that assigns to each latent value z a distribution $p_{\phi(z)}(x)$; this induces the generative model $p(x, z) = p(z)p_{\phi(z)}(x)$. Let us now think of $\ln p_a(x)$ as a

[14] It is immediate that every posterior separable $\mathcal{P}$ has unconstrained margin but not vice versa. An example of $\mathcal{P}$ with unconstrained margin that is not posterior separable is the set of $p(x, z)$, with $z = (z_1, z_2)$, that comply with the DAG $x \to z_1 \to z_2$; this constraint requires the tuple of posteriors $p(x \mid z_1, z_2)$ to be independent of z_2, which is not separable across z.

utility function, reinforcing this by adopting the suggestive notation $u(a,x) = \ln p_a(x)$, and note from Lemma 1 that for the setting of this section, the model-fitting problem simplifies to

$$\max_{\tilde{q}(z),(\tilde{q}(x|z))_z,\tilde{\phi}(z)} \quad \mathrm{E}_{\tilde{q}(z)}\left[\mathrm{E}_{\tilde{q}(x|\hat{z})}\, u\big(\tilde{\phi}(\hat{z}),\hat{x}\big) + \mathrm{H}\left(\tilde{q}(x \mid \hat{z})\right)\right] \tag{13}$$

$$\text{s.t.} \quad \tilde{q}(x \mid z) \in \bar{\mathcal{Q}} \tag{14}$$

$$\mathrm{E}_{\tilde{q}(z)}\, \tilde{q}(x \mid \hat{z}) = q_0(x). \tag{15}$$

This can be formally interpreted as the rational-inattention problem of an agent who learns about x from a signal z (thus reversing our original interpretation of x and z): an agent chooses a distribution $q(z)$ of posteriors $q(x \mid z)$ under the Bayes-plausibility constraint (15), and a choice rule $\phi : z \mapsto a$, to maximize expectation of the payoff $u(a,x)$ augmented with posterior entropy.[15] In addition to the standard rational-inattention problem, constraint (14) restricts the posteriors to the set $\bar{\mathcal{Q}}$.

We apply the concavification technique from Caplin and Dean (2013) to this problem, with one additional step that subsumes constraint (14) by assigning infinite penalty to infeasible posteriors. Accordingly, let $v : \Delta(X) \to \mathbb{R}$ denote the value function

$$v(\rho) = \begin{cases} \max_{a\in A} \mathrm{E}_{\rho(x)} \ln p_a(\hat{x}) + \mathrm{H}\left(\rho\right) & \text{if } \rho \in \bar{\mathcal{Q}}, \\ -\infty & \text{otherwise.} \end{cases}$$

We optimize over distributions $\tilde{\mu}$ of distributions ρ,

$$\max_{\tilde{\mu}\in\Delta(\Delta(X))} \quad \mathrm{E}_{\tilde{\mu}(\rho)}\, v(\hat{\rho}) \tag{16}$$

$$\text{s.t.} \quad \mathrm{E}_{\tilde{\mu}(\rho)}\, \hat{\rho} = q_0,$$

and recall that its solution is given by concavification of v as in Aumann and Maschler (1995) and Kamenica and Gentzkow (2011). Accordingly, let $V(\rho) = \sup\{\xi : (\rho,\xi) \in \mathrm{co}(v)\}$ be the concave closure of v, where $\mathrm{co}(v)$ stands for the convex hull of the graph of v. For a distribution $q_0(x)$, the *tangency points of the concavification* are the tangency points of the function $v(\rho)$ and the hyperplane

[15] Problem (13) subject to (15) is the 'posterior formulation' of the rational-inattention problem by Caplin and Dean (2013). See Matějka and McKay (2015) for an equivalent formulation.

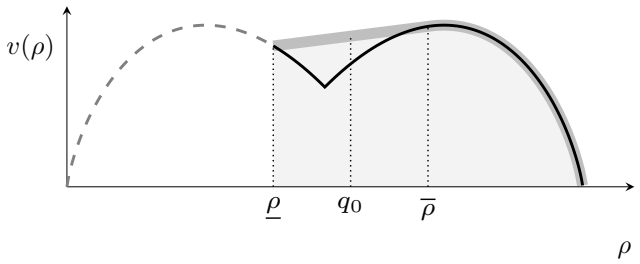

Figure 1: The graph of the value function and its concavification. The value function evaluated at infeasible posteriors is indicated by the dashed curve. The points $\underline{\rho}$ and $\overline{\rho}$ are the tangency points for the given q_0.

that is tangent to $V(\rho)$ at q_0. The *distribution of the tangency points* refers to the weights of the tangency points in their convex combination that equals $q_0(x)$; see Figure 1. The solution to Problem (16) is then determined by the distribution of the tangency points of the concavification of v. By Carathéodory's theorem a solution exists such that its support has size at most $|X| \leq |Z|$.

A solution to Problem (16) corresponds to a class of solutions to the model-fitting problem (13–15) that are equivalent up to permutations of the labels z: Let μ with support size at most $|Z|$ solve (16). Assign arbitrarily to each ρ from the support of μ a distinct label $z = \zeta(\rho)$. For the recognition model, let $q(z) = \mu(\zeta^{-1}(z))$ be the induced distribution of z and let $q(x \mid z) = \zeta^{-1}(z)$ be the distribution $\rho(x)$ that corresponds to each z. For the generative model, let $p(z) = q(z)$. We define $\phi(\rho) \in \arg\max_{a \in A} \mathrm{E}_{\rho(x)} \ln p_a(\hat{x})$ to be the optimal assignment. For each z, let $p(x \mid z) = p_{\phi(q(x|z))}(x)$ be the best fit out of all the primitive distributions from $\bar{\mathcal{P}}$ to $q(x \mid z)$. Consequently, we identify a solution to the model-fitting problem with $\mu(\rho)$ that solves the concavification problem (16).

A special case arises when the updating constraint (14) is removed. Then, the model-fitting problem becomes equivalent to the standard rational-inattention problem with entropic cost. By Proposition 6, this setting corresponds to the setting of White or Berk on asymptotic estimation with the set of the hypotheses $\tilde{p}(x)$ being the convex hull of $\bar{\mathcal{P}}$, coupled with Bayesian updating:

Corollary 1. *The distribution $\mu(\rho)$ solves the model-fitting problem with true process $q_0(x)$, unconstrained updating, and posterior-separable $\mathcal{P}$ with the prim-*

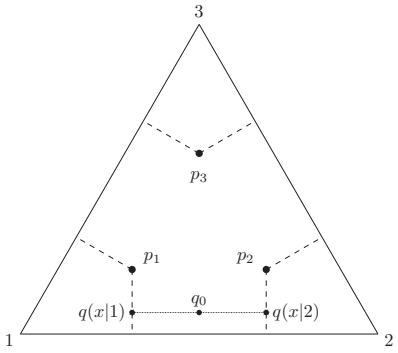

Figure 2: Illustration for Example 4.

itive distributions $\big(p_a(x)\big)_{a\in A}$ *if and only if* $\mu(\rho)$ *is the optimal distribution of posteriors in the rational-inattention problem with the prior* $q_0(x)$ *and the payoff function* $u(a,x) = \ln p_a(x)$.

For a general posterior separable updating constraint, the equivalence of the model-fitting problem (13–15) to the concavification problem (16) has strong implications for the comparative statics with respect to the true process $q_0(x)$. The next result indicates that for certain changes in $q_0(x)$, the agent's choice of posteriors $p(x \mid z)$ and $q(x \mid z)$ is rigid and she adapts only the marginal distribution of the latent variable. When z is the agent's compression of a complex input x—e.g. z is an employer's impression of a job candidate with properties x—the agent's generative and recognition beliefs $p(x \mid z)$ and $q(x \mid z)$ of the actual input x conditional on forming an impression z do not adjust to the base rate $q_0(x)$ of the properties x, akin to the base rate neglect of Tversky and Kahneman (1974).

Proposition 8 (local invariance)**.** *Let* $\mathcal{P}$ *and* $\mathcal{Q}$ *be posterior separable. Consider a true process* $q_0^*(x)$ *in the convex hull of* $\bar{\mathcal{Q}}$ *and denote the associated optimal generative and recognition posteriors by* $p^*(x \mid z)$ *and* $q^*(x \mid z)$*. Then, for all true processes* $q_0(x)$ *in the convex hull of* $\big(q^*(x \mid z)\big)_z$*, a solution to the model-fitting problem exists such that* $p(x \mid z) = p^*(x \mid z)$ *and* $q(x \mid z) = q^*(x \mid z)$.

Proof. The local invariance of $q(x \mid z)$ follows from the analogous local invariance of the tangency points of the concavification of $v(\rho)$. Local invariance of $p(x \mid z) = \phi(q(x \mid z))$ follows from the fact that these are functions of $q(x \mid z)$. □

Example 4 (unconstrained updating)**.** Corollary 1 allows us to translate a rational inattention example from Matyskova and Montes (2023) to the analysis of the model-fitting problem. Let the observable variable x attain values in $X = \{1, 2, 3\}$. The primitive distributions $p_a(x)$ are labeled by $a \in A = \{1, 2, 3\}$ and depicted in Figure 2. The updating is unconstrained.

When the true process is in the convex hull of the primitive distributions, then the agent is well-specified and learns the true process and forms Bayesian updates: $p(x) = q_0(x)$ and $p(x, z) = q(x, z)$. The agent splits $q_0(x)$ into posteriors $p(x \mid z) = q(x \mid z)$ which are equal to the primitive distributions $p_a(x)$.

For true processes that assign nearly all probability to a single value of x (those in neighborhoods of the simplex vertices separated by the dashed lines), the agent declines to employ latent variables in her modeling and chooses her generative model $p(x)$ to be the "nearest" primitive distribution $p_a(x)$ (with $p(z)$ concentrated on an arbitrary value z). This corresponds to a rationally inattentive agent with an extreme prior, who chooses the a priori optimal action without learning.

Finally, for the true process $q_0(x)$ depicted in Figure 2, the generative and recognition models employ two latent values, and the generative model is a mixture of the two nearest primitive distributions. In this case, the impact of a local change in the true process depends on the direction of the change. If $q_0(x)$ stays within the convex hull of the two recognition posteriors, then the optimal posteriors are unaffected. However, small changes of $q_0(x)$ outside of this convex hull induce a change in the recognition posteriors, while the generative posteriors are unaffected. ▲

5.2 Simple Latent Representation

If the set $\bar{\mathcal{Q}}$ of feasible posteriors is convex, then the optimal pair of models p and q exhibit additional simplicity. The number of the employed latent values is bounded by the number of the available primitive distributions and each latent variable has its distinct stochastic meaning. These simplicity properties are analogous to insights from rational inattention, where the optimal signal structure takes the form of a simple action recommendation and the support of the optimal signal is bounded by the number of available actions.

Proposition 9. *If $\bar{\mathcal{Q}}$ is convex, then a solution to the model-fitting model exists such that $q(x \mid z) \neq q(x \mid z')$ and $p(x \mid z) \neq p(x \mid z')$ for each distinct pair z, z'*

from the support of $p(z) = q(z)$. Additionally, the size of the support of $p(z) = q(z)$ is at most equal to the number $|A|$ of available primitive distributions.

The argument in the proof is identical to the argument establishing that each action is taken at a unique posterior in the rational inattention problem.

Proof. Consider a solution $\mu(\rho)$ such that there exist multiple posteriors ρ in its support for which $\phi(\rho) = a$ for some a. Thus, multiple latent values $z = \zeta(\rho)$ with distinct $q(x \mid z) = \rho$ are associated with the same $p(x \mid z) = p_a(x)$. For each such a, replace all these posteriors with a single posterior $\rho' = \mathrm{E}_{\mu(\rho)}[\hat{\rho} \mid \phi(\rho) = a]$ and set $\phi(\rho') = a$. This replacement leads to a solution of (16) because entropy $\mathrm{H}(\rho)$ is a concave function, while the term $\mathrm{E}_{\rho(x)} \ln p_a(\hat{x})$ is linear in ρ for a fixed a. □

When $\bar{\mathcal{Q}}$ is not convex, the simplicity properties from the proposition need not apply. The number of employed latent values may exceed the number $|A|$ of primitive distributions. Moreover, the optimal recognition model may 'hallucinate', i.e., attribute information to differences between latent values that are meaningless under the generative model. Formally, there may exist distinct latent values z and z' such that $p(x \mid z) = p(x \mid z')$ but $q(x \mid z) \neq q(x \mid z')$.

Example 5 (hallucination)**.** The observable variable $x = (x_1, x_2)$ takes values in $\{0,1\}^2$. The agent is endowed with two primitive distributions $p_a(x)$, $a \in \{c, n\}$, where c stands for "correlated" and n for "anticorrelated":

$$\begin{aligned} p_c &= \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right) \\ p_n &= \left(0, \frac{1}{2}, \frac{1}{2}, 0\right), \end{aligned}$$

with the tuples on the right specifying the probabilities of $(x_1, x_2) = 00, 01, 10, 11$. The true process $q_0(x_1, x_2)$ is uniform on $\{0,1\}^2$. The recognition model is constrained to have conditionally independent posteriors: $q(x_1, x_2 \mid z) = q(x_1 \mid z) q(x_2 \mid z)$. Thus, $\bar{\mathcal{Q}}$ is not convex. The set Z is arbitrary, with $|Z| \geq 4$.

The solution to the model-fitting problem consists of $\mu(\rho)$, which splits the true process q_0 into four degenerate posteriors that assign all probability $\rho(x) = 1$ to one of the four states $x = 00, 01, 10, 11$, and an assignment $\phi(00) = \phi(11) = c$, $\phi(01) = \phi(10) = n$.[16] Labeling the set of latent states as

[16] To see this, note that any posterior that does not eliminate all uncertainty over $x \in \{00, 11\}$ vs $x \in \{01, 10\}$ achieves a value $-\infty$. Given that posteriors are restricted to conditional independence, fully informative posteriors are necessary to eliminate this uncertainty.

$Z = \{00, 01, 10, 11\}$, this corresponds to the fully revealing recognition model $q(x \mid z) = \mathbb{1}_{x=z}$ and a partially revealing generative model with $p(x \mid z) = p_c(x)$ for $z = 00, 11$ and $p(x \mid z) = p_n(x)$ for $z = 01, 10$. Thus, the recognition model deems the distinction between $z = 00, 11$ as informative about x, but it is uninformative under the generative model (and similarly for the distinction between $z = 01, 10$). ▲

When the set $\bar{\mathcal{Q}}$ is not convex, the agent would benefit from randomization over the recognition model. If feasible, such randomization transforms $\bar{\mathcal{Q}}$ into its concave closure. Thus, interpreting the posteriors $q(x \mid z)$ as stochastic meanings of the latent values z, the agent may gain by randomizing over these meanings. Randomization over models appears in Spiegler's work on causal reasoning, where mixing arises via equilibrium forces; see Spiegler (2020a) for a review. Our agent may choose to mix in the absence of equilibrium pressures, with the benefit of mixing arising because uncertainty of the recognition model, corresponding to many extended samples, contributes to the good fit. In line with casual observation, Ambuehl and Thysen (2024) experimentally document population heterogeneity in causal reasoning. It remains to be seen whether this variety may usefully be modeled as reflecting mixing.

6 Literature

Our framework rests on two frictions. The first is that an agent finds it difficult to evaluate how well a candidate model fits a set of observations.[17] The second friction arises in forming updated beliefs.

In the machine learning and Bayesian statistics literatures, these frictions represent computational limitations. Calculating the p-likelihood $\prod_{i=1}^{n} p(x_i)$ of a sample $(x_i)_{i=1}^{n}$ under the model $p(x, z)$ is well-acknowledged to be computationally infeasible. In particular, the latent space Z is typically too large to allow for the marginalization $\sum_z p(x, z)$ necessary for the computation of $p(x)$ to be numerically practical. Cinelli et al. note that the likelihood is the obvious standard by which to evaluate a model, but "the computation may not be possible, at least in a viable amount of time" (Cinelli et al., 2021, p. 113). Kingma and

[17] The difficulty does not arise from sampling error. We follow the asymptotic statistics literature (e.g., Van der Vaart (2000)) in assuming the agent has access to an arbitrarily rich set of data. In the parlance of econometrics, we are interested in identification rather than estimation (cf. (Lewbel, 2019, Section 3)).

Welling motivate the model-fitting problem (7) as a procedurally feasible approximation of maximum-likelihood estimation that circumvents the marginalization of "intractable posterior distributions" (Kingma and Welling, 2013, p. 1). Similarly, the point of departure for much of the work in Bayesian statistics is the observation that computing the updated beliefs implied by Bayes' rule is often computationally intractable. Blei et al. (2017) state that "[o]ne of the core problems of modern statistics is to approximate difficult-to-compute probability densities."

The problem of intractable updating was traditionally addressed with Monte Carlo methods (Hastings (1970), Gelfand and Smith (1990)). More recently, research in machine learning has formulated the variational inference problem (2), presented in Section 2.[18] The objective in the variational inference problem (2) is commonly motivated as a computationally feasible lower bound on the "evidence," i.e., the likelihood of the realization of the observed random variable (see Jordan et al. (1999)). The variational inference problem encompasses Bayesian updating and exact evaluation of the model's likelihood as special cases,[19] but also accommodates departures from the standard approach induced by the updating constraint. Strzalecki (2024) represents behavioral updating rules using the variational inference problem and its variants.

Kingma and Welling (2013) introduced the model-fitting problem (7), subsuming the variational inference problem, as a procedurally feasible approximation of maximum-likelihood estimation.[20] This model is known as the variational autoencoder and has become one of the leading approaches for the so-called generative modeling. It's aim is to approximate the true distribution generating the training dataset and to produce new data resembling the original data by drawing from this approximate distribution. Aridor et al. (2020) discuss the neuroscience interpretation of the variational autoencoder. The variational autoencoder is an instance of the information geometry problem of minimizing the divergence between two sets of distributions. In the machine learning literature, the variational autoencoder is addressed by an iterative optimization procedure; Csiszár (1984) provides convergence results.

In economics, the estimation and updating frictions can be interpreted as

[18] See Jordan et al. (1999) for a seminal reference and Blei et al. (2017) and Wainwright et al. (2008) for more recent surveys.

[19] Jordan et al. (1999) note the connection to likelihood evaluation, attributing the observation to Neal and Hinton (1998).

[20] (Cinelli et al., 2021, Chapter 5) provide an introduction to variational autoencoders (7). Doersch (2016) emphasizes their role in providing computationally feasible approaches to otherwise intractable problems.

representing conceptual rather than computational limitations and constraints. For example, it is standard when fitting models to data to restrict attention to models p in which some subset of the observable variables satisfies a linear (or some other parametric) relationship to one another, perturbed by an independent error term.[21] When seeking more flexibility, it is common to simplify the correlation structure exhibited by p, most notably by assuming that some dimensions of x are drawn independently of some others, whether deliberately (a nonparametric estimation exercise must include which variables to include and which to exclude) or inadvertently (see Enke and Zimmermann (2019) for experimental evidence of correlation neglect, and their references for background). These constraints are inevitable—imposing no structure on the set of possible models leaves one with a model that perfectly explains each possible configuration of observable data, precluding useful inference. Even in the hypothetical world of an arbitrarily large sample, the problem remains, as the dimensionality of the observations is in general arbitrarily large. Wolpert and Macready (1997) argue that learning algorithms inevitably pose a trade-off, performing more efficiently in some situations only at the cost of sacrificing performance in some others, and hence invariably involve some friction. Gilboa and Samuelson (2012) identify circumstances under which evaluating models can be fruitless without imposing some constraints on the evaluation.

The manifestation of this evaluation friction is that the set of generative models $\mathcal{P}$ considered by the agent may exclude the true process q_0. A growing literature in economics explores the implications of such misspecified learning. Esponda and Pouzo (2016) examine an equilibrium concept in which a player's belief about her opponent's behavior is learned (via a possibly misspecified model) rather than simply springing to life as part of the equilibrium concept. The result is a generalization of the Nash equilibrium concept that retains the best response criterion for behavior while disciplining beliefs via the (potentially misspecified) learning process. The introduction of behavior endogenizes the data-generating process, which remains exogenous in our framework. Fudenberg et al. (2021) and Heidhues et al. (2021) study the outcomes of individual learning under misspecification, while Bohren (2016) and Bohren and Hauser (2021) examine social learning with misspecified models. For a useful point of entry into the literature, see Frick et al. (2023) and the references

[21] As one recent text (James et al. (2023)[p. 69]) comments, ''[liner regression] has been around for a long time and is the topic of innumerable textbooks ... [it] is still a useful and widely used statistical learning method."

therein.

The second friction involves constraints on the agent's ability to update probabilities. There is ample evidence that updating does not always follow Bayes' rule, instead exhibiting regularities frequently described as heuristics and biases. A large literature has grown out of the early contributions of Tversky et al. (1982) and Tversky and Kahneman (1974). Benjamin (2019) provides a recent and comprehensive survey. Moreover, these departures from Bayes' rule can be economically relevant. For example, Ambuehl and Thysen (2024) report experiments investigating how the type of causal reasoning captured by directed acyclic graphs affects decisions. Andre et al. (2023) show that inflation expectations reflect subjects' narratives that can be represented by directed acyclic graphs. Bhandari et al. (2022) show that inflation and unemployment expectations deviate from rational expectations and find that inserting such irrational expectations can improve the performance of a standard macro model.

In economics, misspecification and updating frictions have traditionally been examined separately, giving rise to what Bohren and Hauser (2023) refer to as the misspecified model approach and the non-Bayesian approach. Bohren and Hauser (2023) investigate the conditions under which these two approaches are equivalent. Their analysis centers around comparing two distributions, one specifying an agent's forecast of her own posteriors and one describing the true distribution of posteriors. Our agent, in general, cannot be represented by a misspecified learning model. Interpreting the generative model as the agent's forecast of updates, and the recognition model as her ex post updates, our agent fails Bohren's and Hansen's "No 'unexpected' beliefs" condition, since her recognition updates may be inconsistent with her generative model. Aina et al. (2023) report experiments in which updated beliefs elicited before signals are observed differ from updates formed after observing these signals. These can be interpreted as empirical counterparts of the generative and recognition models.

The idea that an agent's generative model is tailored to accommodate frictions in her own subsequent updating is novel in economics. Exercises in estimation typically do not address subsequent updating. Conversely, the typical model of updating in economics endows the agent with an exogenously given and unchanging prior belief (or, as in Esponda and Pouzo (2016), a prior over prior beliefs), to which the agent applies Bayes' rule in response to new information (e.g., Baley and Veldkamp (2023)). This dogmatism might be motivated by a result that the prior fades into insignificance with enough data, though a

key point of the misspecification literature is that one cannot always be assured that plentiful data will rescue one from a misspecified model.

In contrast, work in the formative period of Bayesian decision theory readily recognized the link between belief formation and updating. Luce and Raiffa (1957) argue that an agent's anticipated updating plays a role in shaping her prior belief, which arises out of a process of "jockeying—making snap judgments, checking on their consistency, modifying them, again checking on consistency, etc." (Luce and Raiffa, 1957, p. 302). Interestingly, practical variational autoencoder algorithms also involve alternating optimizations over the generative and recognition models. Savage and de Finetti appear to have similar motivations for thinking about probability (though proceeding in a different direction). (Savage, 1972, p. 57) writes that "... the role of the mathematical theory of probability is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected such an inconsistency, he will remove it." (de Finetti, 1937, p. 60; translation in Kyburg and Smokler (1964)) writes, "The practical object of these rules [of probability] is to reduce an evaluation, scarcely accessible directly, to others by which the determination is rendered easier and more precise."

Our work connects to the decision theory literature at a number of points. Our approach is most similar in spirit to that of Spiegler (2016, 2020a,b). We share with Spiegler a focus on procedurally motivated concepts, an interest in frictions, and an attempt to draw economic implications.

There is a growing body of work on non-Bayesian updating, with Epstein (2006) presenting its early axiomatization. Ortoleva (2012) models an agent who updates according to Bayes' rule as long as signals are not too unlikely, but chooses a likelihood-maximizing new prior when confronted with a sufficiently unlikely signal. Schwartzstein and Sunderam (2021) and Aina (2021) study behavioral agents who choose statistical models in the maximum-likelihood estimation. Jakobsen (2021) presents a model of updating in which the agent partitions the simplex, assigns a representative belief to each cell in the partition, and then approximates the true Bayesian posterior with the representative belief of the cell containing the true posterior. This model's collection of representative posteriors is analogous to our set $\mathcal{Q}$. Jakobsen imposes the requirement that the (non-Bayesian) updating in his model must be measurable with respect to Bayesian updating, thereby precluding the type of hallucination that we find in Example 5.

Dominiak et al. (2021) present a model of "conservative updating" that

shares many features with our model.[22] Their agent learns that the probability distribution over states is contained in some set L, and then chooses as her posterior the distribution in L closest to her prior distribution. If "distance" is defined as the Kullback-Leibler divergence and the set L corresponds to having learned that the state is contained in some event E, then this procedure coincides with Bayesian updating. Dominiak et al. (2021) differ from our work in examining distance measures other than the Kullback-Leibler divergence, while we differ in bringing the choice of generative model into the analysis, giving the model-fitting problem (7).

Several papers have brought ideas from machine learning into the economics literature. Zhao et al. (2020) propose and axiomatize a generalization of expected utility theory motivated by machine learning and implemented using a neural net. Aridor et al. (2024) apply the variational autoencoder in a game-theoretic setting as a model of human players' decision process. Caplin et al. (2023) examine data generated by a machine learning algorithm, asking whether the data is consistent with a rationally inattentive agent making decisions subject to costly or constrained information.

7 Discussion

We highlight to several results that have emerged from the analysis. First, it is reassuring that classic models of reasoning emerge as special cases within our framework. An agent free from updating constraints behaves just as in the misspecified learning literature, whereas an agent who is also correctly specified behaves as a perfect Bayesian.

Second, even when beleaguered by frictions, the agent exhibits some familiar properties. For example, under appropriate conditions the agent's optimally chosen beliefs exhibit rational expectations, despite the misspecification and updating frictions.

Third, we have found that ideas from economics and machine learning can be usefully combined. The constrained updating problem, motivated as a computational device in Bayesian statistics, can be interpreted as estimating sample frequencies. The updating constraints, motivated by tractability in the vari-

[22] See Dominiak et al. (2023), Kovach (2021), and Zhao (2022) for related models. In Dominiak et al. (2023) the information learned by the agent corresponds to an event rather than the "general information" allowed by Dominiak et al. (2021). The updating rule in Kovach (2021) is generalized in Dominiak et al. (2021). Zhao (2022) focuses on extending the spirit of Bayesian updating to accommodate more general information.

ational inference literature, can be interpreted in terms of the heuristics and biases common in behavioral economics. When the agent's constraints are posterior separable, the model-fitting problem is isomorphic to a rational inattention problem. This allows us to bring techniques such as concavification to bear on the model-fitting problem. We expect these types of synergies to be useful in a wide range of applications.

Finally, the analysis has produced intriguing implications. We have identified conditions under which an agent who organizes her view of the world according to its causal structure will adopt a simple view of the world, taking the relationship between latent variables to be deterministic for all but those most directly related to the observable variables. If the constraints are posterior separable, the agent will view latent variables as having relatively small support. Moreover, behavioral properties such as correlation neglect and a version of base rate neglect can emerge as implications of the analysis, without being directly assumed.

We have confined ourselves to examining agents' beliefs. An obvious next step would be to consider how agents made decisions based on these beliefs. The resulting actions can then render the data-generating process endogenous, as in Esponda and Pouzo (2016) and Spiegler (2020a), with actions and beliefs satisfying equilibrium consistency conditions. Still within the realm of beliefs, it would be interesting to investigate the implications of distance measures other than the Kullback-Leibler divergence.

A Proofs

Proof of Proposition 2. We prove that

$$\frac{1}{n}\ln \ell^n\left(\tilde{q}\right) \to \mathrm{E}_{\tilde{q}(x,z)} \ln p(x,z) + \mathrm{H}\left(\tilde{q}(x,z)\right) - \mathrm{H}\left(\tilde{q}(x)\right).$$

The limit is the objective in the constrained-updating problem (2) except for the term $-\mathrm{H}\left(\tilde{q}(x)\right)$, which is beyond the agent's control due to the empirical constraint $\tilde{q}(x) = q_0(x)$. Noting that $\frac{1}{n}\ln \ell^n\left(\tilde{q}\right)$ is a monotone transformation of the objective $\ell^n\left(\tilde{q}\right)$ from the estimation problem (6), the proposition then follows from the Maximum theorem.

To prove the limit, observe that

$$\mathcal{N}_n(\tilde{q}) = \prod_{x \in \mathrm{supp}(\tilde{q}(x))} \mathcal{N}'_{\tilde{q}(x)n}\big(\tilde{q}(z \mid x)\big),$$

where $\mathcal{N}'_m\big(\pi(z)\big)$ is the number of the sequences $(z_1, \ldots, z_m)$ of the length m with the empirical distribution $\pi(z)$. This holds because one can compute the number $\mathcal{N}_n(\tilde{q})$ of the sequences $(x_i, z_i)_{i=1}^n$ with distribution $\tilde{q}(x, z)$ that coincide with $(x_i)_{i=1}^n$ on the margin by considering, for each value x, a subsequence $(i_k)_k$ of length $\tilde{q}(x)n$ such that $x_{i_k} = x$ for all k and the empirical distribution of z_{i_k} is $\tilde{q}(z \mid x)$. Then, $\mathcal{N}'_{\tilde{q}(x)n}\big(\tilde{q}(z \mid x)\big)$ is the number of distinct permutations for each such subsequence.

Theorem 11.1.3 in Cover and Thomas (1999) provides the bounds

$$\frac{1}{(m+1)^{|Z|}} \exp\left[m \times \mathrm{H}\left(\pi(z)\right)\right] \leq \mathcal{N}'_m\big(\pi(z)\big) \leq \exp\left[m \times \mathrm{H}\left(\pi(z)\right)\right]. \qquad (17)$$

Substituting these bounds for each x with $m = \tilde{q}(x)n$ gives the bounds

$$\begin{aligned}
&\mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \sum_x \tilde{q}(x)\, \mathrm{H}\left(\tilde{q}(z \mid x)\right) - |Z| \sum_x \frac{\ln\left(\tilde{q}(x)n + 1\right)}{n} \\
&\leq \frac{1}{n} \ln \ell^n\left(\tilde{q}\right) \leq \\
&\mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \sum_x \tilde{q}(x)\, \mathrm{H}\left(\tilde{q}(z \mid x)\right).
\end{aligned}$$

Since $\frac{\ln(\tilde{q}(x)n+1)}{n}$ is nonnegative and at most $\frac{\ln(n+1)}{n}$, both the lower and the upper bounds converge to

$$\mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \sum_x \tilde{q}(x)\, \mathrm{H}\left(\tilde{q}(z \mid x)\right) = \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x, z)\right) - \mathrm{H}\left(\tilde{q}(x)\right),$$

where we applied the chain rule for entropy in the last step. □

Proof of Lemma 1. The negative of the objective of the model-fitting problem

satisfies

$$\begin{aligned} -\operatorname{KL}\left(\tilde{p}(x,z) \parallel \tilde{q}(x,z)\right) &= -\operatorname{KL}\left(\tilde{q}(z) \parallel \tilde{p}(z)\right) - \sum_z \tilde{q}(z) \operatorname{KL}\left(\tilde{q}(x \mid z) \parallel \tilde{p}(x \mid z)\right) \\ &= -\operatorname{KL}\left(\tilde{q}(z) \parallel \tilde{p}(z)\right) + \sum_z \tilde{q}(z) \operatorname{E}_{\tilde{q}(x|z)} \left[\ln \tilde{p}(\hat{x} \mid z) - \ln \tilde{q}(\hat{x} \mid z)\right] \\ &= -\operatorname{KL}\left(\tilde{q}(z) \parallel \tilde{p}(z)\right) + \sum_z \tilde{q}(z) \left(\operatorname{E}_{\tilde{q}(x|z)} \ln \tilde{p}(\hat{x} \mid z) + \operatorname{H}\left(\tilde{q}(x \mid z)\right)\right). \end{aligned}$$

The result follows because the first term on the right vanishes once $\tilde{p}(z)$ is optimized to $p(z) = \tilde{q}(z)$. □

References

Aina, C. (2021). Tailored stories. Technical report, Mimeo.

Aina, C., A. Amelio, and K. Brütt (2023). Contingent belief updating. ECONtribute discussion paper no. 263, University of Bonn and University of Cologne.

Ambuehl, S. and H. C. Thysen (2024). Choosing between causal interpretations: An experimental study. Working paper, University of Zurich and Norwegian School of Economics.

Andre, P., I. Haaland, C. Roth, and J. Wohlfart (2023). Narratives about the macroeconomy. Cesifo working paper number 10535, CESifo.

Aridor, G., R. A. da Silveira, and M. Woodford (2024). Information-constrained coordination of economic behavior. Working paper 32113, NBER. Forthcoming, *Journal of Economic Dynamics and Control*.

Aridor, G., F. Grechi, and M. Woodford (2020). Adaptive efficient coding: A variational auto-encoder approach. biorxiv prepring 2020–05, Cold Spring Harbor Laboratory.

Aumann, R. J. and M. B. Maschler (1995). *Repeated Games with Incomplete Information*. Cambridge, MA: MIT Press.

Baley, I. and L. Veldkamp (2023). Bayesian learning. In *Handbook of Economic Expectations*, pp. 717–748. Elsevier.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, , and D. Laibson (Eds.), *Handbook of Behavioral Economics: Applications and Foundations 1*, Volume 2, pp. 69–186. Elsevier.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics 37*(1), 51–58.

Bhandari, A., J. Borovička, and P. Ho (2022). Survey data and subjective beliefs in business cycle models. Working paper 2763942, SSRN. Forthcoming, *Review of Economic Studies*.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Bohren, J. A. (2016). Informational herding with model misspecification. *Journal of Economic Theory 163*, 222–247.

Bohren, J. A. and D. N. Hauser (2021). Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica 89*(6), 3025–3077.

Bohren, J. A. and D. N. Hauser (2023). Behavioral foundations of model misspecification. Technical report, University of Pennsylvania and Aalto University.

Caplin, A. and M. Dean (2013). Behavioral implications of rational inattention with shannon entropy. Technical report, National Bureau of Economic Research.

Caplin, A., D. Martin, and P. Marx (2023). Modeling machine learning: A cognitive economic approach. Technical report, NYU.

Cinelli, L. P., M. Araújo Marins, E. A. Barros da Silva, and S. Lima Netto (2021). Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pp. 111–149. Springer.

Cover, T. M. and J. A. Thomas (1999). *Elements of Information Theory*. John Wiley & Sons.

Csiszár, I. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions, Dedewicz 1*, 205–237.

de Finetti, B. (1937). La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincare 7*(1), 1–68.

Doersch, C. (2016). Tutorial on variational autoencoders. arxiv preprint arxiv:1606.05908, arXiv.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2021). Minimum distance belief updating with general information. Working paper, Virginia Tech University.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2023). Inertial updating. arxiv preprint arxiv:2303.06336, arXiv.

Enke, B. and F. Zimmermann (2019). Correlation neglect in belief formation. *The review of economic studies 86*(1), 313–332.

Epstein, L. (2006). An axiomatic model of non-bayesian updating. *Review of Economic Studies 73*(2), 413–436.

Epstein, L. G. and M. Schneider (2003). Recursive multiple-priors. *Journal of Economic Theory 113*(1), 1–31.

Esponda, I. and D. Pouzo (2016). Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica 84*(3), 1093–1130.

Eyster, E. and M. Rabin (2005). Cursed equilibrium. *Econometrica 73*(5), 1623–1672.

Frick, M., R. Iijima, and Y. Ishii (2023). Belief convergence under misspecified learning: A martingale approach. *The Review of Economic Studies 90*(2), 781–814.

Fudenberg, D., G. Lanzani, and P. Strack (2021). Limit points of endogenous misspecified learning. *Econometrica 89*(3), 1065–1098.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association 85*(410), 398–409.

Gilboa, I. and L. Samuelson (2012). Subjectivity in inductive inference. *Theoretical Economics 7*(2), 183–216.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Heidhues, P., B. Kőszegi, and P. Strack (2021). Convergence in models of misspecified learning. *Theoretical Economics 16*(1), 73–99.

Jakobsen, A. M. (2021). Coarse bayesian updating. Working paper, University of Calgary.

James, G., D. Witten, T. Hastie, R. Tibshirani, and J. Taylor (2023). Linear regression. In *An introduction to statistical learning: With applications in python*. Springer.

Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic theory 123*(2), 81–104.

Jehiel, P. (2022). Analogy-based expectation equilibrium and related concepts: Theory, applications, and beyond. To appear in the "advances volume" of the twelth World Congress of the Econometric Society, Milan 2020, Paris School of Economics.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine learning 37*, 183–233.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational Bayes. arxiv preprint arxiv:1312.6114, Cornell University.

Kovach, M. (2021). Conservative updating. arxiv preprint arxiv:2102.00152, arXiv.

Kyburg, H. and E. Smokler (1964). *Studies in subjective probability*. Wiley New York.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature 57*(4), 835–903.

Lucas, R. E. (1972). Expectations and the neutrality of money. *Journal of economic theory 4*(2), 103–124.

Luce, D. and H. Raiffa (1957). *Games and Decisions*. New York: John Wiley and Sons.

Matějka, F. and A. McKay (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review 105*(1), 272–298.

Matyskova, L. and A. Montes (2023). Bayesian persuasion with costly information acquisition. *Journal of Economic Theory 211*, 105678.

Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica 29*(3), 315–335.

Neal, R. M. and G. E. Hinton (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. Springer.

Ortoleva, P. (2012). Modeling the change of paradigm: Non-Bayesian reactions to unexpected news. *American Economic Review 102*(6), 2410–2436.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality*. Cambridge university press.

Samuelson, L. and J. Steiner (2023). Growth and likelihood. Technical report, Yale University and University of Zurich/CERGE-EI/CTS.

Savage, L. J. (1972). *The Foundations of Statistics*. New York: Dover Publications. Originally 1954.

Schwartzstein, J. and A. Sunderam (2021). Using models to persuade. *American Economic Review 111*(1), 276–323.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Sloman, S. A. and D. Lagnado (2015). Causality in thought. *Annual review of psychology 66*, 223–247.

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290.

Spiegler, R. (2020a). Behavioral implications of causal misperceptions. *Annual Review of Economics 12*, 81–106.

Spiegler, R. (2020b). Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association 18*(2), 583–617.

Strzalecki, T. (2024). Variational bayes and non-bayesian updating. *arXiv preprint arXiv:2405.08796*.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science 185*(4157), 1124–1131.

Tversky, A., D. Kahneman, and P. Slovic (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wainwright, M. J., M. I. Jordan, et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*(1–2), 1–305.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics 20*(4), 595–601.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Wolpert, D. H. and W. G. Macready (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation 1*(1), 67–82.

Zhao, C. (2022). Pseudo-Bayesian updating. *Theoretical Economics 17*(1), 253–289.

Zhao, C., S. Ke, Z. Want, and S.-L. Hsieh (2020). Behavioral neural networks. Working paper 3633548, SSRN.