CIS-LDRD Project 218313, Final Technical Report. Parsimonious Inference, Information-Theoretic Foundations for a Complete Theory of Machine Learning

Jed A. Duersch (Ch. 1–3) JADUERS@SANDIA.GOV
Thomas A. Catanach (Ch. 1–2) TACATAN@SANDIA.GOV
Sandia National Laboratories
Livermore, CA 94550, United States

Ming Gu (Ch. 3) MGU@BERKELEY.EDU
Department of Mathematics
University of California
Berkeley, CA 94720, United States

SAND Report Number: SAND2020-9834

Keywords: information, rational belief, Bayesian inference, prior belief, complexity, description length, minimum information principle, QR factorization, rank-revealing, low-rank approximation

This work was funded by the U.S. Department of Energy. Original work was funded, in part, by NSF award 1319312 and by the Department of Homeland Security under DOE Contract Number DE-NA0003525 for the management and operation of Sandia National Laboratories. Revised work was supported by NSF award 1760316 and by the Laboratory Directed Research and Development program at Sandia National Laboratories.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Forward

This work examines how we may cast machine learning within a complete Bayesian framework to quantify and suppress explanatory complexity from first principles. Our investigation into both the philosophy and mathematics of rational belief leads us to emphasize the critical role of Bayesian inference in learning well-justified predictions within a rigorous and complete extended logic. The Bayesian framework allows us to coherently account for evidence in the learned plausibility of potential explanations. As an extended logic, the Bayesian paradigm regards probability as a notion of degrees of truth. In order to satisfy critical properties of probability as a coherent measure, as well as maintain consistency with binary propositional logic, we arrive at Bayes' Theorem as the only justifiable mechanism to update our beliefs to account for empiracle evidence.

Yet, in the machine learning paradigm, where explanations are unconstrained algorithmic abstractions, we arrive at a critical challenge: Bayesian inference requires prior belief. Conventional approaches fail to yield a consistent framework in which we could compare prior plausibility among the infinities of potential choices in learning architectures. The difficulty of articulating well-justified prior belief over abstract models is the provinence of memorization in traditional machine learning training practices. This becomes exceptionally problematic in the context of limited datasets, when we wish to learn justifiable predictions from only a small amount of data.

The intuitive solution is to suppress unnecessary complexity, the principle of Occam's Razor. Yet, to suppress complexity, we must have a comprehensive understanding of what complexity is. Our investigation begins in Chapter 1 with analysis of the properties and relationships among various forms of information within a unifed theory of information as a rational measure of change in belief. This theory allows us to understand complexity as information within a wide variety of contexts and analyze the corresponding properties within a rigorous mathematical framework.

Given our theory of information, we are able to return to the question of suppressing complexity in Chapter 2. In this chapter, we examine various notions of algorithmic complexity and universal priors. By applying our theory of information to symbolic descriptions, which articulate learning frameworks, we are able to assign prior plausibility to any architecture that may be cast as Bayesian inference or variational inference. Our numerical experiments demonstrate how our theory suppresses complexity for first principles and arrives at well-justified predictions without cross-validation, thus providing a fundamental capability when datasets are limited. Further, by accounting for multiple potential explanations within our theory, our prototype algorithms also demonstrate natural extrapolation uncertainty from first principles.

Our theories provide a foundation to suppress a wide variety of notions of complexity, including algorithmic communication complexity. By suppressing the amount of communication between slow and fast layers of memory, which we may regard as symbols generated and communicated through elementary operations, we are able to construct algorithms that improve performance and computational feasibility. Chapter 3 demonstrates this potential to improve the performance of algorithmic decision-making by leveraging randomized projection in Randomized QR with Column Pivoting (RQRCP). Randomized projection allows us to control an increase in uncertainty regarding the conditions used to make algorithmic decisions while gaining a substantial performance advantage.

These developments provide new insight into the constraints we must satisfy in order to obtain rigorous justification for the predictions we obtain from data through machine learning. Moreover, we see how these techniques also advance the pursuit of computational feasibility and rational decision-making.

Chapter 1

Generalizing Information to the Evolution of Rational Belief

Information theory provides a mathematical foundation to measure uncertainty in belief. Belief is represented by a probability distribution that captures our understanding of an outcome's plausibility. Information measures based on Shannon's concept of entropy include realization information, Kullback–Leibler divergence, Lindley's information in experiment, cross entropy, and mutual information.

We derive a general theory of information from first principles that accounts for evolving belief and recovers all of these measures. Rather than simply gauging uncertainty, information is understood in this theory to measure change in belief. We may then regard entropy as the information we expect to gain upon realization of a discrete latent random variable.

This theory of information is compatible with the Bayesian paradigm in which rational belief is updated as evidence becomes available. Furthermore, this theory admits novel measures of information with well-defined properties, which we explore in both analysis and experiment. This view of information illuminates the study of machine learning by allowing us to quantify information captured by a predictive model and distinguish it from residual information contained in training data. We gain related insights regarding feature selection, anomaly detection, and novel Bayesian approaches.

1. Introduction

This work integrates essential properties of information embedded within Shannon's derivation of entropy (Shannon, 1948) and the Bayesian perspective (LaPlace, 1774; Jeffreys, 1998; Jaynes, 2003), which identifies probability with plausibility. We pursue this investigation in order to understand how to rigorously apply information-theoretic concepts to the theory of inference and machine learning. Specifically, we would like to understand how to quantify the evolution of predictions given by machine learning models. Our findings are general, however, and bear implications for any situation in which states of belief are updated. We begin in 1.1 with an experiment that illustrates shortcomings with the way standard information measures would partition prediction information and residual information during machine learning training.

1.1 Shortcomings with standard approaches

Let us examine a typical MNIST (LeCun et al., 1998) classifier. This dataset comprises a set of images of handwritten digits paired with labels. Let both $\boldsymbol{x}$ and $\boldsymbol{y}$ denote random variables corresponding respectively to an image and a label in a pair. In this perspective, the training dataset contains independent realizations of such pairs from an unknown joint probability distribution. We would like to obtain a measurement of prediction information that quantifies a shift in belief from an uninformed initial state $\mathbf{q}_0(\boldsymbol{y})$ to model predictions $\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})$. The symmetric uninformed choice for $\mathbf{q}_0(\boldsymbol{y})$ is uniform probability over all outcomes. Note that both $\mathbf{q}_0(\boldsymbol{y})$ and $\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})$ are simply hypothetical states of belief. Some architectures may approximate Bayesian inference, but we cannot always interpret these as the Bayesian prior and posterior.

Two measurements that are closely related to Shannon's entropy are the Kullback–Leibler divergence (Kullback and Leibler, 1951; Kullback, 1997) and Lindley's information in experiment (Lindley, 1956), which are computed respectively as

$$D_{KL}[\,\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x}) \,\|\, \mathbf{q}_0(\boldsymbol{y})\,] = \int d\boldsymbol{y}\, \mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})}{\mathbf{q}_0(\boldsymbol{y})}\right) \quad \text{and}$$

$$D_{L}[\,\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x}) \,\|\, \mathbf{q}_0(\boldsymbol{y})\,] = \int d\boldsymbol{y}\, \mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x}) \log(\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})) - \int d\boldsymbol{y}\, \mathbf{q}_0(\boldsymbol{y}) \log(\mathbf{q}_0(\boldsymbol{y}))\,.$$

Whatever we choose, we would like to use a consistent construction to understand how much information remains unpredicted. After viewing a label outcome $\check{\boldsymbol{y}}$, we let $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$ represent our new understanding of the actual state of affairs, which is a realization assigning full probability to the specified outcome. This distribution captures our most updated knowledge about $\boldsymbol{y}$ and therefore constitutes rational belief. A consistent information measurement should then quantify residual information as the shift in belief from $\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})$ to $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$. For example, the KL version would be $D_{KL}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})\,]$.

In order to demonstrate shortcomings with each approach, some cases are deliberately mislabeled during model testing. We first compute information measurements assuming the incorrect labels hold. Mislabled cases are then corrected to $\hat{\boldsymbol{y}}$ with corrected belief given by $\mathbf{r}(\boldsymbol{y} \mid \hat{\boldsymbol{y}})$. This allows us to compare our first information measurements with corrected versions. An example of each belief state is shown in Figure 1.1 where the incorrect label **3** is changed to **0**.

Ideally, the sum of prediction information and residual information would be a conserved quantity, which would allow us to understand training as simply shifting information from a residual partition to the predicted partition. More importantly, however, prediction information should clearly capture prediction quality. Figure 1.2 shows information measurements corresponding to the example given.

Mislabeling and relabeling shows us that neither formulation of prediction information captures prediction quality. This is because these constructions simply have no affordance to account for our understanding of what is actually correct. Large KL residual information offers some indication of mislabeling and decreases when we correct the label, but total information is not conserved. As such, there is no intuitive notion for what appropriate prediction and residual information should be for a given problem. The Lindley formulation is substantially less satisfying; Although total information is conserved, neither metric changes and we have no indication of mislabeling.

Digit	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\mathbf{q}_0(\boldsymbol{y})$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})$	0.952	0.0	0.045	0.001	0.0	0.001	0.0	0.0	0.0	0.002
$\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
$\mathbf{r}(\boldsymbol{y} \mid \hat{\boldsymbol{y}})$	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

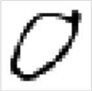

Figure 1.1: Example of the evolution of plausible labels for an image. Without evidence, the probability distribution $\mathbf{q}_0(\boldsymbol{y})$ assigns equal plausibility to all outcomes. A machine learning model processes the image and produces predictions $\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})$. The incorrect label **3** is represented by $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$. After observing the image, shown on the right, the label is corrected to **0** in $\mathbf{r}(\boldsymbol{y} \mid \hat{\boldsymbol{y}})$.

	Original Label			**Corrected Label**		
Information Type	**Prediction**	**Residual**	**Sum**	**Prediction**	**Residual**	**Sum**
Kullback–Leibler	3.02	10.44	13.46	3.02	0.07	3.09
Lindley	3.02	0.30	3.32	3.02	0.30	3.32

Figure 1.2: Information measurements before and after label correction. Neither construction of prediction information allows the computation to account for claimed labels. Residual information, however, decreases in the KL construction when the label is corrected. The Lindley forms are totally unaffected by relabeling.

The problem with these constructions is they do not recognize the gravity of the role of expectation. That is, reasonable expectation must be consistent with rational belief. We hold that our most justified understanding of what may be true provides a sound basis to measure changes in belief. Figure 1.3 gives a preview of information measurements in the framework of this theory. Total information, $\log_2(10)$ bits in this case, is conserved and both prediction information and residual information react intuitively to mislabeling. Determining whether the predictive information is positive or negative provides a clear indication of whether the prediction was informative.

	Original Label			**Corrected Label**		
Information Type	**Prediction**	**Residual**	**Sum**	**Prediction**	**Residual**	**Sum**
Proposed	-7.11	10.44	3.32	3.25	0.07	3.32

Figure 1.3: Information measurements using our proposed framework. Total information is a conserved quantity and when our belief changes, so do the information measurements. Negative prediction information forewarns either potential mislabeling or a poor prediction.

1.2 Our contributions

In the course of pursuing a consistent framework in which information measurements may be understood, we have derived a theory of information from first principles that places all entropic information measures in a unified interpretable context. By axiomatizing the properties of information we desire, we show that a unique formulation follows that subsumes critical properties of Shannon's construction of entropy.

This theory fundamentally understands entropic information as a form of reasonable expectation that measures the change between hypothetical belief states. Expectation is not necessarily taken with respect to the distributions that represent the shift in belief, but rather with respect to a third distribution representing our understanding of what may actually be true. We find compelling foundations for this perspective within the Bayesian philosophy of probability as an extended logic for expressing and updating uncertainty (Cox, 1946; Jaynes, 2003). Our understanding of what may be true, and therefore the basis for measuring information, should be rational belief. Rational belief (Ramsey, 2016; Lehman, 1955; Adams, 1962; Freedman and Purves, 1969; Skyrms, 1987) begins with probabilistically coherent prior knowledge and is subsequently updated to account for observations using Bayes' theorem. As a consequence, information associated with a change in belief is not a fixed quantity. Just as rational belief must evolve as new evidence becomes available, so also does the information we would reasonably assign to previous shifts in belief. By emphasizing the role of rational belief, this theory recognizes that the degree of validity we assign to past states of belief is both dynamic and potentially subjective as our state of knowledge matures.

As a consequence of enforcing consistency with rational belief, a second additivity property emerges; just as entropy can be summed over independent distributions, information gained over a sequence of observations can be summed over intermediate belief updates. Total information over such a sequence is independent of how results are grouped or ordered. This provides a compelling solution to the thought experiment above. Label information in training data is a conserved quantity and we motivate a formulation of prediction information that is directly tied to prediction quality.

Soofi, Ebrahimi, and others (Soofi, 1994, 2000; Ebrahimi et al., 2004, 2010) identify key contributions to information theory in the decade following Shannon's paper that are intrinsically tied to entropy. These are the Kullback–Leibler divergence, Lindley's information in experiment, and Jaynes' construction of entropy-maximizing distributions that are consistent with specified expectations. We show how this theory recovers these measures of information and admits new forms that may not have been previously associated with entropic information, such as the log pointwise posterior predictive measure of model accuracy (Gelman et al., 2013). We also show how this theory admits novel information-optimal probability distributions analogous to that of Jaynes' maximum uncertainty. Having a consistent interpretation of information illuminates how it may be applied and what properties will hold in a given context. Moreover, this theoretical framework enables us to solve multiple challenges in Bayesian learning. For example, one such challenge is understanding how efficiently a given model incorporates new data. This theory provides bounds on the information gained by a model resulting from inference and allows us to characterize the information provided by individual observations.

The rest of this paper is organized as follows. Section 2 discusses notation and background regarding entropic information, Bayesian inference, and reasonable expectation. Section 3 contains postulates that express properties of information we desire as well as the formulation of information that follows and other related measures of information. Section 4 analyzes general consequences and properties of this formulation. Section 5 discusses further implications with respect to Bayesian inference and machine learning. Section 6 explores negative information with computational experiments that illustrate when it occurs, how it may be understood, and why it is useful. Section 6 summarizes these results and offers a brief discussion of future work. Appendix A proves our principal result. Appendix B contains all corollary proofs. Appendix C provides key computations used in experiments.

2. Background and notation

Shannon's construction of entropy (Shannon, 1948) shares a fundamental connection with thermodynamics. The motivation is to facilitate analysis of complex systems which can be decomposed into independent subsystems. The essential idea is simple — when probabilities multiply, entropy adds. This abstraction allows us to compose uncertainties across independent sources by simply adding results. Shannon applied this perspective to streams of symbols called channels. The number of possible outcomes grows exponentially with the length of a symbol sequence, whereas entropy grows linearly. This facilitates a rigorous formulation of the rate of information conveyed by a channel as well as analysis of what is possible in the presence of noise.

The property of independent additivity is used in standard training practices for machine learning. Just as thermodynamic systems and streams of symbols break apart, so does an ensemble of predictions over independent observations. This allows us to partition training sets into batches and compute cross-entropy (Good, 1963) averages. MacKay (MacKay, 2003) gives a comprehensive discussion of information in the context of learning algorithms. Tishby (Tishby and Zaslavsky, 2015) examines information trends during neural network training.

A second critical property of entropy, which is implied by Shannon and further articulated by both Barnard (Barnard, 1951) and Rényi (Rényi, 1961), is that entropy is an expectation. Given a latent random variable $\boldsymbol{z}$, we denote the probability distribution over outcomes as $\mathbf{p}(\boldsymbol{z})$. Stated as an expectation, entropy is defined as

$$S[\,\mathbf{p}(\boldsymbol{z})\,] = \int d\boldsymbol{z}\,\mathbf{p}(\boldsymbol{z}) \log\left(\frac{1}{\mathbf{p}(\boldsymbol{z})}\right) = \mathbb{E}_{\mathbf{p}(\boldsymbol{z})} \log\left(\frac{1}{\mathbf{p}(\boldsymbol{z})}\right).$$

Following Shannon, investigators developed a progression of divergence measures between general probability distributions, $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$. Notable cases include the Kullback–Leibler divergence, Rényi's information of order-α (Rényi, 1961), and Csiszár's f-divergence (Csiszár, 1967). Ebrahimi, Soofi, and Soyer (Ebrahimi et al., 2010) offer an examination of these axiomatic foundations and generalizations with a primary focus on entropy and the KL divergence. Recent work on axiomatic foundations for generalized entropies (Jizba and Arimitsu, 2004) includes constructions that are suitable for strongly-interacting systems (Hanel and Thurner, 2011) and axiomatic derivations of other forms of entropy including SharmaMittal and FrankDaffertshofer entropies (Ilić and Stanković, 2014). Further work relates group-theoretic properties of systems to corresponding notions of entropy and correlation laws (Tempesta, 2011).

2.1 Bayesian reasoning

The Bayesian view of probability, going back to Laplace (LaPlace, 1774) and championed by Jeffreys (Jeffreys, 1998) and Jaynes (Jaynes, 2003), focuses on capturing our beliefs. This perspective considers a probability distribution as an abstraction that attempts to model these beliefs. This view subsumes all potential sources of uncertainty and provides a comprehensive scope that facilitates analysis in diverse contexts.

In the Bayesian framework, the prior distribution $\mathbf{p}(\boldsymbol{z})$ expresses initial beliefs about some latent variable $\boldsymbol{z}$. Statisticians, scientists, and engineers often have well-founded views about real-world systems that from the basis for priors. Examples include physically realistic ranges of model parameters or plausible responses of a dynamical system. In the case of total ignorance, one applies the principle of insufficient reason (Bernoulli, 1713) — we should not break symmetries of belief without justification. Jaynes' construction of maximally uncertainty distributions (Jaynes, 1957) generalizes this principle, which we discuss further in 4.6.

As observations $\boldsymbol{x}$ become available, we update belief from the prior distribution to obtain the posterior distribution $\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$, which incorporates this new knowledge. This update is achieved by applying Bayes' theorem

$$\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{\mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})\mathbf{p}(\boldsymbol{z})}{\mathbf{p}(\boldsymbol{x})} \quad \text{where} \quad \mathbf{p}(\boldsymbol{x}) = \int d\boldsymbol{z}\, \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})\mathbf{p}(\boldsymbol{z}).$$

The likelihood distribution $\mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})$ expresses the probability of observations given any specified value of $\boldsymbol{z}$. The normalization constant $\mathbf{p}(\boldsymbol{x})$ is also the probability of $\boldsymbol{x}$ given the prior belief that has been specified. Within Bayesian inference, this is also called model evidence and it is used to evaluate a model structure's plausibility for generating the observations.

Shore and Johnson (Shore and Johnson, 1980, 1981) provide an axiomatic foundation for updating belief that recovers the principles of maximum entropy and minimum cross-entropy when prior evidence consists of known expectations. For reference, we summarize these axioms as

1. *Uniqueness.* When belief is updated with new observations, the result should be unique.

2. *Coordinate invariance.* Belief updates should be invariant to arbitrary choices of coordinates.

3. *System independence.* The theory should yield consistent results when independent random variables are treated either separately or jointly.

4. *Subset independence.* When we partion potential outcomes into disjoint subsets, the belief update corresponding to conditioning on subset membership first should yield the same result as updating first and conditioning on the subset second.

Jizba and Korbel (Jizba and Korbel, 2019) investigate generalizations of entropy for which the maximum entropy principle satisfies these axioms.

Integrating the maturing notion of belief found within the Bayesian framework with information theory recognizes that our perception of how informative observations are depends on how our beliefs develop, which is dynamic as our state of knowledge grows.

2.2 Probability notation

Random variables are denoted in boldface such as $\boldsymbol{x}$. Typically $\boldsymbol{x}$ and $\boldsymbol{y}$ will imply observable measurements and $\boldsymbol{z}$ will indicate either a latent explanatory variable or unknown observable. Each random variable is implicitly associated with a corresponding probability space including the set of all possible outcomes $\Omega_{\boldsymbol{z}}$, a σ-algebra $\mathcal{F}_{\boldsymbol{z}}$ of measurable subsets, and a probability measure $\mathcal{P}_{\boldsymbol{z}}$ which maps subsets of events to probabilities. We then express the probability measure as a distribution function $\mathbf{p}(\boldsymbol{z})$.

A realization, or specific outcome, will be denoted with either a check $\check{\boldsymbol{z}}$ or, for discrete distributions only, a subscript $\boldsymbol{z}_i$ where $i \in [n]$ and $[n] = \{1, 2, \ldots, n\}$. If it is necessary to emphasize the value of a distribution at a specific point or realization, we will use the notation $\mathbf{p}(\boldsymbol{z} = \check{\boldsymbol{z}})$. Conditional dependence is denoted in the usual fashion as $\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$. The joint distribution is then $\mathbf{p}(\boldsymbol{x}, \boldsymbol{z}) = \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})\mathbf{p}(\boldsymbol{x})$ and marginalization is obtained by $\mathbf{p}(\boldsymbol{x}) = \int d\boldsymbol{z}\, \mathbf{p}(\boldsymbol{x}, \boldsymbol{z})$. When two distributions are equivalent over all subsets of nonzero measure, we use notation $\mathbf{q}_0(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z})$ or $\mathbf{q}_1(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$.

The probability measure allows us to compute expectations over functions $f(\boldsymbol{z})$ which are denoted

$$\mathbb{E}_{\mathbf{p}(\boldsymbol{z})}\, f(\boldsymbol{z}) = \int d\boldsymbol{z}\, \mathbf{p}(\boldsymbol{z}) f(\boldsymbol{z}).$$

The support of integration or summation is implied to be the same as the support of $\mathbf{p}(\boldsymbol{z})$, that is the set of outcomes for which $\mathbf{p}(\boldsymbol{z}) > 0$. For example, in both the discrete case above and continuous cases, such as a distribution on the unit interval $\boldsymbol{z} \in \mathbb{R}_{[0,1]}$, the integral notation should be interpreted respectively as

$$\int d\boldsymbol{z}\, \mathbf{p}(\boldsymbol{z}) f(\boldsymbol{z}) = \sum_{i=1}^{n} \mathbf{p}(\boldsymbol{z} = \boldsymbol{z}_i) f(\boldsymbol{z}_i) \quad \text{and} \quad \int d\boldsymbol{z}\, \mathbf{p}(\boldsymbol{z}) f(\boldsymbol{z}) = \int_0^1 d\check{\boldsymbol{z}}\, \mathbf{p}(\boldsymbol{z} = \check{\boldsymbol{z}}) f(\check{\boldsymbol{z}}).$$

2.3 Reasonable expectation and rational belief

The postulates and theory in this work concern the measurement of a shift in belief from an initial state $\mathbf{q}_0(\boldsymbol{z})$ to an updated state $\mathbf{q}_1(\boldsymbol{z})$. In principle, these are any *hypothetical states of belief.* For example, they could be predictions given by the computational model in 1.1, previous beliefs held before observing additional data, or convenient approximations of a more informed state of belief. A third state $\mathbf{r}(\boldsymbol{z})$, *rational belief*, serves a distinct role as the distribution over which expectation is taken. When we wish to emphasize this role, we also refer to $\mathbf{r}(\boldsymbol{z})$ as the *view of expectation.*

To understand the significance of rational belief, we briefly review work by Cox (Cox, 1946) regarding reasonable expectation from two perspectives on the meaning of probability. The first perspective understands probability as a description of relative frequencies in an ensemble. If we prepare a large ensemble of independent random variables, $\boldsymbol{Z} = \{\boldsymbol{z}_i \,|\, i \in$

$[n]\}$, and each is realized from a proper (normalized) probability distribution $\mathbf{p}(\boldsymbol{z})$, then the relative frequency of outcomes within each subset $\omega \in \Omega_{\boldsymbol{z}}$ will approach the probability measure $\mathcal{P}_{\boldsymbol{z}}(\omega)$ for large n. It follows that the ensemble mean of any transformation $f(\boldsymbol{z})$ will approach the expectation

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{z}_i) = \mathbb{E}_{\mathbf{p}(\boldsymbol{z})}\, f(\boldsymbol{z}).$$

The difficulty arises when we distinguish *what is true* from *what may be known*, given limited evidence. This falls within the purview of the second perspective, the Bayesian view, regarding probability as an extended logic. To illustrate, suppose $\boldsymbol{z}$ is the value of an unknown real mathematical constant. The true probability distribution would be a Dirac delta $\mathbf{p}(\boldsymbol{z}) \equiv \boldsymbol{\delta}(\boldsymbol{z} - \check{\boldsymbol{z}})$ assigning unit probability to the unknown value $\check{\boldsymbol{z}}$. Accordingly, each element in the ensemble above would take the same unknown value. If we have incomplete knowledge $\mathbf{r}(\boldsymbol{z})$ regarding the distribution of plausible values, then we can still compute an expectation $\mathbb{E}_{\mathbf{r}(\boldsymbol{z})}\, f(\boldsymbol{z})$, but we must bear in mind that the result only approximates the unknown true expectation. Since the expectation is limited by the credibility of $\mathbf{r}(\boldsymbol{z})$, we seek to drive belief towards the truth as efficiently as possible from available evidence to fulfill this role.

Within Bayesian Epistemology, rational belief is defined as a belief that is unsusceptible to a Dutch Book. When an agent's beliefs correspond to their willingness to places bets, a Dutch Book (Lehman, 1955; Adams, 1962; Hájek, 2008; Freedman and Purves, 1969; Skyrms, 1987) means that it is possible for a bookie to construct a table of bets that the agent finds acceptable but also guarantees that the agent will lose money. Therefore the existence of such a table corresponds to the agent holding an irrational state of belief. When multiple bets are allowed to be conditioned on a sequence of outcomes, it has been shown that the agent must use Bayes' Theorem to account for previous outcomes in the sequence to update beliefs regarding subsequent outcomes to avoid irrationality (Skyrms, 1987).

For our purposes, it is sufficient to say that if we have a coherent prior belief in a latent variable as well as a likelihood function that implies beliefs about observations, Bayes' theorem incorporates observational evidence to from the posterior distribution representing rational belief. For example, we could measure inference information from prior belief $\mathbf{q}_0(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z})$ to the *first* posterior $\mathbf{q}_1(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$ conditioned on an observation $\boldsymbol{x}$. When we have additional evidence $\boldsymbol{y}$ that complements $\boldsymbol{x}$, then rational belief must corresponds to a *second* inference $\mathbf{r}(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ because retaining the belief $\mathbf{q}_1(\boldsymbol{z})$ would not account for $\boldsymbol{y}$. Likewise, if $\boldsymbol{z}$ is an observable realization, then rational belief must assign full probability to the observed outcome $\check{\boldsymbol{z}}$. This case is specifically denoted as $\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}})$ and in continuous settings it is equivalent to the Dirac delta function $\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}}) \equiv \boldsymbol{\delta}(\boldsymbol{z} - \check{\boldsymbol{z}})$.

2.4 Remarks on Bayesian objectivism and subjectivism

Within the Bayesian philosophy, we may disagree about whether or not rational belief is unique. This disagreement corresponds to objectivist versus subjectivist views of Bayesian epistemology, see (Weisberg, 2011; Jaynes, 2003) for a discussion. Note that this is not the same as the more general view of objective versus subjective probabilities.

In the objectivist's view, one's beliefs must be consistent with the entirety of evidence and prior knowledge must be justified by sound principles of reason. Therefore, anyone with the same body of evidence must hold the same rational belief. In contrast, the subjectivist holds that one's prior beliefs do not need justification. Provided evidence is taken into account using Bayes' theorem, the resulting posterior is rational for any prior as long as the prior is coherent. Note that the subjectivist view does not imply that all beliefs are equally valid. It simply allows validity in the construction of prior belief to be derived from other notions of utility, such as computational feasibility.

While the following postulates in 3 and derivation of Theorem 1 do not require adoption of either perspective, these philosophies influence how we understand reasonable expectation. The objective philosophy implies that an information measurement is justified to the same degree as the view of expectation that defines it, whereas the subjective philosophy entertains information analysis with any view of expectation.

3. Information and evolution of belief

In order to provide context for comparison, we begin by presenting the properties of entropic information originally put forward by Shannon using our notation.

3.1 Shannon's properties of entropy

1. Given a discrete probability distribution $\mathbf{p}(\boldsymbol{z})$ for which $\boldsymbol{z} \in \{\boldsymbol{z}_i \,|\, i \in [n]\}$, the entropy $S[\,\mathbf{p}(\boldsymbol{z})\,]$ is continuous in the probability of each outcome $\mathbf{p}(\boldsymbol{z} = \boldsymbol{z}_i)$.

2. If all outcomes are equally probable, namely $\mathbf{p}(\boldsymbol{z} = \boldsymbol{z}_i) = 1/n$, then $S[\,\mathbf{p}(\boldsymbol{z})\,]$ is monotonically increasing in n.

3. The entropy of a joint random variable $S[\,\mathbf{p}(\boldsymbol{z}, \boldsymbol{w})\,]$ can be decomposed using a chain rule expressing conditional dependence

$$S[\,\mathbf{p}(\boldsymbol{z}, \boldsymbol{w})\,] = S[\,\mathbf{p}(\boldsymbol{z})\,] + \mathbb{E}_{\mathbf{p}(\boldsymbol{z})}\, S[\,\mathbf{p}(\boldsymbol{w} \mid \boldsymbol{z})\,]\,.$$

The first point is aimed at extending Shannon's derivation, which employs rational probabilities, to real-valued probabilities. The second point drives at understanding entropy as a measure of uncertainty; as the number of possible outcomes increases, each realization becomes less predictable. This results in entropy taking positive values. The third point is critical — not only does it encode independent additivity, it implies that entropic information is computed as an expectation.

We note that Fadeeve (Fadeev, 1957) gives a simplified set of postulates. Rènyi (Rényi, 1961) generalizes information by replacing the last point with a weaker version which simply requires independent additivity, but not conditional expectation. This results in α-divergences. Csiszàr (Csiszár, 1967) generalizes this further using convex functions f to obtain f-divergences.

3.2 Postulates

Rather than repeating direct analogs of Shannon's properties in the context of evolving belief, it is both simpler and more illuminating to be immediately forthcoming regarding the key requirement of information in the perspective of this theory.

Postulate 1 *Entropic information associated with the change in belief from* $\mathbf{q}_0(\boldsymbol{z})$ *to* $\mathbf{q}_1(\boldsymbol{z})$ *is quantified as an expectation over belief* $\mathbf{r}(\boldsymbol{z})$*, which we call the view of expectation. As an expectation, it must have the functional form*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z})\,\|\,\mathbf{q}_0(\boldsymbol{z})\,] = \int d\boldsymbol{z}\,\mathbf{r}(\boldsymbol{z}) f(\mathbf{r}(\boldsymbol{z}), \mathbf{q}_1(\boldsymbol{z}), \mathbf{q}_0(\boldsymbol{z}))\,.$$

Postulate 2 *Entropic information is additive over independent belief processes. Taking joint distributions associated with two independent random variables* $\boldsymbol{z}$ *and* $\boldsymbol{w}$ *to be* $\mathbf{q}_0(\boldsymbol{z},\boldsymbol{w}) = \mathbf{q}_0(\boldsymbol{z})\mathbf{q}_0(\boldsymbol{w})$*,* $\mathbf{q}_1(\boldsymbol{z},\boldsymbol{w}) = \mathbf{q}_1(\boldsymbol{z})\mathbf{q}_1(\boldsymbol{w})$*, and* $\mathbf{r}(\boldsymbol{z},\boldsymbol{w}) = \mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w})$ *gives*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w})}[\,\mathbf{q}_1(\boldsymbol{z})\mathbf{q}_1(\boldsymbol{w})\,\|\,\mathbf{q}_0(\boldsymbol{z})\mathbf{q}_0(\boldsymbol{w})\,] = \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z})\,\|\,\mathbf{q}_0(\boldsymbol{z})\,] + \mathbb{I}_{\mathbf{r}(\boldsymbol{w})}[\,\mathbf{q}_1(\boldsymbol{w})\,\|\,\mathbf{q}_0(\boldsymbol{w})\,]\,.$$

Postulate 3 *If belief does not change then no information is gained, regardless of the view of expectation,*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_0(\boldsymbol{z})\,\|\,\mathbf{q}_0(\boldsymbol{z})\,] = 0.$$

Postulate 4 *The information gained from any normalized prior state of belief* $\mathbf{q}_0(\boldsymbol{z})$ *to an updated state of belief* $\mathbf{r}(\boldsymbol{z})$ *in the view of* $\mathbf{r}(\boldsymbol{z})$ *must be nonnegative*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{r}(\boldsymbol{z})\,\|\,\mathbf{q}_0(\boldsymbol{z})\,] \geq 0.$$

The first postulate requires information to be reassessed as belief changes. The most justified state of belief, based on the entirety of observations, will correspond to the most justified view of information. The second postulate is the additive form of Shore and Johnson's Axiom 3, system independence. That is, we need some law of composition, addition in this case, that allows independent random variables to be treated separately and arrive at the same result as treating them jointly.

By combining the first two postulates, it is possible to show that $f(r,q,p) = \log\left(r^{\gamma}q^{\alpha}p^{\beta}\right)$ for constants α, β, γ. See A for details. The third postulate constrains these exponential constants and the fourth simply sets the sign of information.

3.3 Principal result

Theorem 1 ***Information as a measure of change in belief.*** *Information measurements that satisfy these postulates must take the form*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z})\,\|\,\mathbf{q}_0(\boldsymbol{z})\,] = \alpha \int d\boldsymbol{z}\,\mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) \quad \textit{for some} \quad \alpha > 0.$$

Proof is given in A. As Shannon notes regarding entropy, α corresponds to a choice of units. Typical choices are natural units $\alpha = 1$ and bits $\alpha = \log(2)^{-1}$. We employ natural units in analysis and bits in experiments.

Although it would be possible to combine Postulate 1 and Postulate 2 into an analog of Shannon's chain rule as a single postulate, doing so would obscure the reasoning behind the construction. We leave the analogous chain rule as a consequence in Corollary 1. Regarding Shannon's proof that entropy is the only construction that satisfies properties he provides, we observe that he has restricted attention to functionals acting upon a single distribution. The interpretation of entropy is discussed in 4.1.

Normalization of $\mathbf{r}(\boldsymbol{z})$ is a key property of rational belief and reasonable expectation. As for $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$, however, nothing postulated prevents analysis respecting improper or non-normalizable probability distributions. In the Bayesian context, such distributions merely represent relative plausibility among subsets of outcomes. We caution that such analysis is a further abstraction, which requires additional care for consistent interpretation.

We remark that although Rènyi and Csiszàr were able to generalize divergence measures by weakening Shannon's chain rule to independent additivity, inclusion of the first postulate prevents such generalizations. We suspect, however, that if we replace Postulate 1 with an alternative functional that incorporates rational belief into information measurements, or we replace Postulate 2 with an alternative formulation of system independence, then other compelling information theories would follow.

3.4 Regarding the support of expectation

The proof given assumes $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$ take positive values over the support of the integral, which is also the support of $\mathbf{r}(\boldsymbol{z})$. In the Bayesian context, we also have

$$\mathbf{q}_1(\boldsymbol{z}) = \frac{\mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})\mathbf{q}_0(\boldsymbol{z})}{\mathbf{p}(\boldsymbol{x})} \quad \text{and} \quad \mathbf{r}(\boldsymbol{z}) = \frac{\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{x})\mathbf{q}_1(\boldsymbol{z})}{\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x})}$$

Accordingly, if for some $\check{\boldsymbol{z}}$ we have $\mathbf{q}_1(\check{\boldsymbol{z}}) = 0$ it follows that $\mathbf{r}(\check{\boldsymbol{z}}) = 0$. Likewise, $\mathbf{q}_0(\check{\boldsymbol{z}}) = 0$ would imply both $\mathbf{q}_1(\check{\boldsymbol{z}}) = 0$ and $\mathbf{r}(\check{\boldsymbol{z}}) = 0$. This forbids information contributions that fall beyond the scope of the proof. Even so, the resulting form is analytic and admits analytic continuation.

Since both $\lim_{\varepsilon\to 0}\left[\varepsilon \log \varepsilon\right] = 0$ and $\lim_{\varepsilon\to 0}\left[\varepsilon \log \varepsilon^{-1}\right] = 0$, limits of information of the form

$$\lim_{\varepsilon\to 0} \varepsilon \log\left(\frac{q_1}{q_0}\right), \quad \lim_{\varepsilon\to 0} \varepsilon \log\left(\frac{q_1}{\varepsilon}\right), \quad \lim_{\varepsilon\to 0} \varepsilon \log\left(\frac{\varepsilon}{q_0}\right), \quad \text{and} \quad \lim_{\varepsilon\to 0} \varepsilon \log\left(\frac{\varepsilon}{\varepsilon}\right)$$

are consistent with restricting the domain of integration (or summation) to the support of $\mathbf{r}(\boldsymbol{z})$. We gain further insight by considering limits of the form

$$\lim_{\varepsilon\to 0} r \log\left(\frac{\varepsilon}{q}\right) \quad \text{and} \quad \lim_{\varepsilon\to 0} r \log\left(\frac{q}{\varepsilon}\right).$$

Information diverges to $-\infty$ in the first case and $+\infty$ in the second. This is consistent with the fact that no finite amount of data will recover belief over a subset that has been strictly forbidden from consideration, which bears ramifications for how we understand rational belief.

If belief is not subject to influence from evidence, it is difficult to credibly construe an inferred outcome as having rationally accounted for that evidence. Lindley calls this

Cromwell's rule (Lindley, 1980); we should not eliminate a potential outcome from consideration unless it is logically false. The principle of insufficient reason goes further by avoiding unjustified creation of information that is not influenced by evidence.

3.5 Information density

The Radon-Nikodym theorem (Nikodym, 1930) formalizes the notion of density that relates two measures. If we assign both probability and a second measure to any subset within a probability space, then there exists a density function, unique up to subsets of measure zero, such that the second measure is equivalent to the integral of said density over any subset.

Definition 1 ***Information density.*** *We take the Radon-Nikodym derivative to obtain information density of the change in belief from* $\mathbf{q}_0(\boldsymbol{z})$ *to* $\mathbf{q}_1(\boldsymbol{z})$

$$\mathbb{D}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] = \frac{d\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,]}{d\mathbf{r}(\boldsymbol{z})} = \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})} \right).$$

The key property we find in this construction is independence from the view of expectation. As such, information density encodes all potential information outcomes one could obtain from this theory. Furthermore, this formulation is amenable to analysis of improper distributions. For example, it proves useful to consider information density corresponding to constant unit probability density $\mathbf{q}_1(\boldsymbol{z}) \equiv 1$, which is discussed further in 4.1.

3.6 Information pseudometrics

The following pseudometrics admit interpretations as notions of distance between belief states that remain compatible with Postulate 1. This is achieved by simply taking the view of expectation $\mathbf{r}(\boldsymbol{z})$ to be the weight function in weighted-L^p norms of information density. These constructions then satisfy useful properties of pseudometrics:

1. *Positivity*, $\mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] \geq 0$,
2. *Symmetry*, $\mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] = \mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_0(\boldsymbol{z}) \,\|\, \mathbf{q}_1(\boldsymbol{z}) \,]$,
3. *Triangle inequality*,

$$\mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] \leq \mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_1(\boldsymbol{z}) \,] + \mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,].$$

Definition 2 L^p ***information pseudometrics.*** *We may construct pseudometrics that measure distance between states of belief* $\mathbf{q}_0(\boldsymbol{z})$ *and* $\mathbf{q}_1(\boldsymbol{z})$ *with the view of expectation* $\mathbf{r}(\boldsymbol{z})$*, by taking weighted-*$L^p$ *norms of information density where the view of expectation serves as the weight function*

$$\mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] = \left(\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \left| \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})} \right) \right|^p \right)^{1/p} \quad \textit{for some} \quad p \geq 1.$$

Note that taking $p = 1$ results in a pseudometric that is also a pure expectation. The *homogeneity* property of seminorms, $\|\alpha x\| = |\alpha|\|x\|$ for $\alpha \in \mathbb{R}$, implies that these constructions retain the units of measure of information density; if information density is measured in bits, these distances have units of bits as well. Symmetry is obvious from inspection and the other properties follow by construction as seminorms. Specifically, positivity follows from the fact that $|\cdot|^p$ is a convex function for $p \geq 1$. The lower bound immediately follows from Jensen's inequality

$$\mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] \geq \left|\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]\right| \geq 0.$$

A short proof of the triangle inequality is given in B.

We observe that if $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$ are measurably distinct over the support of $\mathbf{r}(\boldsymbol{z})$ then the measured distance must be greater than zero. We may regard states of belief $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$ as weakly equivalent *in the view of* $\mathbf{r}(\boldsymbol{z})$ if their difference is immeasurable over the support of $\mathbf{r}(\boldsymbol{z})$. That is, if $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$ only differ over subsets of outcomes that are deemed by $\mathbf{r}(\boldsymbol{z})$ to be beyond plausible consideration, then in the view of $\mathbf{r}(\boldsymbol{z})$ they are equivalent. As such, these pseudometrics could be regarded as subjective metrics in the view of $\mathbf{r}(\boldsymbol{z})$. The natural definition of information variance also satisfies the properties of a pseudometric and is easily interpreted as a standard statistical construct.

Definition 3 ***Information variance.*** *Information variance between belief states* $\mathbf{q}_0(\boldsymbol{z})$ *and* $\mathbf{q}_1(\boldsymbol{z})$ *in the view of expectation* $\mathbf{r}(\boldsymbol{z})$ *is simply the variance of information density*

$$\mathrm{Var}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \left(\log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) - \varphi\right)^2$$

where $\varphi = \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$.

4. Corollaries and Interpretations

The following corollaries examine primary consequences of Theorem 1. Note that multiple random variables may be expressed as a single joint variable such as $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)$. The following corollaries explore one or two components at a time such as variables $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ or observations $\boldsymbol{x}$ and $\boldsymbol{y}$. Extensions to multiple random variables easily follow.

Note that the standard formulation of conditional dependence holds for all probability distributions in Corollary 1. That is, given an arbitrary joint distribution $\mathbf{q}(\boldsymbol{z}_1, \boldsymbol{z}_2)$, we can compute the marginalization as $\mathbf{q}(\boldsymbol{z}_1) \equiv \int d\boldsymbol{z}_2\, \mathbf{q}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ and conditional dependence follows by the Radon-Nikodym derivative to obtain $\mathbf{q}(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \equiv \frac{\mathbf{q}(\boldsymbol{z}_1, \boldsymbol{z}_2)}{\mathbf{q}(\boldsymbol{z}_1)}$. All proofs are contained in B.

Corollary 1 ***Chain rule of conditional dependence.*** *Information associated with joint variables decomposes as*

$$\begin{aligned}\mathbb{I}_{\mathbf{r}(\boldsymbol{z}_1,\boldsymbol{z}_2)}[\,\mathbf{q}_1(\boldsymbol{z}_1, \boldsymbol{z}_2) \,\|\, \mathbf{q}_0(\boldsymbol{z}_1, \boldsymbol{z}_2)\,] = {}& \mathbb{I}_{\mathbf{r}(\boldsymbol{z}_1)}[\,\mathbf{q}_1(\boldsymbol{z}_1) \,\|\, \mathbf{q}_0(\boldsymbol{z}_1)\,] \\ & + \mathbb{E}_{\mathbf{r}(\boldsymbol{z}_1)}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{z}_2|\boldsymbol{z}_1)}[\,\mathbf{q}_1(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \,\|\, \mathbf{q}_0(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)\,]\,.\end{aligned}$$

Corollary 2 ***Additivity over belief sequences.*** *Information gained over a sequence of belief updates is additive within the same view. Given initial belief* $\mathbf{q}_0(\boldsymbol{z})$*, intermediate states* $\mathbf{q}_1(\boldsymbol{z})$ *and* $\mathbf{q}_2(\boldsymbol{z})$*, and the view* $\mathbf{r}(\boldsymbol{z})$ *we have*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_1(\boldsymbol{z})\,] + \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]\,.$$

Corollary 3 ***Antisymmetry.*** *Information from* $\mathbf{q}_1(\boldsymbol{z})$ *to* $\mathbf{q}_0(\boldsymbol{z})$ *is the negative of information from* $\mathbf{q}_0(\boldsymbol{z})$ *to* $\mathbf{q}_1(\boldsymbol{z})$

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_0(\boldsymbol{z}) \,\|\, \mathbf{q}_1(\boldsymbol{z})\,] = -\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]\,.$$

4.1 Entropy

Shannon's formalization of entropy as uncertainty may be consistently understood as the expectation of information gained by realization. We first reconstruct information contained in realization. We then define the general form of entropy in the discrete case, which is cross entropy, and finally the standard form of entropy follows.

Corollary 4 ***Realization information (discrete).*** *Let* $\boldsymbol{z}$ *be a discrete random variable* $\boldsymbol{z} \in \{\boldsymbol{z}_i \,|\, i \in [n]\}$*. Information gained by realization* $\check{\boldsymbol{z}}$ *from* $\mathbf{q}(\boldsymbol{z})$ *in the view of realization* $\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}})$ *is*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\,\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}}) \,\|\, \mathbf{q}(\boldsymbol{z})\,] = \mathbb{D}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z} = \check{\boldsymbol{z}})\,]\,.$$

Corollary 5 ***Cross entropy (discrete).*** *Let* $\boldsymbol{z}$ *be a discrete random variable* $\boldsymbol{z} \in \{\boldsymbol{z}_i \,|\, i \in [n]\}$ *and* $\check{\boldsymbol{z}}$ *be a hypothetical realization. Expectation over the view* $\mathbf{r}(\check{\boldsymbol{z}})$ *of information gained by realization from belief* $\mathbf{q}(\boldsymbol{z})$ *recovers cross entropy*

$$\mathbb{E}_{\mathbf{r}(\check{\boldsymbol{z}})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\,\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}}) \,\|\, \mathbf{q}(\boldsymbol{z})\,] = \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z})\,] = S_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}(\boldsymbol{z})\,]\,.$$

Corollary 6 ***Entropy (discrete).*** *Let* $\boldsymbol{z}$ *be a discrete random variable* $\boldsymbol{z} \in \{\boldsymbol{z}_i \,|\, i \in [n]\}$ *and* $\check{\boldsymbol{z}}$ *be a hypothetical realization. Expectation over plausible realizations* $\mathbf{q}(\check{\boldsymbol{z}})$ *of information gained by realization from belief* $\mathbf{q}(\boldsymbol{z})$ *recovers entropy*

$$\mathbb{E}_{\mathbf{q}(\check{\boldsymbol{z}})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z})\,] = \mathbb{I}_{\mathbf{q}(\boldsymbol{z})}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z})\,] = S[\,\mathbf{q}(\boldsymbol{z})\,]\,.$$

Shannon proved that this is the only construction as a functional acting on a single distribution $\mathbf{q}(\boldsymbol{z})$ that satisfies his properties. As mentioned earlier, the information notation $\mathbb{I}_{\mathbf{q}(\boldsymbol{z})}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z})\,]$ requires some subtlety of interpretation. Probability density 1 over all discrete outcomes $\boldsymbol{z} \in \Omega_{\boldsymbol{z}}$ is not generally normalized. Although these formulas are convenient abstractions that result from formal derivations as expectations in the discrete case, nothing prevents us from applying them in continuous settings, which recovers the typical definitions in such cases.

In the continuous setting, we must emphasize that this definition of entropy is not consistent with taking the limit of a sequence of discrete distributions that converges in probability density to a continuous limiting distribution. The entropy of such a sequence diverges to infinity, which matches our intuition; the number of bits required to specify a continuous (real) random variable also diverges.

4.2 Information in an observation

As discussed in 2.3, we may regard $\mathbf{q}_0(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z})$ as prior belief and $\mathbf{q}_1(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$ as the posterior conditioned on the observation of $\boldsymbol{x}$. Without any additional evidence, we must hold $\mathbf{r}(\boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$ to be rational belief and we recover the Kullback–Liebler divergence as the rational measure of information gained by the observation of $\boldsymbol{x}$, but with a caveat; once we obtain additional evidence $\boldsymbol{y}$ then information in the observation of $\boldsymbol{x}$ must be recomputed as $\mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, \mathbf{p}(\boldsymbol{z})\,]$. In contrast, this theory holds that Lindley's corresponding measure

$$D_L[\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, \mathbf{p}(\boldsymbol{z})\,] = S[\,\mathbf{p}(\boldsymbol{z})\,] - S[\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})\,]$$

is not the information gained by the observation of $\boldsymbol{x}$; it is simply the difference in uncertainty before and after the observation.

4.3 Potential information

We now consider expectations over hypothetical future observations $\boldsymbol{w}$ that would influence belief in $\boldsymbol{z}$ as a latent variable. Given belief $\mathbf{p}(\boldsymbol{z})$, the probability of an observation $\boldsymbol{w}$ is $\mathbf{p}(\boldsymbol{w}) = \int d\boldsymbol{z}\, \mathbf{p}(\boldsymbol{w} \mid \boldsymbol{z})\mathbf{p}(\boldsymbol{z})$ as usual.

Corollary 7 ***Consistent future expectation.*** *Let the view $\mathbf{p}(\boldsymbol{z})$ express present belief in the latent variable $\boldsymbol{z}$ and $\boldsymbol{w}$ represent a future observation. The expectation over plausible $\boldsymbol{w}$ of information in the belief-shift from $\mathbf{q}_0(\boldsymbol{z})$ to $\mathbf{q}_1(\boldsymbol{z})$ in the view of rational future belief $\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w})$ is equal to information in the present view*

$$\mathbb{E}_{\mathbf{p}(\boldsymbol{w})}\, \mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{w})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \mathbb{I}_{\mathbf{p}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]\,.$$

Corollary 8 ***Mutual information.*** *Let the view $\mathbf{p}(\boldsymbol{z})$ express present belief in the latent variable $\boldsymbol{z}$ and $\boldsymbol{w}$ represent a future observation. Expectation of information gained by a future observation $\boldsymbol{w}$ is mutual information*

$$\mathbb{E}_{\mathbf{p}(\boldsymbol{w})}\, \mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{w})}[\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w}) \,\|\, \mathbf{p}(\boldsymbol{z})\,] = \mathbb{I}_{\mathbf{p}(\boldsymbol{z},\boldsymbol{w})}[\,\mathbf{p}(\boldsymbol{z},\boldsymbol{w}) \,\|\, \mathbf{p}(\boldsymbol{z})\mathbf{p}(\boldsymbol{w})\,]\,.$$

Corollary 9 ***Realization limit.*** *Let $\boldsymbol{z}$ be a latent variable and $\check{\boldsymbol{z}}$ be the limit of increasing observations to obtain arbitrary precision over plausible values of $\boldsymbol{z}$. Information gained from $\mathbf{q}_0(\boldsymbol{z})$ to $\mathbf{q}_1(\boldsymbol{z})$ in the realization limit $\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}})$ is pointwise information density*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \mathbb{D}[\,\mathbf{q}_1(\boldsymbol{z} = \check{\boldsymbol{z}}) \,\|\, \mathbf{q}_0(\boldsymbol{z} = \check{\boldsymbol{z}})\,]\,.$$

4.4 Consistent optimization analysis

Bernardo (Bernardo, 1979) shows that integrating entropy-like information measures with Bayesian inference provides a logical foundation for rational experimental design. He considers potential utility functions, or objectives for optimization, which are formulated as kernels of expectation over posterior belief updated by the outcome of an experiment. Bernardo then distinguishes the belief a scientist *reports* from belief that is *justified* by inference.

For a utility function to be *proper*, the Bayesian posterior must be the unique optimizer of expected utility over all potentially reported beliefs. In other words, a proper utility function must not provide an incentive to lie. His analysis shows that Lindley information is a proper utility function. Corollary 10 holds that information in this theory also provides proper utility. Thus information measures are not simply ad hoc objectives; they facilitate consistent optimization-based analysis that recovers rational belief.

Corollary 10 ***Information is a proper utility function.*** *Taking the rational view* $\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})$ *over the latent variable* $\boldsymbol{z}$ *conditioned upon an experimental outcome* $\boldsymbol{x}$*, the information* $\mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{x})}[\,\mathbf{q}(\boldsymbol{z}) \,\|\, \mathbf{p}(\boldsymbol{z})\,]$ *from prior belief* $\mathbf{p}(\boldsymbol{z})$ *to reported belief* $\mathbf{q}(\boldsymbol{z})$ *is a proper utility function. That is, the unique optimizer recovers rational belief*

$$\mathbf{q}^{*}(\boldsymbol{z}) \equiv \underset{\mathbf{q}(\boldsymbol{z})}{\mathit{argmax}}\, \mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{x})}[\,\mathbf{q}(\boldsymbol{z}) \,\|\, \mathbf{p}(\boldsymbol{z})\,] \equiv \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x}).$$

We would like to go a step further and show that when information from $\mathbf{q}_0(\boldsymbol{z})$ to $\mathbf{q}_1(\boldsymbol{z})$ is positive in the view of $\mathbf{r}(\boldsymbol{z})$, we may claim that $\mathbf{q}_1(\boldsymbol{z})$ is closer to $\mathbf{r}(\boldsymbol{z})$ than $\mathbf{q}_0(\boldsymbol{z})$. For this claim to be consistent we must show that any perturbation that unambiguously drives belief $\mathbf{q}_1(\boldsymbol{z})$ toward the view $\mathbf{r}(\boldsymbol{z})$ must also increase information. The complementary perturbation response with respect to $\mathbf{q}_0(\boldsymbol{z})$ immediately follows by Corollary 3.

Corollary 11 ***Proper perturbation response.*** *Let* $\mathbf{q}_1(\boldsymbol{z})$ *be measurably distinct from the view* $\mathbf{r}(\boldsymbol{z})$ *and* $\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$ *be finite. Let the perturbation* $\boldsymbol{\eta}(\boldsymbol{z})$ *preserve normalization and drive belief toward* $\mathbf{r}(\boldsymbol{z})$ *on all measurable subsets. It follows*

$$\lim_{\varepsilon \to 0} \frac{\partial}{\partial \varepsilon} \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) + \varepsilon \boldsymbol{\eta}(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] > 0.$$

It bears repeating, by Corollary 8, that mutual information captures expected proper utility, which provides a basis for rational experimental design and feature selection.

4.5 Discrepancy functions

Ebrahimi, Soofi, and Soyer (Ebrahimi et al., 2010) discuss information discrepancy functions, which have two key properties. First, a discrepancy function is nonnegative

$$\mathcal{D}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] \geq 0 \quad \text{with equality if and only if} \quad \mathbf{q}_1(\boldsymbol{z}) \equiv \mathbf{q}_0(\boldsymbol{z}).$$

Second, if we hold $\mathbf{q}_0(\boldsymbol{z})$ fixed then $\mathcal{D}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$ is convex in $\mathbf{q}_1(\boldsymbol{z})$. One of the reasons information discrepancy functions are useful is that they serve to identify independence. Random variables $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent if and only if $\mathbf{p}(\boldsymbol{x}) \equiv \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})$. Therefore, we have $\mathcal{D}[\,\mathbf{p}(\boldsymbol{x}, \boldsymbol{z}) \,\|\, \mathbf{p}(\boldsymbol{x})\mathbf{p}(\boldsymbol{z})\,] \geq 0$ with equality if and only if $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent, noting that $\mathbf{p}(\boldsymbol{x}, \boldsymbol{z}) \equiv \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})\mathbf{p}(\boldsymbol{z})$. This has implications regarding sensible generalizations of mutual information.

Theorem 1 does not satisfy information discrepancy properties unless the view of expectation is taken to be $\mathbf{r}(\boldsymbol{z}) \equiv \mathbf{q}_1(\boldsymbol{z})$, which is the KL divergence. We note, however, that information pseudometrics and information variance given in 3.6 satisfy a weakened formulation. Specifically, $\mathbb{L}^{p}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] \geq 0$ with equality if and only if $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$ are weakly equivalent in the view of $\mathbf{r}(\boldsymbol{z})$. Likewise, these formulations are convex in information density $\mathbb{D}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$.

4.6 Jaynes maximal uncertainty

Jaynes uses entropy to analytically construct a unique probability distribution for which uncertainty is maximal while maintaining consistency with a specified set of expectations. This construction avoids unjustified creation of information and places the principle of insufficient reason into an analytic framework within which the notion of symmetry generalizes to informational symmetries conditioned upon observed expectations.

We review how Jaynes constructs the resulting distribution $\mathbf{r}^*(\boldsymbol{z})$. Let such kernels of expectation be denoted $f_i(\boldsymbol{z})$ for $i \in [n]$ and the observed expectations be $\mathbb{E}_{\mathbf{r}(\boldsymbol{z})}[f_i(\boldsymbol{z})] = \varphi_i$. The objective of optimization is

$$\mathbf{r}^*(\boldsymbol{z}) = \underset{\mathbf{r}(\boldsymbol{z})}{\operatorname{argmax}}\, S[\,\mathbf{r}(\boldsymbol{z})\,] \quad \text{subject to} \quad \mathbb{E}_{\mathbf{r}(\boldsymbol{z})}\, f_i(\boldsymbol{z}) = \varphi_i \quad \forall\, i \in [n].$$

The Lagrangian, which captures both the uncertainty objective and expectation constraints, is

$$\mathcal{L}[\,\mathbf{r}(\boldsymbol{z}), \lambda\,] = \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \left(\log\!\left(\frac{1}{\mathbf{r}(\boldsymbol{z})} \right) - \sum_{i=1}^{n} \lambda_i \left(f_i(\boldsymbol{z}) - \varphi_i \right) \right).$$

where $\lambda \in \mathbb{R}^n$ is the vector of Lagrange multipliers. This Lagrangian formulation satisfies the variational principle in both $\mathbf{r}(\boldsymbol{z})$ and λ. Variational analysis yields the optimizer

$$\mathbf{r}^*(\boldsymbol{z}) \propto \exp\!\left(\sum_{i=1}^{n} \lambda_i f_i(\boldsymbol{z}) \right).$$

Information-critical distributions

Rather than maximizing entropy, we may minimize $\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{r}(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$ while maintaining consistency with specified expectations. Since the following corollary holds for general distributions $\mathbf{q}_0(\boldsymbol{z})$, including the improper case $\mathbf{q}_0(\boldsymbol{z}) \equiv 1$, this includes Jaynes' maximal uncertainty as a minimization of negative entropy.

Corollary 12 ***Minimal information.*** *Given kernels of expectation $f_i(\boldsymbol{z})$ and specified expectations $\mathbb{E}_{\mathbf{r}(\boldsymbol{z})}[f_i(\boldsymbol{z})] = \varphi_i$ for $i \in [n]$, the distribution $\mathbf{r}^*(\boldsymbol{z})$ that satisfies these constraints while minimizing information $\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{r}(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$ is given by*

$$\mathbf{r}^*(\boldsymbol{z}) \propto \mathbf{q}_0(\boldsymbol{z}) \exp\!\left(\sum_{i=1}^{n} \lambda_i f_i(\boldsymbol{z}) \right) \quad \textit{for some} \quad \lambda \in \mathbb{R}^n.$$

4.7 Remarks on Fisher information

Fisher provides an analytic framework to assess the suitability of a pointwise latent description of a probability distribution (Fisher, 1925). As Kullback and Leibler note, the functional properties of information in Fisher's construction are quite different from Shannon's and thus we do not regard Fisher information as a form of entropic information. Fisher's construction, however, can be rederived and understood within this theory. He begins with the assumption that there is some latent realization $\tilde{\boldsymbol{z}}$ for which $\mathbf{p}(\boldsymbol{x} \mid \tilde{\boldsymbol{z}})$ is an exact description of the true distribution of $\boldsymbol{x}$. We can then define the *Fisher score* as the gradient of information from any independent prior belief $\mathbf{q}_0(\boldsymbol{x})$ to a pointwise latent description $\mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})$, in the view $\mathbf{p}(\boldsymbol{x} \mid \check{\boldsymbol{z}})$

$$\boldsymbol{f} = \nabla_{\boldsymbol{z}} \mathbb{I}_{\mathbf{p}(\boldsymbol{x}|\check{\boldsymbol{z}})}[\, \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{x}) \,] \,.$$

Note that $\check{\boldsymbol{z}}$ is fixed by assumption, despite remaining unknown. By the variational principle, the score must vanish at the optimizer $\boldsymbol{z}^*$. By Corollary 10, the optimizer must be $\boldsymbol{z}^* = \check{\boldsymbol{z}}$. We can then assess the sensitivity of information to the parameter $\boldsymbol{z}$ at the optimizer $\boldsymbol{z}^*$ by computing the Hessian. This recovers an equivalent construction of the Fisher matrix within this theory

$$F_{ij} = \frac{\partial^2}{\partial \boldsymbol{z}_i \partial \boldsymbol{z}_j} \mathbb{I}_{\mathbf{p}(\boldsymbol{x}|\check{\boldsymbol{z}})}[\, \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{x}) \,] \,.$$

The primary idea behind this construction is that high-curvature in $\boldsymbol{z}$ implies that a pointwise description is both suitable and a well-conditioned optimization problem.

Generalized Fisher matrix

We may eliminate the assumption of an exact pointwise description and generalize analogous formulations to arbitrary views of expectation.

Definition 4 ***Generalized Fisher score.*** *Let* $\mathbf{r}(\boldsymbol{x})$ *be the view of expectation regarding an observable* $\boldsymbol{x}$. *The gradient with respect to* $\boldsymbol{z}$ *of information from independent prior belief* $\mathbf{q}_0(\boldsymbol{x})$ *to a pointwise description* $\mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})$ *gives the score*

$$\boldsymbol{f} = \nabla_{\boldsymbol{z}} \mathbb{I}_{\mathbf{r}(\boldsymbol{x})}[\, \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{x}) \,] \,.$$

Definition 5 ***Generalized Fisher matrix.*** *Let* $\mathbf{r}(\boldsymbol{x})$ *be the view of expectation regarding an observable* $\boldsymbol{x}$. *The Hessian matrix with respect to components of* $\boldsymbol{z}$ *of information from independent prior belief* $\mathbf{q}_0(\boldsymbol{x})$ *to the pointwise description* $\mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z})$ *gives the generalized Fisher matrix*

$$F_{ij} = \frac{\partial^2}{\partial \boldsymbol{z}_i \partial \boldsymbol{z}_j} \mathbb{I}_{\mathbf{r}(\boldsymbol{x})}[\, \mathbf{p}(\boldsymbol{x} \mid \boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{x}) \,] \,.$$

Again, a local optimizer $\boldsymbol{z}^*$ must satisfy the variational principle and yield a score of zero. The generalized Fisher matrix would typically be evaluated at such an optimizer $\boldsymbol{z}^*$.

5. Information in inference and machine learning

We now examine model information and predictive information provided by inference. Once we have defined these information measurements, we derive upper and lower bounds between them that we anticipate being useful for future work. Finally, we show how inference information may be constrained, which addresses some challenges in Bayesian inference.

5.1 Machine learning information

Akaike (Akaike, 1974) first introduced information-based complexity criteria as a strategy for model selection. These ideas were further developed by Schwarz, Burnham, and Gelman (Schwarz, 1978; Burnham and Anderson, 2001; Gelman et al., 2013). We anticipate these notions will prove useful in future work to both understand and control the problem of memorization in machine learning training. Accordingly, we discuss how this theory views model complexity and distinguishes formulations of predictive information and residual information.

In machine learning, observations correspond to matched pairs of inputs and labels $\boldsymbol{Y} = \left\{ \left(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)} \right) \middle| \, j \in [T] \right\}$. For each sample j of T training examples, we would like to map the input $\boldsymbol{x}^{(j)}$ to an output label $\boldsymbol{y}^{(j)}$. Latent variables $\boldsymbol{\theta}$ are unknown model parameters from a specified model family or computational structure. A *model* refers to a specific parameter state and the predictions that the model computes are $\mathbf{p}(\boldsymbol{y}^{(j)} \mid \boldsymbol{x}^{(j)}, \boldsymbol{\theta})$. Since the definitions and derivations that follow hold with respect to either single cases or the entire training ensemble, we will use shorthand notation $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})$ to refer to both scenarios.

We denote the initial state of belief in model parameters as $\mathbf{q}_0(\boldsymbol{\theta})$ and updated belief during training as $\mathbf{q}_i(\boldsymbol{\theta})$ for $i \in [n]$. We can then compute predictions from any state of model belief by marginalization $\mathbf{q}_i(\boldsymbol{y}) \equiv \int d\boldsymbol{\theta} \, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \mathbf{q}_i(\boldsymbol{\theta})$.

Definition 6 ***Model information.*** *Model information from initial belief $\mathbf{q}_0(\boldsymbol{\theta})$ to updated belief $\mathbf{q}_i(\boldsymbol{\theta})$ in the view of $\mathbf{r}(\boldsymbol{\theta})$ is given by*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{\theta})}[\, \mathbf{q}_i(\boldsymbol{\theta}) \, \| \, \mathbf{q}_0(\boldsymbol{\theta}) \,] \quad \textit{for} \quad i \in [n].$$

When we compute information contained in training labels, the label data obviously provide the rational view. This is represented succinctly by $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$, which assigns full probability to specified outcomes. Again, if we need to be explicit then this could be written as $\mathbf{r}(\boldsymbol{y} \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)})$ for each case in the training set.

Definition 7 ***Predictive label information.*** *The realization of training labels is the rational view $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$ of label plausibility. We compute information from prior predictive belief $\mathbf{q}_0(\boldsymbol{y})$ to predictive belief $\mathbf{q}_i(\boldsymbol{y})$ in this view as*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\, \mathbf{q}_i(\boldsymbol{y}) \, \| \, \mathbf{q}_0(\boldsymbol{y}) \,] \, .$$

In the continuous setting, this formulation is closely related to *log pointwise predictive density* (Gelman et al., 2013). We can also define complementary label information that is not contained in the predictive model.

Definition 8 ***Residual label information (discrete).*** *Residual information in the label realization view $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$ is computed as*

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\, \mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \, \| \, \mathbf{q}_i(\boldsymbol{y}) \,] \, .$$

Residual information is equivalent to cross-entropy if the labels are full realizations. We note, however, that if training labels are probabilistic and leave some uncertainty then replacing both occurances of $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$ above with a general distribution $\mathbf{r}(\boldsymbol{y})$ would correctly calibrate residual information so that if predictions were to match label distributions then residual information would be zero.

As a consequence of Corollary 2, the sum of predictive label information and residual label information is always constant. This allows us to rigorously frame predictive label information as a fraction of the total information contained in training labels. Moreover, Corollary 11 assures us that model perturbations that drive predictive belief toward the label view must increase predictive information. In the continuous setting, just as the limiting form of entropy discussed in 4.1 diverges, so also does residual information diverge. Predictive label information, however, remains a finite alternative. This satisfies our initial incentive for this investigation.

There is a second type of predictive information we may rationally construct, however. Rather than considering predictive information with respect to specified label outcomes, we might be interested in the information we expect to obtain about new samples from the generative process. If we regard marginalized predictions $\mathbf{q}_i(\boldsymbol{y})$ as our best approximation of this process, then we would simply measure change in predictive belief in this view.

Definition 9 ***Predictive generative approximation.*** *We may approximate the distribution of new outcomes from model belief $\mathbf{q}_i(\boldsymbol{\theta})$ using the predictive marginalization $\mathbf{q}_i(\boldsymbol{y}) \equiv \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{q}_i(\boldsymbol{\theta})$. If we hold this to be the rational view of new outcomes from the generative process, predictive information is*

$$\mathbb{I}_{\mathbf{q}_i(\boldsymbol{y})}[\, \mathbf{q}_i(\boldsymbol{y}) \,\|\, \mathbf{q}_0(\boldsymbol{y}) \,] \,.$$

5.2 Inference information bounds

In Bayesian inference, we have prior belief in model parameters $\mathbf{q}_0(\boldsymbol{\theta}) \equiv \mathbf{p}(\boldsymbol{\theta})$ and the posterior inferred from training data $\mathbf{q}_1(\boldsymbol{\theta}) \equiv \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})$. The predictive marginalizations are called the *prior predictive* and *posterior predictive* distributions respectively

$$\mathbf{p}(\boldsymbol{y}) \equiv \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{p}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \equiv \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}).$$

We derive inference information bounds for Bayesian networks (Pearl, 2014). Let $\boldsymbol{y}$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ represent a directed graph of latent variables. In general, the joint distribution can always be written as $\mathbf{p}(\boldsymbol{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{y})\mathbf{p}(\boldsymbol{\theta}_1 \mid \boldsymbol{y})\mathbf{p}(\boldsymbol{y})$. The property of *local conditionality* (Goodfellow et al., 2016) means $\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \check{\boldsymbol{y}}) \equiv \mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$. That is, belief dependence in $\boldsymbol{\theta}_2$ is totally determined by that of $\boldsymbol{\theta}_1$ just as belief in $\boldsymbol{\theta}_1$ is computed from $\check{\boldsymbol{y}}$.

Corollary 13 ***Joint local inference information.*** *Inference information in $\boldsymbol{\theta}_1$ gained by having observed $\check{\boldsymbol{y}}$ is equivalent to the inference information in both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.*

$$\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}[\, \mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \,] = \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})}[\, \mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1) \,] \,.$$

Corollary 14 ***Monotonically decreasing local inference information.*** *Inference information in $\boldsymbol{\theta}_2$ gained by having observed $\check{\boldsymbol{y}}$ is bound above by inference information in $\boldsymbol{\theta}_1$.*

$$\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}[\, \mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_2) \,] \leq \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})}[\, \mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1) \,] \,.$$

This shows that inference yields nonincreasing information as we compound inference on locally conditioned latent variables, which is relevant for sequential predictive computational models such as neural networks. We observe that the inference sequence from training data $\check{\boldsymbol{y}}$ to model parameters $\boldsymbol{\theta}$ to new predictions $\boldsymbol{y}$ is also a locally conditioned sequence. If belief in a given latent variable is represented as a probability distribution, this places bounds on what transformations are compatible with the progression of information. For example, accuracy measures which snap the maximum probability outcome of a neural network to unit probability impose an unjustified creation of information.

Corollary 15 ***Inferred information upper bound.*** *Model information in the posterior view is less than or equal to predictive label information resulting from inference*

$$\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})\,\|\,\mathbf{p}(\boldsymbol{\theta})\,] \leq \mathbb{I}_{\mathbf{r}(\boldsymbol{y}\mid\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y}\mid\check{\boldsymbol{y}})\,\|\,\mathbf{p}(\boldsymbol{y})\,]\,.$$

This is noteworthy because it tells us that inference always yields a favorable tradeoff between increased model complexity and predictive information. Combining Corollary 14 and Corollary 15, we have upper and lower bounds on model information due to inference

$$\mathbb{I}_{\mathbf{p}(\boldsymbol{y}\mid\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y}\mid\check{\boldsymbol{y}})\,\|\,\mathbf{p}(\boldsymbol{y})\,] \leq \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})\,\|\,\mathbf{p}(\boldsymbol{\theta})\,] \leq \mathbb{I}_{\mathbf{r}(\boldsymbol{y}\mid\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y}\mid\check{\boldsymbol{y}})\,\|\,\mathbf{p}(\boldsymbol{y})\,]\,.$$

5.3 Inference information constraints

Practitioners of Bayesian inference often struggle when faced with inference problems for models structures that are not well suited to the data. An under expressive model family is not capable of representing the process being modeled. As a consequence, the posterior collapses to a small set of outcomes that are least inconsistent with the evidence. In contrast, an over expressive model admits multiple sufficient explanations of the process.

Both model and predictive information measures offer means to understand and address these challenges. By constraining the information gained by inference, we may solve problems associated with model complexity. In this section, we discuss explicit and implicit approaches to enforcing such constraints.

Explicit information constraints

Our first approach to encode information constraints is to explicitly solve a distribution that satisfies expected information gained from the prior to the posterior. We examine how information-critical distributions can be constructed from arbitrary states of belief $\mathbf{q}_i(\boldsymbol{\theta})$ for $i \in [n]$. Again, we may obtain critical distributions with respect to uncertainty by simply setting $\mathbf{q}_0(\boldsymbol{\theta}) \equiv 1$. By applying this to inference, so that $n = 1$ and $\mathbf{q}_1(\boldsymbol{\theta}) \equiv \mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$, we recover likelihood annealling as a means to control model information.

Corollary 16 ***Constrained information.*** *Given states of belief $\mathbf{q}_i(\boldsymbol{\theta})$ and information constraints $\mathbb{I}_{\mathbf{r}(\boldsymbol{\theta})}[\,\mathbf{q}_i(\boldsymbol{\theta})\,\|\,\mathbf{q}_0(\boldsymbol{\theta})\,] = \varphi_i$ for $i \in [n]$, the distribution $\mathbf{r}^*(\boldsymbol{\theta})$ that satisfies these constraints while minimizing $\mathbb{I}_{\mathbf{r}(\boldsymbol{\theta})}[\,\mathbf{r}(\boldsymbol{\theta})\,\|\,\mathbf{q}_0(\boldsymbol{\theta})\,]$ has the form*

$$\mathbf{r}^*(\boldsymbol{\theta}) \propto \mathbf{q}_0(\boldsymbol{\theta}) \prod_{i=1}^{n} \left(\frac{\mathbf{q}_i(\boldsymbol{\theta})}{\mathbf{q}_0(\boldsymbol{\theta})}\right)^{\boldsymbol{\lambda}_i} \quad \textit{for some} \quad \boldsymbol{\lambda} \in \mathbb{R}^n.$$

Corollary 17 ***Information-annealed inference.*** *Annealed belief* $\mathbf{r}(\boldsymbol{\theta})$ *for which information gained from prior to posterior belief is fixed* $\mathbb{I}_{\mathbf{r}(\boldsymbol{\theta})}[\,\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta})\,] = \varphi$ *and information* $\mathbb{I}_{\mathbf{r}(\boldsymbol{\theta})}[\,\mathbf{r}(\boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{\theta})\,]$ *is minimal must take the form*

$$\mathbf{r}(\boldsymbol{\theta}) \propto \mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})^{\boldsymbol{\lambda}}\mathbf{p}(\boldsymbol{\theta}) \quad \textit{for some} \quad \boldsymbol{\lambda} \in \mathbb{R}.$$

Note that the bounds in 5.2 still apply if we simply include $\boldsymbol{\lambda}$ as a fixed model parameter in the definition of the likelihood function so that $\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}) \mapsto \mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \equiv \mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})^{\boldsymbol{\lambda}}$. This prevents the model from learning too much, which may be useful for under-expressive models or for smoothing out the posterior distribution to aid exploration during learning.

Implicit information constraints

Our second approach introduces hyper-parameters, $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$, into the Bayesian inference problem, which allows us to define a prior on those hyper-parameters that implicitly encodes information constraints. This approach gives us a way to express how much we believe we can learn from the data and model that we have in hand. Doing so may prevent overconfidence when there are known modeling inadequacies or underconfidence from overly broad priors.

As above, $\boldsymbol{\lambda}$ parameters influence the likelihood and can be though of as controlling annealing or an embedded stochastic error model. The $\boldsymbol{\psi}$ parameters control the prior on the model parameters $\boldsymbol{\theta}$. For example, these parameters could be the prior mean and covariance if we assume a Gaussian prior distribution. Therefore the inference problem takes the form

$$\mathbf{p}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\psi} \mid \check{\boldsymbol{y}}) = \frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}})}.$$

In order to encode the information constraints, we must construct the hyper-prior distribution $\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\psi}) = g\left(\varphi_\theta, \varphi_y\right)$ where

$$\begin{aligned}
\varphi_\theta &= \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}, \boldsymbol{\lambda}, \boldsymbol{\psi})}[\,\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}, \boldsymbol{\lambda}, \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\,] \quad \text{and} \\
\varphi_y &= \mathbb{I}_{\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \check{\boldsymbol{y}}, \boldsymbol{\lambda}, \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\psi})\,]
\end{aligned}$$

control model information and predictive information, respectively. The function $g\left(\varphi_\theta, \varphi_y\right)$ is the likelihood of $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$ given the specified model and prediction complexities. For example, this could be an indicator function as to whether the information gains are within some range. Note that we may also consider other forms of predictive information such as the predictive generative approximation.

The posterior distribution on model parameters and posterior predictive distribution can be formed by marginalizing over hyper-parameters

$$\begin{aligned}
\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) &= \int d\boldsymbol{\lambda} d\boldsymbol{\psi}\, \mathbf{p}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\psi} \mid \check{\boldsymbol{y}}) \quad \text{and} \\
\mathbf{p}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) &= \int d\boldsymbol{\lambda} d\boldsymbol{\psi}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})\mathbf{p}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\psi} \mid \check{\boldsymbol{y}}).
\end{aligned}$$

6. Negative information

The possibility of negative information is a unique property of this theory in contrast to divergence measures such as Kullback–Leibler divergence, α-divergences, and f-divergences. It provides an easily interpreted notion of whether a belief update is consistent with our best understanding. Negative information can be consistently associated with misinformation in the view of rational belief. That is, if $\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$ is negative then $\mathbf{q}_0(\boldsymbol{z})$ is a better approximation of rational belief than $\mathbf{q}_1(\boldsymbol{z})$. The consistency of this interpretation would be violated if we could construct $\mathbf{q}_1(\boldsymbol{z})$ that is unambigously better than $\mathbf{q}_0(\boldsymbol{z})$ at approximating $\mathbf{r}(\boldsymbol{z})$. Corollary 11 shows, however, that this is not possible. If we construct $\mathbf{q}_1(\boldsymbol{z})$ by integrating perturbations from $\mathbf{q}_0(\boldsymbol{z})$ that drive belief towards $\mathbf{r}(\boldsymbol{z})$ on all measurable subsets, then information must be positive. The following experiments illustrate examples of negative information and motivate its utility.

6.1 Negative information in continuous inference

In the following set of experiments we have a latent variable $\boldsymbol{\theta} \in \mathbb{R}^2$, which is distributed as $\mathcal{N}(\boldsymbol{\theta} \mid 0, \boldsymbol{I})$. Each sample $\boldsymbol{y}^{(j)} \in \mathbb{R}^2$ corresponds to realization of an independent latent variable $\boldsymbol{x}^{(j)} \in \mathbb{R}^2$ so that $\boldsymbol{y}^{(j)} = \boldsymbol{\theta} + \boldsymbol{x}^{(j)}$. Each $\boldsymbol{x}^{(j)}$ is distributed as $\mathcal{N}(\boldsymbol{x}^{(j)} \mid 0, \sigma_1^2 \boldsymbol{I})$ where $\sigma_1 = 1/2$. Both prior belief in plausible values of $\boldsymbol{\theta}$ and prior predictive belief in plausible values of $\boldsymbol{y}$ are visualized in Figure 1.4. Deciles separate annuli of probability 1/10. The model information we expect to gain by observing 10 samples of $\boldsymbol{y}$, which is also mutual information from Corollary 8, is $\mathbb{I}_{\mathbf{p}(\boldsymbol{y},\boldsymbol{\theta})}[\,\mathbf{p}(\boldsymbol{y},\boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y})\mathbf{p}(\boldsymbol{\theta})\,] = 5.36$ bits. See C for details.

The first observation consists of 10 samples of $\boldsymbol{y}$ followed by inference of $\boldsymbol{\theta}$. Subsequent observations each add another 10, 20, and 40 samples respectively. A typical inference sequence is shown in Figure 1.5. Model information gained by inference from the first observation in the same view is 5.72 bits. As additional observations become available the model information provided by first inference is eventually refined to 5.10 bits. Typically the region of plausible models $\boldsymbol{\theta}$ resulting from each inference is consistent with what was previously considered plausible.

By running 1 million independent experiments, we construct a histogram of the model information provided by first inference in subsequent views. This is shown in Figure 1.6. As a consequence of Postulate 4, the model information provided by first inference must always be positive before any additional observations are made. The change in model covariance in this experiment provides a stronger lower bound of 3.95 bits after first inference, which can be seen in the first view on the left. This bound is saturated in the limit when the inferred mean is unchanged. Additional observations may indicate that the first inference was less informative than initially believed. We may regard the rare cases showing negative information as being misinformed after first inference. The true value of the model $\boldsymbol{\theta}$ may be known to arbitrary precision if we collect enough observations. This is the realization limit on the right. Under this experimental design, this limit converges to the Laplace distribution centered at mutual information $\mathcal{L}(\mu, (\log 2)^{-1})$ computed in the prior view.

From these million experiments, we can select the most unusual cases for which the information provided by first inference is later found to an extreme. Figure 1.7 visualizes the experiment for which the information provided by first inference is found to be the minimum after observing 160 total samples from the generative process. Although model information assessed following the first observation is a fairly typical value, additional samples quickly show that the first samples were unusual. This becomes highly apparent in the fourth view, which includes 80 samples in total.

Figure 1.8 visualizes the complementary case in which we select the experiment for which the information provided by first inference is later found to be the maximum. The explantory characteristic of this experiment is the rare value that the true model has taken. High information in inference shows a high degree of surprise from what the prior distribution deemed plausible. Each inference indicates a range of plausible values of $\boldsymbol{\theta}$ that is quite distant from the plausible region indicated by prior belief. The change in belief due to first inference is confirmed by additional data in fourth inference.

Finally, we examine a scenario in which the first 10 samples are generated from a different process than subsequent samples. We proceed with inference as before and assume a single generative process, despite the fact that this assumption is actually false. Figure 1.9 shows the resulting inference sequence. After first inference, nothing appears unusual because there is no data that would contradict inferred belief. As soon as additional data become available, however, information in first inference becomes conspicuously negative. Note that the *one-in-a-million* genuine experiment exhibiting minimum information, Figure 1.7, gives -11.53 bits after 70 additional samples. In contrast, this experiment yields -47.91 bits after only 10 additional samples.

By comparing this result to the information distribution in the realization limit, we see that the probability of a genuine experiment exhibiting information this negative would be less than 2^{-155}. This shows how highly negative information may flag anomalous data. We explore this further in the next section.

6.2 Negative information in MNIST model with mislabeled data

We also explore predictive label information in machine learning models by constructing a small neural network to predict MNIST digits (LeCun et al., 1998). This model was trained with 50,000 images with genuine labels. Training was halted using cross-validation from 10,000 images that also had genuine labels. To investigate how predictive label information serves as an indicator of prediction accuracy, we randomly mislabeled a fraction of unseen cases. Prediction information was observed on 10,000 images for which 50% had been randomly relabeled, which resulted in 5,521 original labels and 4,479 mismatched labels.

The resulting distribution of information outcomes is plotted in Figure 1.10, which shows a dramatic difference between genuine labels and mislabeled cases. In all cases, prediction information is quantified from the uninformed probabilities $\mathbf{q}_0(\boldsymbol{y} = \boldsymbol{y}_i) = 1/10$ for all outcomes $i \in [10]$ to model predictions, which are conditioned on the image input $\mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x})$, in the view of the label $\mathbf{r}(\boldsymbol{y} \mid \boldsymbol{y}_i)$. Total label information, the sum of predictive label information and residual label information, is

$$\begin{aligned}\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\boldsymbol{y}_i)}[\, \mathbf{r}(\boldsymbol{y} \mid \boldsymbol{y}_i) \,\|\, \mathbf{q}_0(\boldsymbol{y}) \,] &= \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\boldsymbol{y}_i)}[\, \mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x}) \,\|\, \mathbf{q}_0(\boldsymbol{y}) \,] + \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\boldsymbol{y}_i)}[\, \mathbf{r}(\boldsymbol{y} \mid \boldsymbol{y}_i) \,\|\, \mathbf{q}_1(\boldsymbol{y} \mid \boldsymbol{x}) \,] \\ &= \log_2(10)\,\text{bits}\end{aligned}$$

or roughly 3.32 bits for each case. Both Figure 1.11 and Figure 1.12 show different forms of anomaly detection using negative information.

Figure 1.11 shows that genuine labels may exhibit negative information when predictions are poor. Only 1.1% of correct cases exhibit negative predictive label information. The distribution mean is 3.2 bits for this set. Notably, the first two images appear to be genuinely mislabeled in the original dataset, which underscores the ability of this technique to detect anomalies.

In contrast, over 99.1% of mislabeled cases exhibit negative information with the distribution mean at −18 bits. The top row of Figure 1.12 shows that information is most negative when the claimed label is not plausible and model predictions clearly match the image. Similarly to 6.1, strongly negative information indicates anomalous data. When incorrect predictions match incorrect labels, however, information can be positive as shown in the bottom row. The cases appear to share identifiable features with the claim.

7. Conclusion

Just as belief matures with accumulation of evidence, we hold that the information associated with a shift in belief must also mature. By formulating principles that articulate how we may regard information as a reasonable expectation that measures change in belief, we derived a theory of information that places existing measures of entropic information in a coherent unified framework. These measures include Shannon's original description of entropy, cross-entropy, realization information, Kullback–Leibler divergence, and Lindley information (uncertainty difference) due to an experiment.

Moreover, we found other explainable information measures that may be adapted to specific scenarios from first principles including the log pointwise predictive measure of model accuracy. We derived useful properties of information including the chain rule of conditional dependence, additivity over belief updates, consistency with respect expected future observations, and expected information in future experiments as mutual information. We also showed how this theory generalizes information-critical probability distributions that are consistent with observed expectations analogous to that of Jaynes'. In the context of Bayesian inference, we showed how information constraints recover and illuminate useful annealed inference practices.

We also examined the phenomenon of negative information, which occurs when a more justified point of view, based on a broader body of evidence, indicates that a previous change of belief was misleading. Experiments demonstrated that negative information reveals anomalous cases of inference or anomalous predictions in the context of machine learning.

The primary value of this theoretical framework is the consistent interpretation and corresponding properties of information that guide how it may be assessed in a given context. The property of additivity over belief updates within the present view allows us to partition information in a logically consistent manner. For machine learning algorithms, we see that total information from the uninformed state to a label-informed state is a constant that may be partitioned into the predicted component and the residual component. This insight suggests new approaches to model training, which will be the subject of continuing research.

7.1 Future work

The challenges we seek to address with this theory relate to real-world applications of inference and machine learning. Although Bayesian inference provides a rigorous foundation for learning, poor choices of prior or likelihood can lead to results that elude or contradict human intuition when analyzed after the fact. This only becomes worse as the scale of learning problems increases, as in deep neural networks, where human intuition cannot catch inconsistencies. Information provides a metric to quantify how well a model is learning that may be useful when structuring learning problems. Some related challenges include:

1. Controlling model complexity in machine learning to avoid memorization,
2. Evaluating the influence of different experiments and data points to identify outliers or poorly supported inferences,
3. Understanding the impact of both model-structure and fidelity of variational approximations on learnability.

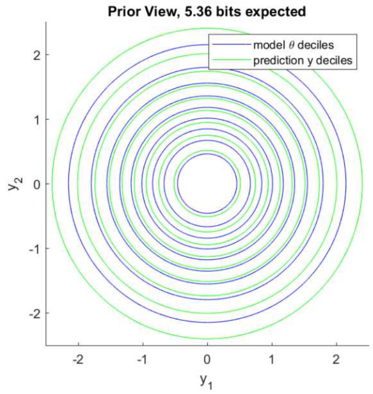

Figure 1.4: Prior distribution of $\boldsymbol{\theta}$ and prior predictive distribution of individual $\boldsymbol{y}$ samples. The domain of plausible $\boldsymbol{\theta}$ values is large before any observations are made.

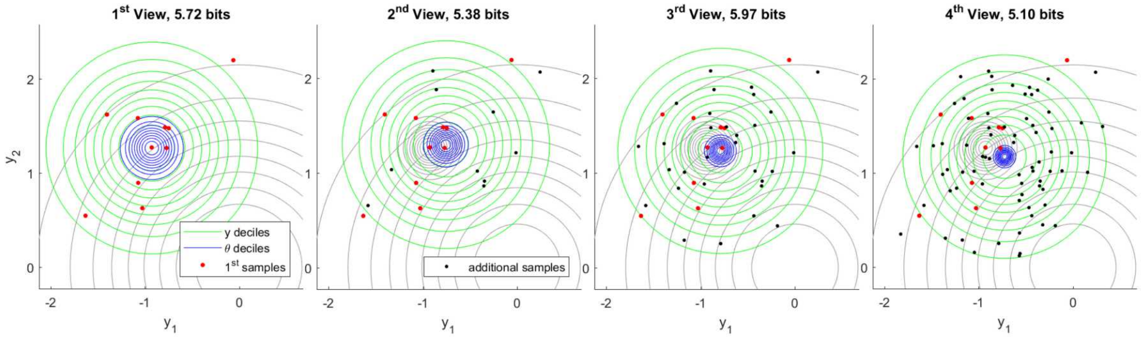

Figure 1.5: Typical inference of $\boldsymbol{\theta}$ from observation of 10 samples of $\boldsymbol{y}$ (left) followed by 10, 20, and 40 additional samples respectively. Both prior belief and first inference deciles of $\boldsymbol{\theta}$ are shown in gray. As observations accumulate, the domain of plausible $\boldsymbol{\theta}$ values tightens.

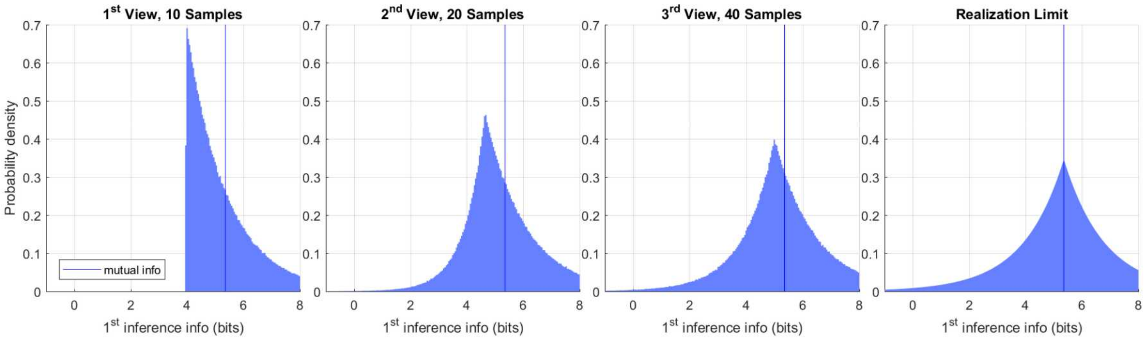

Figure 1.6: Histogram of first inference information in observation sequence. The vertical line at 5.36 bits is mutual information. Information is positive after first inference, but may drop with additional observations. The limiting view of infinite samples (realization) is shown on the right.

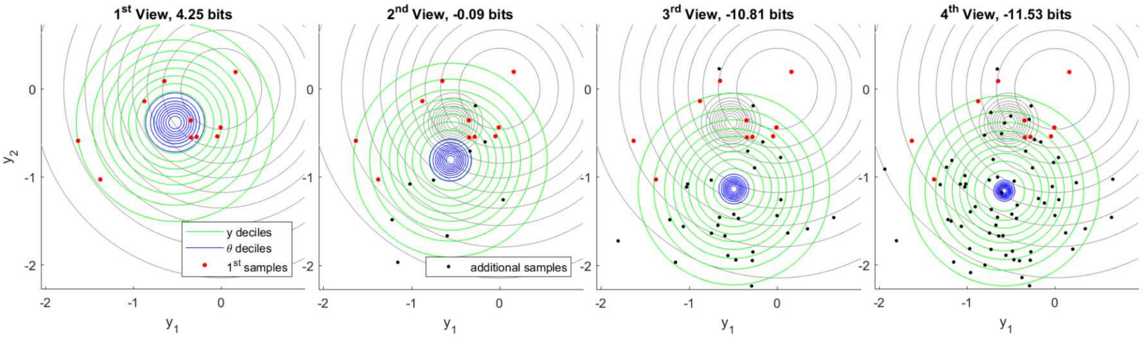

Figure 1.7: Minimum first inference information out of 1 million independent experiments. This particularly rare case shows how first samples can mislead inference, which is later corrected by additional observations. The fourth inference (right) bears remarkably little overlap with the first.

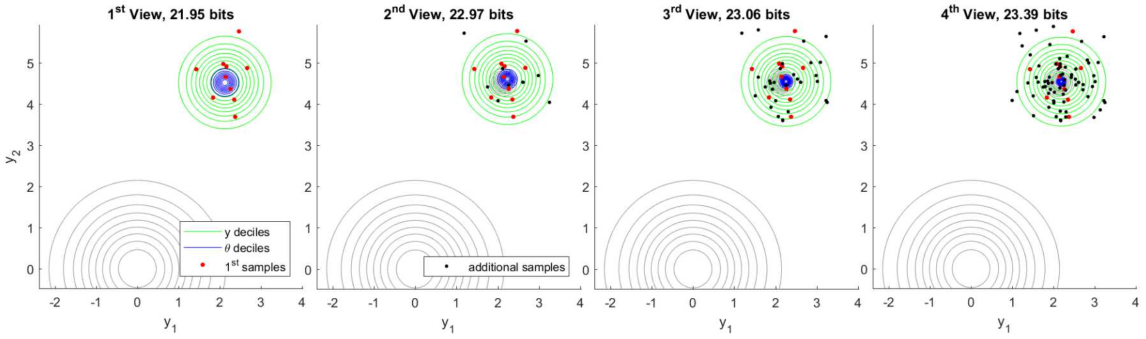

Figure 1.8: Maximum first inference information out of 1 million independent experiments. The true value of $\boldsymbol{\theta}$ has taken an extremely rare value. As evidence accumulates, plausible ranges of $\boldsymbol{\theta}$ confirm the first inference.

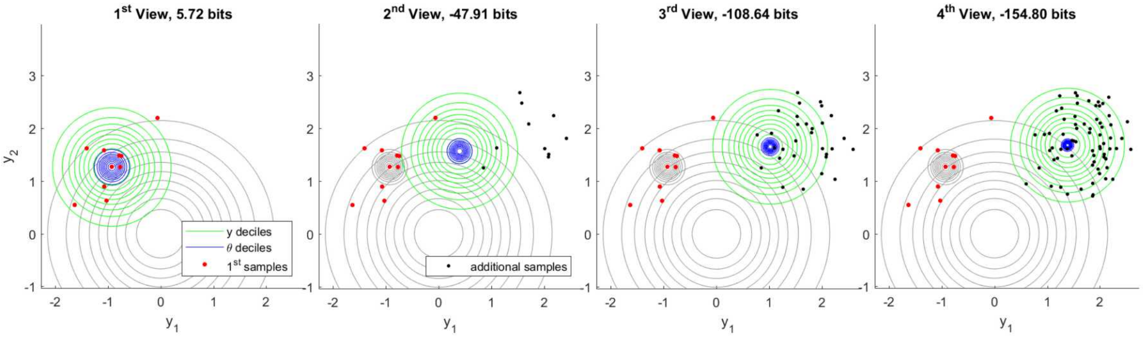

Figure 1.9: Inconsistent inference. The first 10 samples are drawn from a different ground truth than subsequent samples, but inference proceeds as usual. As additional data become available, first inference information becomes markedly negative.

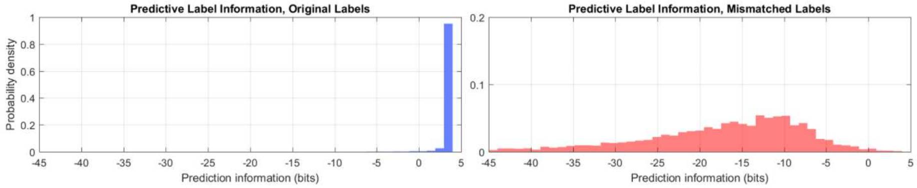

Figure 1.10: Histogram of information outcomes for mismatched and original labels. Correct label information is highly concentrated at 3.2 bits, which is 95.9% of the total information contained in labels. Mislabeled cases have mean information at -18 bits and information is negative for 99.15% of mislabeled cases.

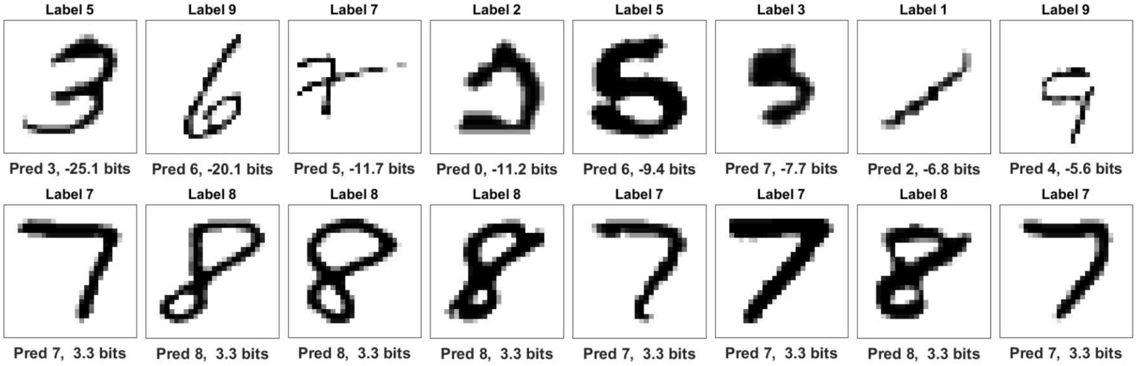

Figure 1.11: Original MNIST labels. The top row shows lowest predictive label information among original labels. Notably, the two leading images appear to be genuinely mislabeled in the original dataset. Subsequent predictions are poor. The bottom row shows the highest information among original labels. Labels and predictions are consistent in these cases.

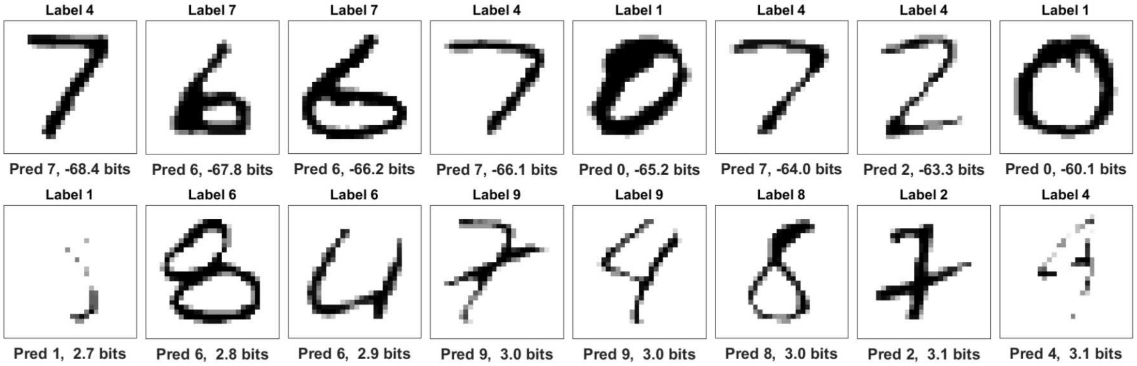

Figure 1.12: Mislabeled digits. The top row shows the lowest predictive label information among mislabeled cases. In each case, the claimed label is implausible and the prediction is correct. The bottom row shows the highest prediction information among mislabeled cases. Although claimed labels are incorrect, most images share identifiable features with the claim.

Chapter 2

Parsimonious Inference

Occam's Razor conveys an intuitive principle; the simplest sufficient explanation tends to be most predictive. Building upon our work regarding information as a rational measure of change in belief, we develop parsimonious inference, an information-theoretic formulation of Bayesian inference over arbitrary predictive architectures that formalizes Occam's Razor. Ideally, a learning problem consists of a description of prior belief followed by Bayesian inference to account for evidence in updated belief over models that could explain our data. Yet, it may be especially difficult to articulate well-founded justification for prior belief in the machine learning context, where algorithms contain high parameter dimensions and practitioners frequently modify computational architectures to improve outcomes. Our universal hyperprior expands on the core relationships between program length, Kolmogorov complexity, and Solomonoff's algorithmic probability to assign plausibility to prior descriptions that are encoded as sequences of discrete symbols. We then cast learning as an information minimization problem over our composite change in belief when an architecture is specified, training data are observed, and model parameters are inferred. This framework allows us to distinguish model complexity from prediction information and quantify the phenomenon of memorization.

Information optimization allows us to control learned explanatory complexity from first principles, rather than resorting to heuristic strategies such as cross-validation. Although this theory is general, it is most critical when we seek to apply machine learning to limited training datasets. Our optimization approaches require both efficient encodings of potential models and prudent sampling strategies to construct predictive ensembles from small datasets. Experiments include novel algorithms for polynomial regression and random forests in the presence of extremely small or skewed datasets. These algorithms demonstrate how model complexity adapts to the amount of evidence contained within available data in order to avoid memorization without cross-validation. Our theory addresses a fundamental challenge in how we may efficiently use data to obtain rational predictions from the otherwise arbitrary architectures encountered in machine learning.

1. Introduction

We began this investigation desiring to understand the relationship between prior belief and the resulting uncertainty in predictions obtained from inference in the hope that new insights would provide a sound basis to improve prediction credibility in machine learning. The mathematical and epistemological foundations of rational belief, from which the laws of probability and Bayesian inference are derived as an extended logic from binary propositional logic (Cox, 1946), lead us to assert the central role of Bayesian inference in obtaining rigorous justification for uncertainty in predictions. Although this foundation of reason holds generally, it is critical when we need to learn robust predictions from limited datasets. Yet, applying Bayesian inference within the machine learning context requires addressing a fundamental challenge: inference requires prior belief. When the amount of evidence contained within a dataset regarding a phenomenon of interest is extremely limited, specifying prior belief is not merely an inconvenience; it is the dominant source of uncertainty in predictions. Examples of such data limitations include having few observations, noisy measurements, skewed or highly imbalanced labels of interest, or even a degree of mislabeling in the data.

When predictive models integrate well-understood physical principles, they are often accompanied by physically plausible parameter ranges that provide a strong basis for prior belief. Likewise, canonical priors are acceptable for simple approximations with relatively few unconstrained parameters in comparison to the size of the dataset intended for inference. Kass and Wasserman (1996) give a thorough survery of related work. In contrast, the machine learning paradigm seeks to instrument arbitrary algorithms with high parameter dimensionality. A typical architecture may have tens of thousands, or perhaps millions, of free parameters. In this setting, the sensitivity of predictions to an arbitrary choice of prior belief may be unacceptable for applications of consequence (Owhadi et al., 2015).

1.1 Our contributions

Expanding on the work of Solomonoff (1964a,b, 2009), Kolmogorov (1965), Rissanen (1983, 1984), and Hutter (2007), we develop a theoretical framework that assigns plausibility to arbitrary inference architectures. Just as Solomonoff derives algorithmic probability from program length, widely understood as the number of bits needed to encode a program for a specific interpreter, we observe that a similar approach yields a universal hyperprior over symbolic encodings of ordinary priors. We may regard an ordinary prior, that which is typically used in Bayesian inference, as a restricted state of belief from a general universe of potential explanatory models. Within our framework, every choice of computational architecture, and associated prior over model parameters, is just a restriction of prior belief. Our hyperprior provides a means to measure and control the complexity of such choices.

We show how our theory of information (Duersch and Catanach, 2020), Theorem 1, allows us to derive a training objective from the information that is created when we select a prior representation, observe the training data, and either infer the posterior distribution or construct a variational approximation of it. Zhang et al. (2018) provide a thorough survey. Our main result, Theorem 2, expresses this objective in three components:

- Encoding information contained within a symbolic description of prior belief;
- Inference information gained regarding plausible models from observations;
- Predictive information gained regarding the observed labels from plausible models.

Notably, our derivation generates the first two terms with negative signs and the third with a positive sign, which reveals how this theory suppresses complexity as an intrinsic tradeoff against increased agreement with observed labels. We demonstrate this theory with two learning prototypes.

Our first algorithm casts polynomial regression within this framework, predicting a distribution over continuous outcomes from a continuous input. By setting the maximum polynomial degree to be much higher than the data merits, standard machine learning training strategies are susceptible to memorization, as we demonstrate by applying gradient based training with leave-one-out cross-validation. In contrast, our prototype discovers much simpler models from the same high-degree basis. When we aggregate predictions over an ensemble of polynomial representations, our prototype further demonstrates the natural increase in uncertainty we intuitively associate with extrapolation.

Our second algorithm samples ensembles of decision trees that are constructed using the parsimonious inference objective. These models aim to predict discrete labels through a sequence of partitions on continuous feature coordinates. Our random forest prototype demonstates the ability to learn credible prediction uncertainty from extremely small and heavily skewed datasets, which we contrast with a standard decision tree model and bootstrap aggregation. Both of these algorithms achieve superior prediction uncertainty by accounting for many alternative explanations, according to their degree of plausibility within the Bayesian paradigm of reason, from prior belief that is derived to both quantify and naturally suppress complexity over arbitrary explanations.

1.2 Organization

Section 2 begins with a discussion and illustration of the severe inadequacies of traditional machine learning training approaches that depend on cross-validation. We then briefly review the critical connections between scientific principles, rational belief, and Bayesian inference, which provide a sound theory to obtain rigorously justified uncertainty in predictions. When placed in the machine learning context, however, we explore how principled justification for prior belief over abstract models, including justification for our unavoidable disregard for an infinite number of alternatives, remains a critical challenge. Further, we summarize how our theory of information is derived to satisfy key properties that allow us to relate the various forms of complexity that follow in the parsimonious inference objective.

Section 3 continues with our main contributions, including a discussion of generalized description length, a coherent complexity hyperprior, and the principles of minimum information and maximum entropy. These notions culminate in the parsimonious inference objective, providing a suitable framework to understand and control model complexity over arbitrary learning architectures. We also show how this objective allows us to quantify memorization. Our theory allows us to apply these concepts within a wide variety of approaches to solving learning problems, including variational inference techniques.

Section 4 examines implementation details within our prototype algorithms, including efficient encodings and training strategies for polynomial regression and decision trees.

Section 5 concludes with a discussion of our theory's relationship to other work, a pathway to frame and address computability à priori, open questions for further investigation, and a summary of our findings.

2. Background

In order to place trust in machine learning predictions, we need to understand where trust originates in science. Yet, in order to motivate the need for a review, we begin by illustrating the severe deficiencies of standard machine learning practices. Machine learning models are typically trained using some variation of stochastic gradient descent (Robbins and Monro, 1951).

Gradient-based training algorithms are subject to overtraining, causing model predictions to memorize or artifically hew to training data even as predictions on unseen cases deteriorate. Cross-validation (Allen, 1974) attempts to prevent memorization, by monitoring predictions on a holdout dataset that is not directly used to tune model parameters. By monitoring predictions on a validation dataset, practitioners are able to tune hyperparameters, such as regularization weights and learning-rate schedules, and estimate the optimal model discovered over a training trajectory. These methods work well when data are abundant, but they are problematic in limited data regimes.

2.1 Memorization

The term memorization is often conflated with overtraining, however, we distinguish these terms as follows. Overtraining is characterized by degradation in prediction quality on unseen data that occurs after an initial stage of improvements. In contrast, memorization refers more generally to any predictive algorithm that exhibits unjustifiable confidence, or low prediction uncertainty, in the training dataset labels that were used to adjust model parameters.

Conflating these terms lends to an incorrect picture of the problem; to avoid memorization, we must merely halt training at the correct moment. Cross-validation is a data-driven method to address overtraining, but it fails to prevent memorization on small datasets. The obvious difficulty presented by cross-validation is the inherent tradeoff between using as much data as possible to train parameters, but also having a reliable estimator for when to halt training to prevent overfitting.

For limited data, standard practices apply some form of k-fold cross-validation (Hastie et al., 2009). We form k distinct partitions of the dataset, train k models respectively, and aggregate predictions by averaging. Leave-one-out cross-validation uses the same number of partitions as datapoints. Each partition reserves only one observation to estimate the best model over each training trajectory.

Figure 2.1 demonstrates this technique using polynomial regression, fitting 20th degree polynomials with only 12 points. The top-left shows an example of a single model trained by holding out one point for validation, shown in green. The average predictions over 12 such models appears at the bottom-left. Yet, suppose we could train with all 12 points while remaining highly confident that we will halt training at the correct moment. This ideal is demonstrated as a thought experiment in the middle column of Figure 2.1 by sampling 1000 extra data from the *generative process*, the ground truth mechanism that creates observations. We see that eliminating the tradeoff between training and validation would not prevent artifacts from developing that confidently hew to scant observations, memorization.

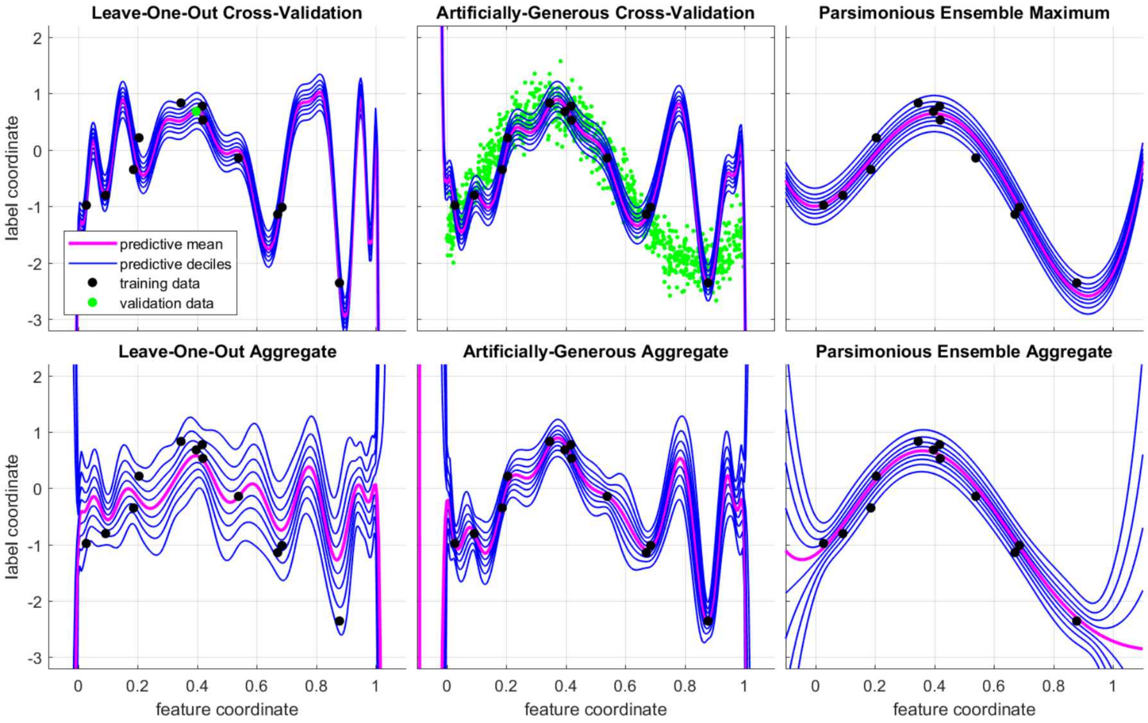

Figure 2.1: Illustration of standard training shortcomings. Top-left: training optimum obtained from holding out the green point. Bottom-left: mean predictions over all single-holdout sets. Leave-one-out does not prevent development of complex artifacts that hew to the data. Top-middle: idealized training using all original data as well as 1000 extra validation points from ground truth. Bottom-middle: mean predictions starting from 12 random initializations, $\mathcal{N}(\theta_i \mid \mu = 0, \sigma = 0.2)$. Removing the tradeoff between validation count and training data does not prevent complexities in predictions. Top-right: optimal model discovered using our theoretical framework and prototype algorithm. Bottom-right: our aggregate accounts for many plausible models, improving robustness and demonstrating natural extrapolation uncertainty.

This experiment demonstrates how neither of the competing cross-validation objectives are at the core of the problem with learning from limited data. Memorization is often framed in terms of a bias-variance tradeoff; predictions should avoid fluctuating rapidly, but also remain flexible enough to extract predictive patterns. In our theoretical framework, however, memorization is more comprehensively and rigorously understood as unparsimonious model complexity, i.e. increases in model information that are not justified by only small improvements to training predictions.

We conclude that stochastic gradient optimization never even explores low-complexity models. Regularization strategies attempt to address this heuristically by penalizing excessive freedom in learning parameters, for example attaching an ℓ_1 or ℓ_2 norm to the training objective. While many of these approaches can be equivalently cast as choices of prior belief, they lack a clear foundation that would allow us to compare alternatives or architectures without having a sufficient amount of data to apply cross-validation, thus failing to address the core challenge.

2.2 Scientific Reasoning and Bayesian Inference

In order to reiterate the concrete relationship between Bayesian inference and scientific reasoning, we review the epistemological foundations of reason at the center of the scientific method. These foundations bear decisive consequences regarding the valid forms of analysis we may pursue in order to obtain rational predictions. At its core, the scientific method relies on coherent mathematical models of observable phenomena that have been informed over centuries of physical measurements. Within the field of epistemology, this is the naturalist view of rational belief (Brandt, 1985). It holds that validity is ultimately derived from consistency, which can be understood in three key components:

1. Rational beliefs must be logical, avoiding internal contradictions;
2. Rational beliefs must be empirical, accounting for all available evidence;
3. Rational beliefs must be predictive, continually reassessing validity by how well predictions agree with new observations.

The third point is really nothing more than a restatement of the second point, placing emphasis on the evolving nature of rational beliefs as new data become available. The critical significance of the first point is that it provides a path to elevate the second and third points to a rigorous extended logic: Bayesian inference.

Building on the rich body of work by many scholars—including Ramsey (2016, original 1926), De Finetti (1937), and Jeffreys (1998, original 1939)—Cox (1946) shows that for a mathematical framework analyzing degrees of truth, belief as an extended logic, to be consistent with binary propositional logic, that formalism must satisfy the laws of probability:

1. Probability is nonnegative.
2. Only impossibility has probability zero.
3. Only certainty has maximum probability, which we normalize to one.
4. To revise the degree of credibility we assign to a model upon reviewing empirical evidence, we must apply Bayes' theorem.

Consequently, the Bayesian paradigm provides a uniquely rigorous approach to quantify uncertainty in the predictions we derive through inductive reasoning. Therefore, the only logically correct path to quantify and suppress memorization in learning must be cast within the Bayesian perspective.

Within the Bayesian formalism, a probability distribution called the prior $\mathbf{p}(\boldsymbol{\theta})$ quantifies our lack of information, our initial uncertainty, in plausible explanatory models. Here, $\boldsymbol{\theta}$ is any specific parameter state within a model class, or computational architecture. When we need to emphasize the prior dependence on a model class, or choice of architecture, and the distribution of parameters within that class, we will write the prior as $\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})$, where a description $\boldsymbol{\psi}$, or hyperparameters, provides such details. We will examine how $\boldsymbol{\psi}$ plays a key role regarding model complexity in detail in Section 3. The empirical data are expressed as a set of ordered pairs $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i \in [n]\}$ that have been sampled from the generative process $\mathbf{g}(\boldsymbol{x}, \boldsymbol{y})$. The features $\boldsymbol{x}_i$ are used to predict labels $\boldsymbol{y}_i$ from a computational structure and parameters. We write the predicted distribution over all potential labels as $\mathbf{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$.

If the ordered pairs in $\mathcal{D}$ represent independent samples from the underlying process, the likelihood is evaluated as $\mathbf{p}(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{i\in[n]} \mathbf{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$, which expresses the probability of observing $\mathcal{D}$ if a hypothetical explanation $\boldsymbol{\theta}$ held. Then, we update our beliefs according to Bayes' theorem. In our picture, we hold that having $\boldsymbol{\theta}$ alone is sufficient to evalute predictions. When we explicate the role of prior descriptions $\boldsymbol{\psi}$, that means inference can be written as

$$\mathbf{p}(\boldsymbol{\theta} \mid \mathcal{D}, \boldsymbol{\psi}) = \frac{\mathbf{p}(\mathcal{D} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})}{\mathbf{p}(\mathcal{D} \mid \boldsymbol{\psi})} \quad \text{where} \quad \mathbf{p}(\mathcal{D} \mid \boldsymbol{\psi}) = \int d\boldsymbol{\theta}\, \mathbf{p}(\mathcal{D} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})$$

is the model-class evidence. If we have a hyperprior $\mathbf{p}(\boldsymbol{\psi})$ over potential descriptions, we can also infer the hyperposterior

$$\mathbf{p}(\boldsymbol{\psi} \mid \mathcal{D}) = \frac{\mathbf{p}(\mathcal{D} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})}{\mathbf{p}(\mathcal{D})} \quad \text{where} \quad \mathbf{p}(\mathcal{D}) = \int d\boldsymbol{\psi}\, \mathbf{p}(\mathcal{D} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi}).$$

The central point of inference is that it does not attempt to identify a single explanation matching the data, as with stochastic gradient optimization and cross-validation. Rather, inference naturally adheres to the Epicurean principle, that we should retain multiple explanations according to their respective degrees of plausibility, within a coherent mathematical framework. We obtain rational predictions by evaluating the posterior predictive integral, or even the hyperposterior predictive integral, respectively constructed as

$$\begin{aligned}\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}, \boldsymbol{\psi}) &= \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \mathcal{D}, \boldsymbol{\psi}) \quad \text{and} \\ \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) &= \int d\boldsymbol{\psi}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}, \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi} \mid \mathcal{D}).\end{aligned}$$

The resulting predictions meet the exigent standard of rational belief for meaningful uncertainty quantification.

2.3 The Universal Scope of Prior Belief

The pervasive objection to the Bayesian paradigm is the lack of clear provenance for prior belief. In addition to objections based on subjectivity, translating our intuitive beliefs into distributions can be non-trivial. This problem is most pronounced in the domain of machine learning, where computational models are abstract, driven only by practical utility rather than well-understood physical principles. The premise of machine learning is that we do not need to integrate expert knowledge and specialized scientific theory into algorithms to obtain useful predictions from data, which eludes the traditional view of priors, that we must express our beliefs.

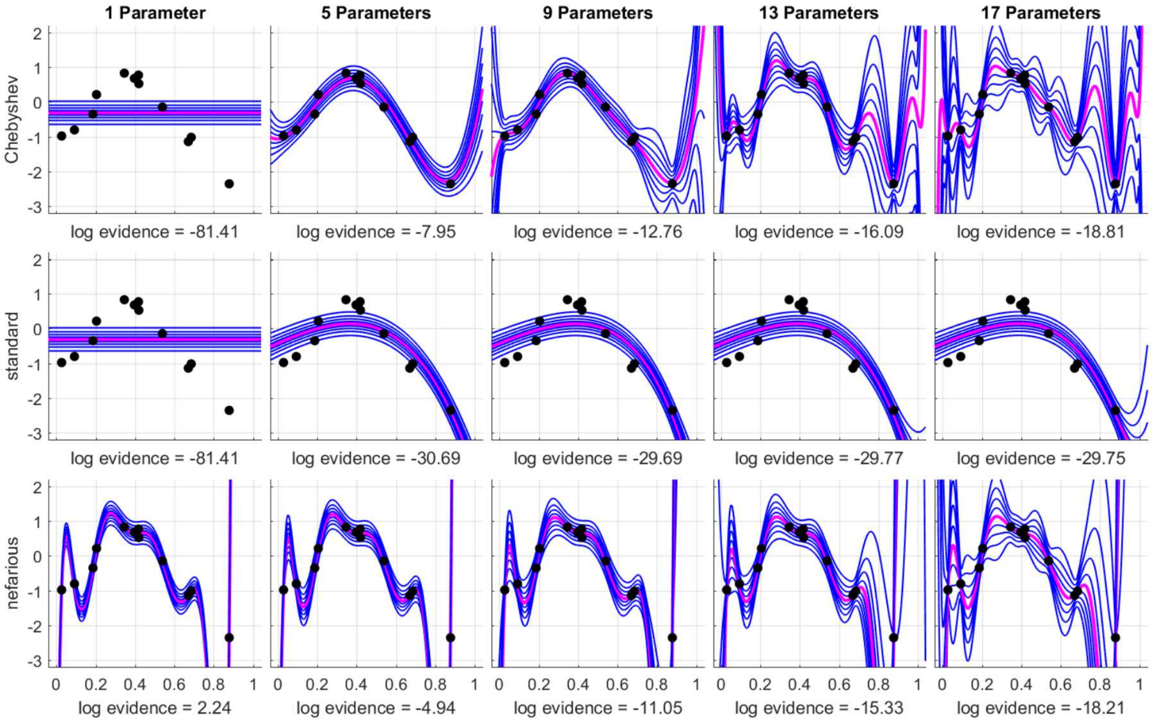

Figure 2.2: Illustration of prediction sensitivity to prior belief. The first row uses a Chebyshev polynomial basis and the second uses the standard basis. The third row exacerbates the problem of basis dependence by inserting a polynomial that has already partially memorized the data as the first basis function, followed by the Chebyshev basis. Priors are Gaussian with zero mean and identity covariance matrix. The choice of basis clearly matters and inference alone does not prevent the artifacts we associate with memorization from developing as parameter dimensions increase; we conclude that to control memorization, we require principled constraints on prior belief.

Prediction sensitivity to prior belief is most apparent when the number of parameters approaches or exceeds the size of our dataset, as illustrated in Figure 2.2. If we have n observations and $k > n$ differentiable parameters, every point in parameter space must have at least $k - n$ perturbable dimensions in which the likelihood gradient is zero. Because the likelihood remains constant as we move through these dimensions, each point lives within a $(k - n)$-dimensional submanifold wherein prior belief entirely determines the structure of posterior belief. Thus, within these submanifolds, the contribution of parameter uncertainty to prediction uncertainty is not affected by evidence. Clearly, we cannot be satisfied with meeting only the bare conditions for technically rational belief; we require concrete philosophical justification for prior belief.

In order to appreciate the solution, we must grasp the full severity of this problem by framing it in the most arduous scope. We can define the model universe as the set containing every computational architecture that could produce coherent predictions over $\boldsymbol{y}$ from $\boldsymbol{x}$. By considering inference over the model universe, we see that every architectural design decision is actually a choice of support for prior belief, i.e. the subdomain in which prior belief is nonzero. Similarly, parameter regularization strategies simply correspond to the shape of the prior distribution within an architecture. By capturing potential choices of prior belief using model-class descriptions $\boldsymbol{\psi}$, these problems are subsumed by investigating the correct form of a hyperprior $\mathbf{p}(\boldsymbol{\psi})$.

This picture also alludes to a second significant challenge in the machine learning setting, the problem of dimensionality in high-parameter families. Even if we obtain an attractive hyperprior, we can always construct increasingly complicated architectures. It is not possible to explore all of them. Occam's Razor provides a compelling path to a solution; explanations should not exhibit more complexity than what is required to explain the evidence. We conclude that a comprehensive hyperprior must compute and surpress a principled formulation of complexity. Moreover, our learning framework must justify disregarding infinite dimensions from inference and simultaneously address how to feasibly construct or approximate restricted posteriors.

2.4 Controlling Complexity

Bayesian theorists have had a persistent interest in articulating core principles for constructing priors over abstract models, particularly within the objectivist Bayesian philosophy. Examples of such priors include maximum entropy priors (Jaynes, 1957; Good, 1963) and Jeffrey's priors (Jeffreys, 1946). Other approaches use information criteria to determine a suitable number of parameters, such as the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). In contrast to these approaches that explicitly compare model classes, in which a parameter is either present or absent, Automatic Relevance Determination (ARD) priors (MacKay, 1995; Neal, 2012) take a softer approach. ARD uses a hyperprior to express uncertain relevance of different parameters and features in a model. It postulates that most model parameters should be close to zero because only a limited number of features are relevant for prediction. Through inference, these relevant model parameters can be identified automatically. ARD priors are are particularly relevant for machine learning because they were developed for neural networks. We will discuss the relationship between ARD priors and our theory in Section 5.3.

Perhaps the most principled approach to a universal prior is Solomonoff's work on algorithmic probability (Solomonoff, 1964a,b, 2009). Solomonoff derives a prior over all possible programs based upon their lengths using binary encodings subject to a Turing complete interpreter. Hutter (2007), reviewing prevalent principles of reason, goes further to show how Solomonoff's framework solves important philisophical problems in the Bayesian setting, including predictive and decision-theoretic performance bounds under the assumption that the generative process is a program. Potapov et al. (2012) also discuss Solomonoff's algorithmic probability, emphasizing the importance of retaining many alternative models to not only learn robust predictions, but also maintain adaptability in decision making.

Kolmogorov's work on mapping complexity (Kolmogorov, 1965) is closely related and we will examine it in detail in Section 3.1. Rissanen's work on universal priors and Minimum Description Length (MDL) (Rissanen, 1983, 1984) is also related. We examine his universal prior on integers in Section 4.1 and the relationship between our theory and MDL in Section 5.4. The key advantage of Solomonoff's approach is that it applies generally to any model we can program, thus eliminating artificial constraints on computational architectures. Solomonoff does not separate the model $\boldsymbol{\theta}$ from the model class $\boldsymbol{\psi}$, since any model from any model class can be expressed as a program. As this complexity-based information-theoretic prior provides a strong base for our work, we provide a detailed discussion in Section 3.2.

Fundamentally, the notion of complexity for arbitrary architectures concerns the amount of information that is contained within an encoding, first investigated by Shannon as entropy (Shannon, 1948). In order to rigorously understand how information in our datasets relates to Bayesian inference and encoding complexity, we developed a theory of information (Duersch and Catanach, 2020) rooted in understanding information as an expectation over rational belief. Given an arbitrary latent variable $\boldsymbol{z}$, we would like to measure the information gained by shifting belief between hypothetical states, i.e. from $\mathbf{q}_0(\boldsymbol{z})$ to $\mathbf{q}_1(\boldsymbol{z})$. We require this measurement to be taken relative to a third state of belief, $\mathbf{r}(\boldsymbol{z})$, which we hold to be valid. The precise reasoning by which validity of $\mathbf{r}(\boldsymbol{z})$ is derived is an important epistemological question. For our purposes, $\mathbf{r}(\boldsymbol{z})$ will either be rational belief, expressing our present understanding of the actual state of affairs, or a choice, representing a hypothetical state of affairs following a decision.

Rational belief is defined as the posterior distribution resulting from inference, which reserves some nonnegative probability for every outcome that is plausible. In contrast, we regard a choice as a restriction on the support of belief, effectively confining probability to any distribution that we can describe. This occurs when we must adhere to a course of action from a set of mutually incompatible options.

Key results are summarized in the following postulates and theorem.

1. Information gained by changing belief from $\mathbf{q}_0(\boldsymbol{z})$ to $\mathbf{q}_1(\boldsymbol{z})$ is quantified as an expectation over a third state $\mathbf{r}(\boldsymbol{z})$, called the view of expectation.

2. Information is additive over independent belief processes.

3. If belief does not change then no information is gained, regardless of the view of expectation.

4. Information gained from any normalized prior state of belief $\mathbf{q}_0(\boldsymbol{z})$ to an updated state of belief $\mathbf{r}(\boldsymbol{z})$ in the view of $\mathbf{r}(\boldsymbol{z})$ must be nonnegative

Information, measured in bits, satisfying these postulates must take the form

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log_2 \left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})} \right) \text{bits.}$$

When the view of expectation is the same as the target belief, we recover the Kullback-Leibler divergence (Kullback and Leibler, 1951)

$$D_{KL}[\,\mathbf{r}(\boldsymbol{z}) \,\|\, \mathbf{q}(\boldsymbol{z})\,] = \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{r}(\boldsymbol{z}) \,\|\, \mathbf{q}(\boldsymbol{z})\,]\,.$$

The entropy of a distribution $\mathbf{p}(\boldsymbol{z})$ over discrete outcomes $\boldsymbol{z} \in \{\boldsymbol{z}_i \mid i \in [n]\}$ is equivalent to the expected information gained upon realization in the view of the realization

$$S[\,\mathbf{p}(\boldsymbol{z})\,] = \sum_{i=1}^{n} \mathbf{p}(\boldsymbol{z}_i) \log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{z}_i)}\right) \text{ bits.}$$

Our theory allows us to relate and analyze changes in belief regarding our data, model parameters, and model-class descriptions within a unified framework. This theory provides a pathway to rigorously define memorization, Corollary 23, and prevent it by allowing us to quantify the benefit of changes to the model structure; an increase in model complexity that benefits multiple predictions is parsimonious, a worthwhile investment. In contrast, we regard memorization as an increase in model information benefiting few training instances. Further, when it becomes necessary to select discrete approximations of plausible model distributions in the pursuit of computational feasibility, the domain of variational inference, our theory allows us to analyze potential choices within the same formalism.

3. Complexity and Parsimony

We present our theoretical learning framework in three components. First, we analyze a modest generalization of Kolmogorov's notion of program length to sequences of symbols drawn from arbitrary alphabets that may be conditioned on previously realized symbols. This simplifies our ability to assign complexity to arbitrary descriptions of prior belief. Second, we use description length to derive a hyperprior that extends Solomonoff's formulation of algorithmic probability to general inference architectures. Third, we cast learning as an information minimization principle. We show how learning balances the two forms of information contained within models, due to both prior descriptions and inference, against the information the models provide about our dataset. Not only does this formalism allow us to analyze the utility of potential choices of restriction on prior belief, we also recover variational inference optimization from the same principle when we desire to approximate the posterior distribution.

3.1 Program Length and Kolmogorov Complexity

Kolmogorov's discussion of complexity begins with a countable set of objects $\mathcal{A} = \{\boldsymbol{a}\}$ that are indexed with binary sequences. For Kolmogorov, an object $\boldsymbol{a}$ is a program and the length of the program $\ell(\boldsymbol{a})$ is taken to be the number of binary digits in a corresponding binary sequence $\boldsymbol{\psi}(\boldsymbol{a})$. Given a domain of program inputs $\mathcal{X}$ and a codomain of outputs $\mathcal{Y}$, a Turing complete interpreter $\boldsymbol{\varphi}(\cdot,\cdot)$ accepts the program $\boldsymbol{a}$ and an input $\boldsymbol{x} \in \mathcal{X}$ and returns an output $\boldsymbol{y} = \boldsymbol{\varphi}(\boldsymbol{a}, \boldsymbol{x}) \in \mathcal{Y}$. The Kolmogorov complexity of an ordered pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ is the length of the shortest program that is capable of reproducing the pair

$$\mathcal{K}_{\boldsymbol{\varphi}}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{a} \in \mathcal{B}} \ell(\boldsymbol{a}) \quad \text{where} \quad \mathcal{B} = \{\boldsymbol{a} \mid \boldsymbol{y} = \boldsymbol{\varphi}(\boldsymbol{a}, \boldsymbol{x})\} \subset \mathcal{A}.$$

An ordered pair may be understood to enforce multiple function values or even the entire mapping that defines a function. Further, Kolmogorov's framework easily captures the complexity of a singleton $\boldsymbol{y}$ by taking an empty input $\boldsymbol{x} = \emptyset$.

Because Turing complete interpreters can simulate one another, the choice of interpreter only affects complexity by a small constant offset. That said, the choice of interpreter is an important problem that is also resolvable within our theory. Consequently, we revisit this issue in Section 5.1.

The descriptions of interest to us, however, may not admit perfectly efficient binary codes. For example, we may wish to represent the outcome of rolling of a balanced six-sided die. Rather than solving for an optimal binary encoding (Huffman, 1952), it is convenient to extend the notion of the length to finite sequences of symbols drawn from multiple alphabets. For example, this extension supports efficient descriptions of feature domain partitions within decision trees, discussed in Section 4.3.

Let a description be composed of sequence of symbols represented as $\boldsymbol{\psi} = (\boldsymbol{s}_i)_{i=1}^n$. When we wish to draw attention to the role of a sequence as an encoding of an object, we write $\boldsymbol{\psi}(\boldsymbol{a})$. For our purposes, these objects do not necessarily need to be programs, but rather are simply descriptions of belief, such as $\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})$ discussed earlier. For each $i \in [n]$, a symbol $\boldsymbol{s}_i$ is selected from an alphabet $\boldsymbol{\Sigma}_i$. We emphasize that the each alphabet is allowed to depend on previously realized symbols in the sequence so that the sequence of alphabets is not fixed. To avoid cumbersome notation and excessive indexing variables, subsequences expressed as $(\boldsymbol{s})_1^j$ are interpreted with the natural indexing $(\boldsymbol{s}_i)_{i=1}^j$. It is also useful to regard $(\boldsymbol{s})_1^0$ as the empty subsequence. Likewise, the conditional dependence of each alphabet on previous symbols is left implied so that we may simply write $\boldsymbol{\Sigma}_i$ rather than $\boldsymbol{\Sigma}_i\left[(\boldsymbol{s})_1^{i-1}\right]$. As with Kolmogorov, the length of an object is derived from a sequence, $\ell(\boldsymbol{a}) = \ell(\boldsymbol{\psi}(\boldsymbol{a}))$.

If we treat each symbol in an encoding $\boldsymbol{\psi}(\boldsymbol{a})$ as a random variable, we have

$$\mathbf{p}(\boldsymbol{\psi}(\boldsymbol{a})) = \prod_{i=1}^{n} \mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1}).$$

The entropy corresponding to each potential symbol, or the information we expect to gain upon realization, is the maximum if and only if the probability of each symbol is uniform over its alphabet, $\mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1}) = \frac{1}{|\boldsymbol{\Sigma}_i|}$. That is,

$$\sum_{\boldsymbol{s}_i \in \boldsymbol{\Sigma}_i} \mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1}) \log_2\!\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1})}\right) \leq \log_2(|\boldsymbol{\Sigma}_i|).$$

Our construction of generalized length in Definition 10 invokes the principle of maximum entropy to remove restrictions on the kinds of codes we can consider, while recovering Kolmogorov's length when all symbols are binary digits.

Definition 10 ***Generalized Length as Maximum Entropy Encoding.*** *The generalized length of an arbitrary sequence $\boldsymbol{\psi}$ is the upper bound on entropy of the corresponding sequence of alphabets from which each symbol is drawn,*

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log_2(|\boldsymbol{\Sigma}_i|) \textit{ bits.}$$

Corollary 18 shows that this length saturates the information lower bound in Shannon's source coding theorem (Shannon, 1948). We provide all proofs in Appendix D. Furthermore, when we develop an efficient encoding for the kinds of objects we would like to use, Corollary 19 allows us to naturally derive the probability of an object from the encoding.

Corollary 18 ***Generalized Length Lower Bound.*** *Given a set of objects $\mathcal{A}$ and probabilities $\mathbf{p}(\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathcal{A}$, the expected generalized length of an object is bound from below by the entropy*

$$\mathbb{E}_{\mathbf{p}(\boldsymbol{a})}\, \ell(\boldsymbol{\psi}(\boldsymbol{a})) \geq \mathbb{E}_{\mathbf{p}(\boldsymbol{a})} \log_2 \left(\frac{1}{\mathbf{p}(\boldsymbol{a})} \right).$$

Corollary 19 ***Probability from Length.*** *Given a set of objects $\mathcal{A}$ and a maximum entropy encoding with $\mathbf{p}(\boldsymbol{a}) = \mathbf{p}(\boldsymbol{\psi}(\boldsymbol{a}))$ for all $\boldsymbol{a} \in \mathcal{A}$, the generalized length satisfies*

$$\mathbf{p}(\boldsymbol{a}) = 2^{-\ell(\boldsymbol{a})}.$$

We remark that by allowing symbol probabilities to deviate from the maximum entropy limit, subjective prior beliefs could take the form of non-uniform probabilities within each alphabet. In this view, just as generalized length is the limit of expected information gained by realization from an encoding, the corresponding probability in Corollary 19 may be regarded as a limit of subjective priors subject to a given encoding. Subjective prior beliefs may also be taken into account through the choice of interpreter that we deem to be valid.

3.2 Algorithmic Probability

During the same time period that Kolmogorov worked on mapping complexity, Solomonoff developed a related theoretical framework for inductive inference and algorithmic probability. He was specifically interested in programs capable of reproducing a binary sequence $\boldsymbol{y}$, i.e. the subset of programs $\mathcal{B} = \{\boldsymbol{a} \mid \boldsymbol{y} = \boldsymbol{\varphi}(\boldsymbol{a}, \emptyset)\} \subset \mathcal{A}$, and he derived the probabilistic contribution of each program to plausible continuations of the sequence

$$\mathbf{p}(\boldsymbol{a} \mid \boldsymbol{y}) \propto \begin{cases} 2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))} & \boldsymbol{a} \in \mathcal{B} \\ 0 & \boldsymbol{a} \notin \mathcal{B} \end{cases}$$

where, as with Kolmogorov's picture, length corresponds to a binary encoding subject to a Turing complete interpreter. We view his result as Bayesian inference wherein Corollary 19 provides prior belief and a program has unit likelihood if it reproduces the sequence and zero likelihood otherwise. It follows that the optimizing program that yields Kolmogorov's complexity is simply the MAP estimator in the same picture.

In the information-theoretic perspective, the Kolmogorov complexity is the minimum amount of information that is possible to gain by restricting belief to a discrete program that is capable of reproducing a desired ordered pair. If, however, we allow distributions of belief over many programs, so that $\mathbf{r}(\boldsymbol{a}) \geq 0$ for any $\boldsymbol{a} \in \mathcal{B}$, Corollary 20 shows that Solomonoff's algorithmic probability is the minimizer of information gain, improving beyond the Kolmogorov complexity.

Corollary 20 ***Information Optimality of Solomonoff Programs.*** *If we measure the change in belief from all possible programs according to Corollary 19 to distributions $\mathbf{r}(\boldsymbol{a})$ that restrict belief to programs capable of reproducing an input-output pair $(\boldsymbol{x}, \boldsymbol{y})$, the minimizer*

$$\mathbf{r}^{*}(\boldsymbol{a}) = \underset{\mathbf{r}(\boldsymbol{a})}{\mathit{argmin}}\; D_{KL}[\,\mathbf{r}(\boldsymbol{a}) \,\|\, \mathbf{p}(\boldsymbol{a})\,] \quad \textit{subject to} \quad \mathbf{r}(\boldsymbol{a}) = 0 \quad \forall \quad \boldsymbol{a} \notin \mathcal{B} = \{\boldsymbol{a} \mid \boldsymbol{y} = \boldsymbol{\varphi}(\boldsymbol{a}, \boldsymbol{x})\}\,,$$

is uniquely given by

$$\mathbf{r}^{*}(\boldsymbol{a}) = \frac{2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}}{\mathbf{p}(\mathcal{B})} \quad \forall \quad \boldsymbol{a} \in \mathcal{B} \quad \textit{where} \quad \mathbf{p}(\mathcal{B}) = \sum_{\boldsymbol{a}\in\mathcal{B}} 2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}.$$

Solomonoff's picture is even more general than it first appears. There is no need to confine our attention to binary programs that reproduce sequences. Instead, we may apply the same framework to programs that generate coherent probabilities on any given dataset, $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{a})$. Since any algorithm we write is ultimately a program, this induces universal prior belief over arbitrary algorithms. Then, Bayesian inference yields rational belief as a posterior distribution over all such algorithms.

Yet, this approach is fundamentally difficult because it requires us to efficiently explore the posterior over suitable programs via their discrete sequences. Since it is not always possible to anticipate how a program will respond to given inputs in finite time, we arrive at the problem of uncomputability. Moreover, even if we discover seemingly high posterior probability programs, we cannot guarantee that our sample adequately represents the posterior for the purpose of obtaining credible uncertainty in predictions. We discuss this problem and a potential solution further in Section 5.2.

Rather than restricting our attention to programs, however, we would like analyze more general inference architectures, which we easily accomplish by relaxing $\boldsymbol{\psi}$ to be merely a description of prior belief $\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})$. A description may still be a program, a series of instructions, subject to a Turing complete interpreter, wherein we implicitly assign full probability to the program that was written, but we can also consider interpreters that merely require a few hyperparameters of a distribution to be specified, rather than a full set of instructions. Then, standard inference drives the result from a single description, rather than being forced to only account for posterior probability among separate programs. Letting our data drive updates in belief from a short description is often substantially more information-efficient than accounting for a full program that computes the equivalent result. In doing so, however, we lose the ability of interpreters to simulate one another, the critical property of Turing complete decoders that bounds variations in Kolmogorov complexity resulting from the choice of interpreter. Yet, our framework still allows us to compute the relative plausibility of interpreters within the same theory by identifying a common language describing interpreters. We discuss the role of interpreters further in Section 5.1.

The parsimonious hyperprior over corresponding model classes from Corollary 19, $\mathbf{p}(\boldsymbol{\psi}) = 2^{-\ell(\boldsymbol{\psi})}$, is enough to complete the Bayesian framework with well-founded justification for how we arrive at prior belief over arbitrary architectures. When we perform Bayesian inference from a parsimonious prior or hyperprior, we call the result parsimonious rational belief. To achieve computational feasiblity, however, we still need to investigate principled restrictions of belief.

3.3 The Principle of Information Minimization

We develop this paradigm with the intention of providing a path to computationally feasibility predictions with well-founded theoretical justification. As alluded to in Corollary 20, we can cast learning as an information minimization problem over our total change in belief due to observing the training data $\mathcal{D}$, selecting model classes $\boldsymbol{\psi}$, and solving for distributions over models $\boldsymbol{\theta}$ within each class. The principle of minimum information (Evans, 1969), based on the closely related principle of maximum entropy (Jaynes, 1957), intuitively states that driving the information gained upon viewing the training data to be as low as possible, we obtain better predictions. If a dataset contains a highly predictive pattern, then once that pattern is known we can obtain strong predictions. As a consequence, the information gained by observing new labels drops because the outcomes are easy to predict. In contrast, the information gained by observing new labels will remain high when there is no discernable pattern. We derive Theorem 2 from a rigorous formulation of the minimum information principle using our work regarding information as a rational measure of change in belief and show how this information objective can be manipulated into three terms that provide insight into how we may understand and control complexity during learning as a constrained optimization problem with the aim of improving computational feasibility.

Theorem 2 ***Parsimonious Inference Optimization.*** *Let our training dataset be represented as an ordered pair* $(\boldsymbol{x}, \boldsymbol{y})$*. A model* $\boldsymbol{\theta}$ *computes coherent probabilities over potential labels* $\boldsymbol{y}$ *from features* $\boldsymbol{x}$*, or* $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$*. The shorthand* $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})$ *leaves dependence on* $\boldsymbol{x}$ *implied. A description of the model class is represented by a sequence* $\boldsymbol{\psi}$ *and provides a restriction on prior belief* $\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})$*. The parsimonious hyperprior over potential sequences derives from generalized length as* $\mathbf{p}(\boldsymbol{\psi}) = 2^{-\ell(\boldsymbol{\psi})}$*. Our joint belief in potential sequences, models, and labels is given by* $\mathbf{p}(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})$*. Viewing the labels* $\check{\boldsymbol{y}}$ *changes our belief in labels to* $\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})$*, a distribution assigning full probability to the observed outcomes. Let our choice of belief over descriptions after viewing the data be represented as* $\mathbf{r}(\boldsymbol{\psi})$*. Likewise, within the model class* $\boldsymbol{\psi}$*, our chosen posterior approximation over plausible models is written as* $\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})$*, which may or may not be the exact posterior* $\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}, \boldsymbol{\psi})$*. The total information gained is given by the Kullback-Leibler divergence*

$$D_{KL}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})\,]\,.$$

The minimizer of this objective is equivalent to the maximizer of the following parsimony objective, expressed in three parts

$$\begin{aligned}\omega = {} & \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi})}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] && \textit{(prediction information)}\\ & - \mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})}\, D_{KL}[\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\,] && -\textit{(inference information)}\\ & - \mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})}\, \ell(\boldsymbol{\psi}) + S[\,\mathbf{r}(\boldsymbol{\psi})\,]\,, && -\textit{(description information)}\end{aligned}$$

where $\boldsymbol{\theta}_0$ *anchors predictive information gained regarding labels relative to any fixed baseline model.*

The first term, prediction information, is the expected information gained about training data resulting from our belief in explanations. Both secondary terms, inference information and description information, account for model complexity. Anchoring predictive information to any fixed model $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)$ allows us to coherently interpret label information as that which is gained relative to $\boldsymbol{\theta}_0$. Any fixed predictive distribution suffices, including the prior predictive

$$\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \int d\boldsymbol{\psi}\, d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi}).$$

However, because the prior predictive may be difficult (or impossible) to compute, it is much simpler to use a naïve model $\boldsymbol{\theta}_0$. If we disregard the role of $\boldsymbol{\psi}$ and only account for prediction information and inference information from prior belief $\mathbf{p}(\boldsymbol{\theta})$, i.e. from the first two terms of the parsimonious inference objective, then we recover a form of the Bayesian Occam's Razor (MacKay, 1992),

$$\begin{aligned}\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y})\,\|\,\mathbf{p}(\boldsymbol{y}\mid\boldsymbol{\theta}_0)\,] &= \log_2\!\left(\frac{\mathbf{p}(\check{\boldsymbol{y}})}{\mathbf{p}(\check{\boldsymbol{y}}\mid\boldsymbol{\theta}_0)}\right)\\ &= \mathbb{E}_{\mathbf{p}(\boldsymbol{\theta}|\check{\boldsymbol{y}})}\log_2\!\left(\frac{\mathbf{p}(\check{\boldsymbol{y}}\mid\boldsymbol{\theta})}{\mathbf{p}(\check{\boldsymbol{y}}\mid\boldsymbol{\theta}_0)}\right) - D_{KL}[\,\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})\,\|\,\mathbf{p}(\boldsymbol{\theta})\,]\,,\end{aligned}$$

provided we use the exact posterior $\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})$ for $\mathbf{r}(\boldsymbol{\theta}\mid\boldsymbol{\psi})$. Otherwise, we recover a variational inference objective that is equivalent to maximizing the conventional Evidence Lower Bound (ELBO). These constructions show a tradeoff between information due to inference and information inference provides about our data, but they leave the provenance of $\mathbf{p}(\boldsymbol{\theta})$ unaddressed.

Our description information terms show how our theory subsumes the Principle of Maximum Entropy. If we were to disregard the critical role that the parsimonious hyperprior plays in controlling complexity, dropping expected length, we would be left with an optimization objective that drives increases in entropy within our chosen distribution of descriptions. Doing so, however, would mean that long and complicated descriptions would be just as plausible as short and simple descriptions, as long as they are all equally capable of explaining the data. By accounting for the complexity of descriptions, the parsimonious inference objective seeks to drive up information gained about our dataset while simultaneously suppressing a complete notion of model complexity. Thus, we have a complete and rigorous formulation of the intuition in Occam's Razor.

Critically, Corollaries 21 and 22 show that unconstrained optimization recovers, and is therefore consistent with, Bayesian inference and parsimonious rational belief. As demonstrated in Section 4.3, some prior beliefs facilitate exact inference and easily allow us to take $\mathbf{r}(\boldsymbol{\theta}\mid\boldsymbol{\psi}) = \mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}},\boldsymbol{\psi})$. Yet, unconstrained optimization of $\mathbf{r}(\boldsymbol{\psi})$ would need to account for unlimited varieties of programs and model classes. It is not clear how that would ever be achievable. Instead, we can restrict the support of prior belief to a feasible exploration manifold. We can also constrain the types of distributions we are willing to consider to approximate the posterior. Theorem 2 provides a consistent framework to compare the utility of such restrictions.

Corollary 21 ***Optimality of Inference.*** *Given a single description $\boldsymbol{\psi}$ specifying prior belief $\mathbf{p}(\boldsymbol{\theta}\mid\boldsymbol{\psi})$, the conditionally optimal distribution over models,*

$$\mathbf{r}^{*}(\boldsymbol{\theta}\mid\boldsymbol{\psi}) = \underset{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})}{\mathit{argmax}}\quad \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y}\mid\boldsymbol{\theta})\,\|\,\mathbf{p}(\boldsymbol{y}\mid\boldsymbol{\theta}_0)\,] - D_{KL}[\,\mathbf{r}(\boldsymbol{\theta}\mid\boldsymbol{\psi})\,\|\,\mathbf{p}(\boldsymbol{\theta}\mid\boldsymbol{\psi})\,]\,,$$

is the posterior distribution $\mathbf{r}^{*}(\boldsymbol{\theta}\mid\boldsymbol{\psi}) = \mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}},\boldsymbol{\psi})$.

Corollary 22 ***Optimality of Hyper Inference.*** *Applying the optimizer from Corollary 21 to the objective in Theorem 2 produces the second optimization problem*

$$\mathbf{r}^{*}(\boldsymbol{\psi}) = \underset{\mathbf{r}(\boldsymbol{\psi})}{\mathit{argmax}}\quad \mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})}\log_2\!\left(\frac{\mathbf{p}(\check{\boldsymbol{y}}\mid\boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}}\mid\boldsymbol{\theta}_0)}\right) - D_{KL}[\,\mathbf{r}(\boldsymbol{\psi})\,\|\,\mathbf{p}(\boldsymbol{\psi})\,]\,.$$

The optimizer is the hyperposterior distribution, $\mathbf{r}^*(\boldsymbol{\psi}) = \mathbf{p}(\boldsymbol{\psi} \mid \tilde{\boldsymbol{y}})$.

As an information objective, we hold the given views of expectation to be valid because they represent the actual distributions that will be used in practice to compute predictions. We also observe that this construction of information, rather than the reversed divergence $D_{KL}[\,\mathbf{p}(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\psi}) \,\|\, \mathbf{r}(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\psi})\,]$, is necessary to avoid multiple infinities. For each description $\boldsymbol{\psi}$ with $\mathbf{r}(\boldsymbol{\psi}) = 0$, the reversed divergence is infinite. Moreover, we cannot avoid eliminating an infinite number of such cases from consideration.

The parsimony objective allows us to understand and quantify memorization of training data in Corollary 23 as a bound on the increase in model complexity that is required to achieve increased agreement between predictions and our training data. This bound holds even when we restrict the support of prior belief to a computationally feasible manifold. Likewise, we may also restrict our attention to allowable variational approximations of the posterior distribution and the bound still holds relative to the optimizer over feasible states of belief. As demonstrated in our experiments, restricting our attention to classes of simple descriptions provides a tractable means to discover models and control complexity.

Corollary 23 ***Quantifying Memorization.*** *We can write the combined model complexity terms as* $\boldsymbol{\chi}[\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\psi})] = D_{KL}[\,\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta}, \boldsymbol{\psi})\,]$ *and let* $\mathbf{r}^*(\boldsymbol{\psi}, \boldsymbol{\theta})$ *be the constrained optimizer of the parsimony objective, restricted to whatever feasible set of distributions we are willing to consider. Let the optimal predictions be written as*

$$\mathbf{r}^*(\boldsymbol{y}) = \int d\boldsymbol{\psi}\, d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \mathbf{r}^*(\boldsymbol{\theta}, \boldsymbol{\psi}).$$

Every feasible alternative $\mathbf{r}(\boldsymbol{\psi}, \boldsymbol{\theta})$ *must satisfy*

$$\boldsymbol{\chi}[\mathbf{r}(\boldsymbol{\psi}, \boldsymbol{\theta})] - \boldsymbol{\chi}[\mathbf{r}^*(\boldsymbol{\psi}, \boldsymbol{\theta})] \geq \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\psi})}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{y} | \tilde{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{r}^*(\boldsymbol{y})\,]\,,$$

showing that any increased agreement with training data can only be achieved by a still greater increase in model complexity.

4. Implementation

The parsimony objective acts on opportunities for compression to reduce the complexity of our belief over models through both the description of prior belief and the information gained due to inference. While there are many ways to encode the concepts we need to articulate prior belief, compression is only possible if the interpreter admits a range of code lengths. Consequently, it is important to review some efficient encodings, capable of expressing increasing degrees of specificity with longer codes, that are needed by our prototype implementations. Then we discuss our algorithms for polynomial regression followed by decision trees.

4.1 Useful Encodings

Sometimes we need to identify one of multiple states without any principle that would allow us to break the symmetry among potential outcomes. For example, our decision tree algorithm requires a feature dimension to be specified from n possibilities. Laplace's principle of insufficient reason indicates that our encoding should not break symmetry among hypothetical permutations of the features. We can easily handle this case by representing each state with a single symbol from an alphabet of n possibilities.

As the cardinality of the set increases, however, the information provided by realizing a symbol increases logarithmically. Thus, this approach cannot hold when we have countably infinite sets, such as with integers or rational numbers, or information would diverge. Instead, we must break symmetry with either some notion of magnitude, some notion of precision, or both.

Nonnegative integers

Rissanen's universal prior over integers (Rissanen, 1983) can be derived by counting outcomes over binary sequences of increasing length. Provided the sequence length is known, any nonnegative integer z can be encoded with $\lfloor \log_2(z+1) \rfloor$ binary digits, as shown in Table 2.1. Yet, the sequence length is also a nonnegative integer, thus a recursive encoding of arbitrary nonnegative integers will have length approaching

$$\begin{aligned}\log_2^*(z) = &\lfloor \log_2(z+1) \rfloor + \lfloor \log_2 \left(\lfloor \log_2(z+1) \rfloor + 1 \right) \rfloor \\ &+ \lfloor \log_2 \left(\lfloor \log_2 \left(\lfloor \log_2(z+1) \rfloor + 1 \right) \rfloor + 1 \right) \rfloor + \cdots .\end{aligned}$$

Table 2.2 shows how the first few Rissanen codes are formed. This encoding becomes very efficient for large integers, but the number of length recursions must be set high enough. If most of the integers we need are small, a unary encoding can also provide good compression. Table 2.3 provides a comparison.

Sequence		0	1	00	01	10	11	000	001	$\cdots$
Index	0	1	2	3	4	5	6	7	8	$\cdots$

Table 2.1: Enumeration of binary sequences of increasing length.

$\boldsymbol{\psi}_0$	0	1	1	1	1	1	1
z_0	0	1	1	1	1	1	1
$\boldsymbol{\psi}_1$		0	0	1	1	1	1
z_1	0	1	1	2	2	2	2
$\boldsymbol{\psi}_2$		0	1	00	01	10	11
z_2	0	1	2	3	4	5	6

Table 2.2: The first sequence $\boldsymbol{\psi}_0$ has an implied length of 1 bit. The represented outcome z_0 indicates the length of $\boldsymbol{\psi}_1$ and so on. Rissanen$_i$ codes are formed by concatenation $(\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_i)$. With three length recursions, numbers 0 through 126 are compressed to use between 1 and 9 bits.

z	0	1	2	3	4	5	6	7
Unary	0	10	110	1110	11110	111110	1111110	11111110
Rissanen$_1$	0	10	11	n.r.	n.r.	n.r.	n.r.	n.r.
Rissanen$_2$	0	100	101	1100	1101	1110	1111	n.r.
Rissanen$_3$	0	1000	1001	10100	10101	10110	10111	1100000

Table 2.3: Nonnegative integers with unary codes and Rissanen codes. Integers that have no representation with the given encoding are indicated by n.r. .

Binary Fractions

It will also be useful to represent a dense distribution of fractions on the open unit interval $\boldsymbol{q} \in (0, 1)$, thus allowing us to approximate any real number to arbitrary precision by a variety of potential transformations. Binary fractions, with a denominator that is an integer power of 2 and a numerator that is odd, provide such a set with a convenient encoding. These fractions can be written as

$$q = \frac{2i-1}{2^{z+1}} \quad \text{where} \quad i \in [2^z]$$

and z is a nonnegative integer. If we desire all fractions of a specific precision, corresponding a fixed z, to have the same encoding length, then the numerator may be regarded as a single symbol with 2^z outcomes, providing z bits. We simply need to encode the precision z with one of the integer encodings above. The unary encoding provides a simple and attractive solution for this purpose because it does not require us to commit to a maximum precision. Table 2.4 shows some examples. If we translate and scale q to represent an angle on the real Riemann circle, the corresponding real numbers are $r = \tan(\pi(q - 1/2))$, but other choices are also possible, such as inverting the normal cumulative distribution function $r = \sqrt{2}\,\mathrm{erf}^{-1}(2q-1)$. We can also easily set the scale of outcomes by multiplying the result by some $\sigma > 0$.

q	1/2	1/4	3/4	1/8	3/8	5/8	7/8	1/16	$\cdots$
code	0	100	101	11000	11001	11010	11011	1110000	$\cdots$

Table 2.4: Leading binary fractions on the open unit interval with a unary encoding of z.

4.2 Polynomial Regression

Our regression prototype directly encodes a polynomial within a description and captures uncertainty with a hyperposterior ensemble. Since any nth degree polynomial can be written as a linear combination of $n+1$ basis functions of ascending degree, we first need to identify the degree of polynomial. A unary encoding serves this purpose without placing a limit on the maximum polynomial degree and ensures that each additional degree increases model complexity by the same amount. Coefficients are represented in the Chebyshev basis. Since critical points equioscillate in this basis, the corresponding polynomial coefficients are interpretable as the length scales of oscillation. Because this basis is known à priori, we expect all $n+1$ coefficients to take nontrival values. For this reason, the encoding implicitly reserves a code segment for each coefficient rather than attempting to use a sparse encoding. Still, natural sparsity arises from binary fractional codes, representing angles on the Riemann circle, after transforming them to the corresponding real numbers.

Algorithm 1 is a nonreversible sequence of reversible samples over all polynomial coefficients that we are willing to consider, which satisfies the requirement to obtain an ensemble that asymptotically approaches that of the posterior. Our sampler proposes polynomials up to 20th degree by constructing the set of all perturbations of both the leading nonzero coefficient, which determines the polynomial degree, and other coefficients in a randomly permuted order. All other basis coefficients are held fixed in each proposal set. For each coefficient, the sampler considers all binary fractions with $z \leq 4$. Since the polynomial representation automatically reserves symbols to describe coefficients up to the last nonzero coefficient, the posterior is most sensitive to perturbations of the last nonzero coefficient. This technique allows us to periodically sample all joint perturbations of the leading coefficient and any other coefficient. Figure 2.3 shows results.

Algorithm 1 Parsimonious Polynomial Regression Gibb's Sampler

```
Require: Vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ provide observed abscissas and ordinates, respectively. The length scale
    of $\boldsymbol{y}$ is set so that the known intrinsic stochasticity of the process is $\sigma = 1$.
Ensure: $\boldsymbol{\Psi} = \{\boldsymbol{\psi}\}$ is an ensemble of hyperposterior polynomial descriptions $\boldsymbol{\psi}_i \sim \mathbf{p}(\boldsymbol{\psi} \mid \boldsymbol{X}, \boldsymbol{y})$.
function ParsimoniousRegression($\boldsymbol{x}, \boldsymbol{y}$)
    Initialize $\boldsymbol{\psi}$ to the zero polynomial
    for each sample iteration $i = 1, 2, \ldots$ do
        Determine randomly permuted order of polynomial coefficients.
        for each permuted basis coefficient $j = 1, 2, \ldots, b$. do
            Identify the leading nonzero coefficient $k$ in $\boldsymbol{\psi}$.
            Form tensor product of all representable perturbations over both $j$ and $k$.
            Update $\boldsymbol{\psi}$ by sampling the conditional posterior and add it to the ensemble.
        end for
    end for
end function
```

4.3 Decision Trees

Decision trees learn to predict discrete classifications, labels, as a sequence of binary decisions. Each case to be predicted is represented by a feature vector $(\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_k)$ and each feature must be comparable to a learned threshold, thus facilitating a binary decision.

The decision process begins within the root node, representing the full domain of potential input features. Each binary branching decision must identify both a feature category, or mode, and the comparison threshold. We classify the outcomes of the comparsion as indicating membership in either left or right child nodes. The sequence of binary decisions effectively filters the case through a series of increasingly restrictive partitions, each

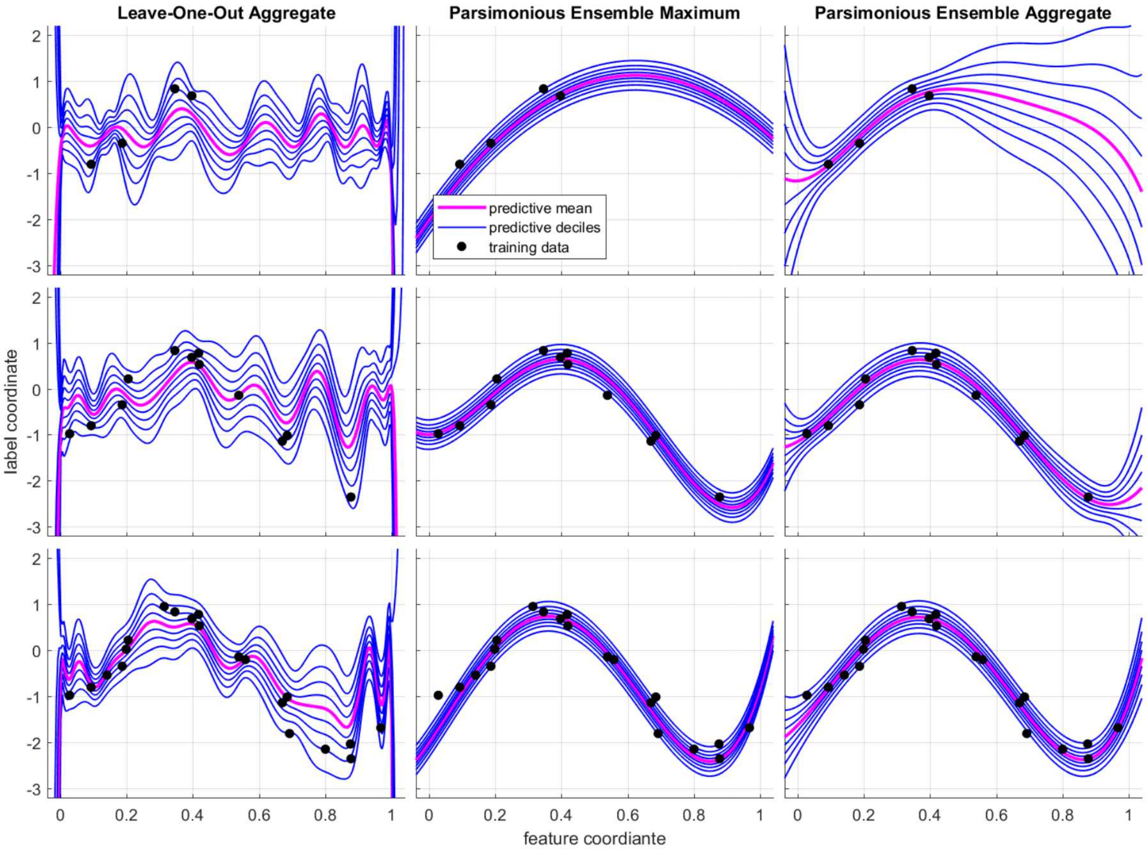

Figure 2.3: Polynomial regression experiments comparing leave-one-out cross validation, the hyper MAP, and the hyper posterior aggregate. All data are drawn from the same ground truth. The first row shows how the hyper MAP is much simpler than the leave-one-out aggregate. The hyper posterior aggregate demonstrates extrapolation risk as increasing uncertainty as we move away from the data. Increasing the number of observations provides modest increases in complexity and reduces uncertainty in the aggregate.

of which ideally provides some simplification to the classification problem. This filtration process terminates at a leaf node, which then issues label probabilities.

Decision trees are grown from training data in a recursive process, beginning at the root node. We consider the set of training cases that are members of a given node and we must either construct a branch decision or compute final leaf probabilities. The conventional procedure is to evaluate every potential splitting outcome with some utility function and choose the optimizer. While there are many varieties of utility functions used in practice, a standard information-theoretic approach maximizes the reduction in entropy due to the splitting. Given d label outcomes, let $\boldsymbol{c}_i$ represent the count of cases with a specific label, indexed by $i \in [d]$, within the domain partition corresponding to a given node. If the given node were a leaf node, the conventional frequentist approach to predicting label probabilities uses the sample mean

$$\boldsymbol{\mu}_i = \frac{\boldsymbol{c}_i}{c} \quad \text{where} \quad c = \sum_{i=1}^{d} \boldsymbol{c}_i.$$

Likewise, we denote the corresponding variables regarding hypothetical left and right child nodes, resulting from a potential splitting, using superscripts such as $\boldsymbol{c}_i^{(L)}$ and $\boldsymbol{c}_i^{(R)}$, respectively. The reduction in entropy corresponding to a potential splitting, weighted by the fraction of cases that appear within the respective child feature domains, is

$$\Delta S = \frac{c^{(L)}}{c} \sum_{i=1}^{d} \boldsymbol{\mu}_i^{(L)} \log_2(\boldsymbol{\mu}_i^{(L)}) + \frac{c^{(R)}}{c} \sum_{i=1}^{d} \boldsymbol{\mu}_i^{(R)} \log_2(\boldsymbol{\mu}_i^{(R)}) - \sum_{i=1}^{d} \boldsymbol{\mu}_i \log_2(\boldsymbol{\mu}_i).$$

The recursion typically continues until there is no reduction in entropy, i.e. each leaf node contains only a single label.

Bootstrap aggregation constructs a forest of decision trees by resampling the dataset. This consists of forming a new dataset, the same size as the original, by sampling the orginal dataset uniformly with replacement. Predictions are aggregated among trees uniformly, by taking the average of predictions from each decision tree.

Similarly, our approach encodes prior belief as a sequence of splitting decisions that define the partition of feature coordinates. Within any node, we can translate and scale features in a given dimension to the unit interval $[0, 1]$. Since it is never useful to split at either 0 or 1, we can represent the split location using a binary fraction on the open unit interval.

Within a leaf node, let $\boldsymbol{\theta}$ represent the vector of label probabilities over all potential outcomes. Before any data are observed, the leaf node has a flat Dirichlet distribution over the simplex of all potential models. We can perform exact inference using label counts $\boldsymbol{c}_i$ to recover Laplace's rule of succession

$$\mathbb{E}_{\mathbf{p}(\boldsymbol{\theta}|\check{\boldsymbol{y}})}\left[\boldsymbol{\theta}_i\right] = \frac{\boldsymbol{c}_i + 1}{c + d}.$$

Because we are performing exact inference, we could skip the following information analysis and compute the log likelihood of label counts directly

$$\log_2\left(\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\psi})\right) = \log_2\left(\frac{\Gamma(d)}{\Gamma(c+d)}\right) + \sum_{i=1}^{d} \log_2(\Gamma(\boldsymbol{c}_i + 1)).$$

We can use this opportunity, however, to demonstrate how the information analysis would proceed if a variational approximation were used. The amount of information due to change in model belief is

$$\begin{aligned} & D_{KL}[\,\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta})\,] \\ & = \frac{1}{\log(2)}\left(\log\left(\frac{\Gamma(c+d)}{\Gamma(d)}\right) - c\digamma(c+d) + \sum_{i=1}^{d} \boldsymbol{c}_i \digamma(\boldsymbol{c}_i + 1) - \log(\Gamma(\boldsymbol{c}_i + 1))\right) \end{aligned}$$

where $\digamma(x) = \frac{d}{dx}\log(\Gamma(x))$ is the digamma function. The prediction information gained about labels from a uniform prior is

$$\mathbb{E}_{\mathbf{p}(\boldsymbol{\theta}|\check{\boldsymbol{y}})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y}\mid\boldsymbol{\theta})\,\|\,\mathbf{p}(\boldsymbol{y}\mid\boldsymbol{\theta}_0)\,] = \frac{1}{\log(2)}\left(c\log(d) - c\digamma(c+d) + \sum_{i=1}^{d}\boldsymbol{c}_i\digamma(\boldsymbol{c}_i+1)\right).$$

Subtracting inference information from predictive information recovers the log likelihood up to an additive constant. The variational versions would simply replace $\mathbf{p}(\boldsymbol{\theta}\mid\check{\boldsymbol{y}})$ with any restricted distribution $\mathbf{r}(\boldsymbol{\theta})$.

Our parsimonious decision trees are formed by calling Algorithm 2 on the full training dataset to form the root node. Internal child nodes are constructed recursively. Aggregating predictions over an ensemble of many trees approximates the hyperposterior. By noting both the proposal probability and the posterior probability, up to an unknown normalization constant, we use importance weighting to approximate the posterior integral. For a tree t with description $\boldsymbol{\psi}_t$, the weights that compose a convex combination of predictions are

$$w_t \propto \frac{\mathbf{p}(\boldsymbol{\psi}_t\mid\check{\boldsymbol{y}})}{\mathbf{s}(\boldsymbol{\psi}_t)} \quad\text{so that}\quad \sum_t w_t = 1$$

where $\mathbf{s}(\boldsymbol{\psi}_t)$ is the composite probability of the sequence of samples that generated $\boldsymbol{\psi}_t$.

Algorithm 2 Parsimonious Node Construction

```
Require: $\boldsymbol{X}$ is a matrix of feature coordinates and $\boldsymbol{y}$ is the corresponding vector of labels for data
    in a partition of the feature domain.
Ensure: $\boldsymbol{\psi}$ and $\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\psi})$ sample annealed hyperposterior.
function ParsimonyNode($\boldsymbol{X}, \boldsymbol{y}$)
    for each enumerated node outcome $i = 1, 2, \dots$ do
        Construct leaf description or splitting $\boldsymbol{\psi}_i$.
        Approximate likelihood $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\psi}_i)$, assuming children will be leaf nodes.
        Compute annealed posterior $\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\psi}_i)^\alpha \mathbf{p}(\boldsymbol{\psi}_i)$ (unnormalized).
    end for
    Sample a description $\boldsymbol{\psi}_i$ from annealed posterior approximations.
    Record sample probability.
    if $\boldsymbol{\psi}_i$ is a branch node then
        Partition data into $(\boldsymbol{X}_{\text{left}}, \boldsymbol{y}_{\text{left}})$ and $(\boldsymbol{X}_{\text{right}}, \boldsymbol{y}_{\text{right}})$.
        Recursively construct left child: ParsimonyNode($\boldsymbol{X}_{\text{left}}, \boldsymbol{y}_{\text{left}}$)
        Recursively construct right child: ParsimonyNode($\boldsymbol{X}_{\text{right}}, \boldsymbol{y}_{\text{right}}$)
        Concatenate descriptions and posteriors from each partition.
    else
        Infer label posterior $\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ from a flat prior within this feature partition.
    end if
end function
```

Our first set of decision tree experiments, Figure 2.4, examines learning from a generative process that cleanly partitions the data into regions containing a single label. Row one compares models that learn from a single sample. Rows two and three learn from 25 and 100 samples, respectively. The leading two columns compare a conventional decision tree with a typical parsimonious decision tree. The last two columns compare bootstrap aggregation with our hyperposterior aggregate.

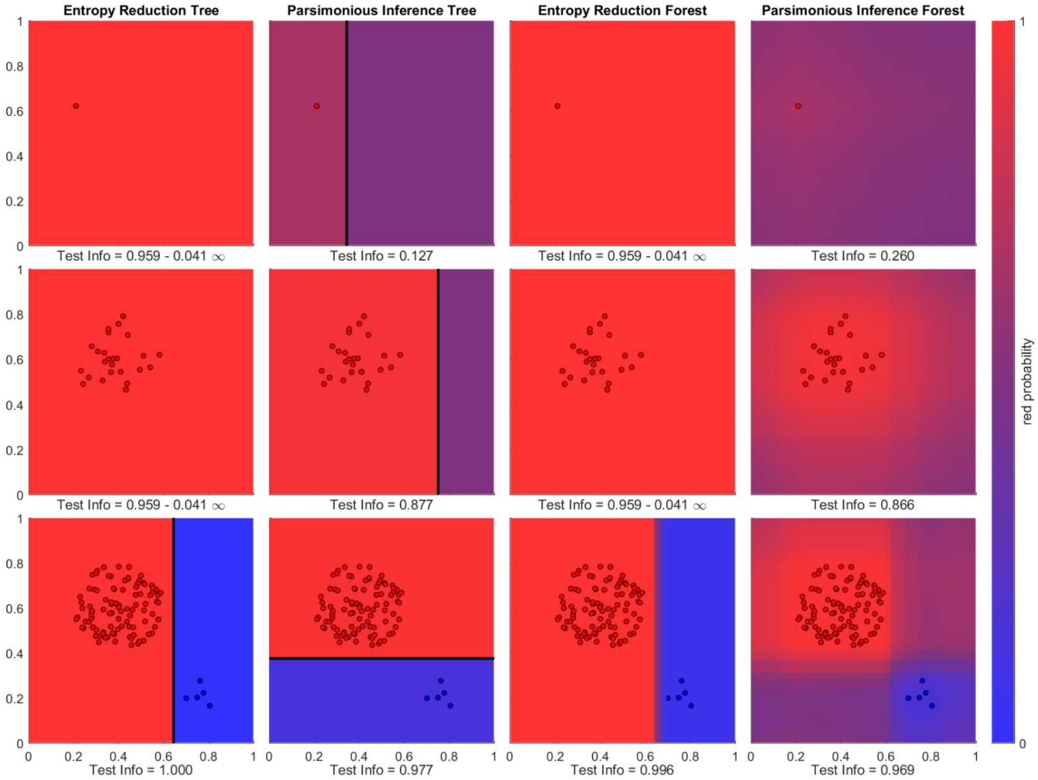

Figure 2.4: This first set of decision tree experiments uses a highly-skewed generative process with well-separated label domains. Conventional decision trees and random forests, columns 1 and 3, obtain highly confident predictions despite having few data, whereas our parsimonious trees and hyperposterior aggregates, columns 2 and 4, only gradually reduce uncertainty. Our aggregates demonstrate extrapolation uncertainty as the prediction domain departs from the data.

This is a highly skewed generative process. Even with 25 samples, the second row still has no realizations of a blue label. Yet, the parsimonious aggregate predictions in both rows 1 and 2 naturally increase uncertainty as the prediction domain deviates from the training data. In stark contrast, the conventional approaches generate completely certain predictions in regions that lack any data. Blue labels finally appear in the third row, showing how the parsimonious forest reacts to skewed data. This particular generative process is incapable of generating data in the off-diagonal regions, but without any way of knowing that, we should be highly skeptical of predictions bearing certainty in the absense of evidence.

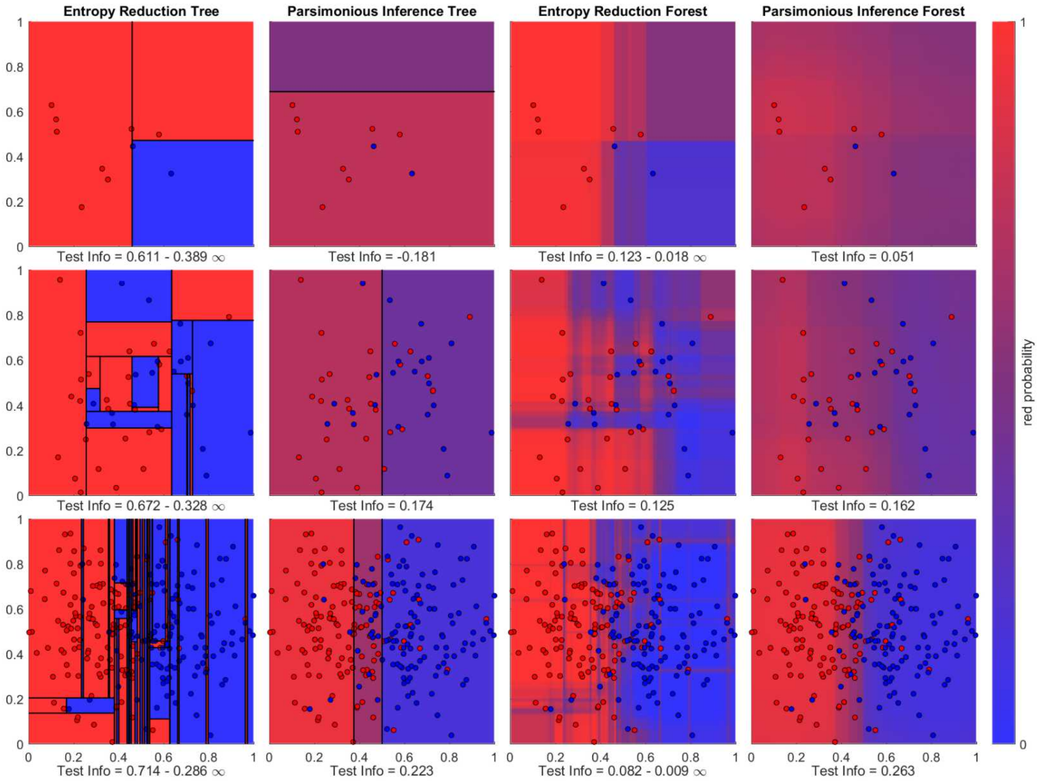

Figure 2.5: This second set of decision tree experiments demonstrates a generative process with smooth mixing of labels. Conventional trees increase complexity rapidly with increasing data. In contrast, parsimonious trees remain simple. Conventional forests exhibit confident predictive artifacts that hew to few datapoints, whereas the parsimonious forests only gain confidence with sufficient evidence.

The second set of experiments, Figure 2.5, examines a generative process that mixes labels. There is nonzero probability of generating a point of either label at any location, but red labels are more likely to appear on the left and blue on the right. We first observe that typical decision trees are more complicated than their parsimonious counterparts, as expected. Moreover, the complexity increases severely as the number of training observations increases. In contrast, parsimonious decision trees increase complexity more gradually. We also observe how the typical approach generates confident artifacts in the predictive structure that hew to few data points. In contrast, the parsimonious trees and parsimonious forests only gradually reduce uncertainty.

5. Discussion

Because our formulation of complexity accounts for information within arbitrary descriptions, it already contains the functionality needed to address a wide variety of challenges. First, we examine how to include other sources of prior belief and prior belief in interpreters within the same framework. Second, we discuss how changing the scope of descriptions to account for symbols generated and communicated by elementary operations during the evaluation of predictions provides a mechanism to prefer fast algorithms. Third, we explore how other Bayesian hyperpriors relate to description complexity. Fourth, we compare the non-Bayesian treatment of probability in Rissanen's Minimum Description Length to our approach. Finally, we offer our concluding remarks.

5.1 Comparing and Inferring Interpreters

While the difficulty of expressing prior belief for abstract complex models motivates this research, when we have access to additional information that could constrain prior beliefs, that information may be very impactful. Therefore, we should also be able to integrate these prior beliefs within the general complexity framework. Let our complexity-based prior belief be denoted as $\mathbf{p}(\boldsymbol{a} \mid \mathcal{C}) = 2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}$. If, additionally, we have other prior beliefs about the model, $\mathbf{p}(\boldsymbol{a} \mid \mathcal{B})$, we can form the combined prior-beliefs as $\mathbf{p}(\boldsymbol{a} \mid \mathcal{B},\mathcal{C})$. For example, $\mathcal{B}$ may express physical laws or previously observed data. One approach would be to use an interpreter that implicity embeds $\mathcal{B}$ within viable encodings so that $\mathbf{p}(\boldsymbol{a} \mid \mathcal{B},\mathcal{C}) = 2^{-\ell(\boldsymbol{\psi}_{\mathcal{B}}(\boldsymbol{a}))}$. Alternatively, if we assume that belief derived from $\mathcal{B}$ is conditionally independent of our complexity-based belief $\mathcal{C}$, then we have

$$\begin{aligned}\mathbf{p}(\boldsymbol{a} \mid \mathcal{B},\mathcal{C}) &= \frac{\mathbf{p}(\mathcal{B} \mid \boldsymbol{a},\mathcal{C})\mathbf{p}(\boldsymbol{a} \mid \mathcal{C})}{\mathbf{p}(\mathcal{B} \mid \mathcal{C})} = \frac{\mathbf{p}(\mathcal{B} \mid \boldsymbol{a})\mathbf{p}(\boldsymbol{a} \mid \mathcal{C})}{\mathbf{p}(\mathcal{B} \mid \mathcal{C})} \\ &= \frac{\mathbf{p}(\boldsymbol{a} \mid \mathcal{B})\mathbf{p}(\mathcal{B})\mathbf{p}(\boldsymbol{a} \mid \mathcal{C})}{\mathbf{p}(\boldsymbol{a})\mathbf{p}(\mathcal{B} \mid \mathcal{C})} \propto \mathbf{p}(\boldsymbol{a} \mid \mathcal{B})\mathbf{p}(\boldsymbol{a} \mid \mathcal{C}).\end{aligned}$$

Here we hold that $\mathbf{p}(\boldsymbol{a})$ must be uniform over all $\boldsymbol{a}$ since $\mathcal{B}$ and $\mathcal{C}$ have been constructed to capture all our beliefs. Thus, the composite prior is easily expressed up to a constant of proportionality.

We can also compare choices of interpreter within this framework by identifying a common language in which each interpreter is defined. When different interpreters are used, $\ell(\boldsymbol{\psi}(\boldsymbol{a}))$ will differ for a fixed state of belief and thus have a different prior probability. To compare different interpreters, or infer the interpreter from data, we must construct the prior $\mathbf{p}(\boldsymbol{\varphi})$. Fortunately, we already have the machinery to do this since $\boldsymbol{\varphi}$ is just a program taking $\boldsymbol{\psi}$ as input and equating it to a state of belief over explanatory models. Thus we define $\mathbf{p}(\boldsymbol{\varphi}) = 2^{-\ell(\boldsymbol{\psi}^*(\boldsymbol{\varphi}))}$, where $\boldsymbol{\psi}^*(\boldsymbol{\varphi})$ is constructed relative to a common language $\boldsymbol{\varphi}^*$. If potential interpreters are enumerated as $\boldsymbol{\varphi}_i$ for $i \in [n]$ and the corresponding descriptions are $\boldsymbol{\psi}_i(\boldsymbol{a})$, then resulting objective is equivalent to augmenting our model descriptions with a description of the interpreter to obtain composite description lengths $\ell(\boldsymbol{\psi}_i(\boldsymbol{a}))+\ell(\boldsymbol{\psi}^*(\boldsymbol{\varphi}_i))$.

The benefit of this view is it allows us to see how short and simple interpreters are generally more plausible. Nefarious interpreters, such as the third basis in Figure 2.2, effectively transfer complexity from an otherwise long encoding, subject to a simple interpreter, into a short encoding, subject to a long interpreter. Yet, when we shift the derivation of plausibility to a common language, the excessive complexity becomes visible. We conclude that an objectivist view of credible interpreters, in the absense of additional justification for prior belief, should be short and simple.

This approach is a formulation of grammar discovery, which is a core aspect of learning. If we have several inference problems that we believe should be well-explained within a common language, then we can infer an efficient interpreter from these datasets. An interpreter that represents common functions between the different learning problems efficiently will be more likely than one that solves a single problem well by hiding complex functions within shortcut codes.

5.2 The Imperative of Utility

It is simply not viable to have a theory that requires infinite computational power to obtain useful predictions. Practical models must be discoverable and predictions must be computable. The Kolmogorov complexity is well-known to be uncomputable, thus raising a natural concern that generalizing prior belief to arbitrary descriptions only exacerbates the problem. Yet, the primary purpose of Theorem 2 is to show how information theory allows us to restrict our attention to feasible manifolds of belief, while simultaneously allowing us to compare outcomes from different choices of restriction. Because long descriptions are already exponentially suppressed à priori, the information we generate by refusing to consider long descriptions becomes small as the descriptions we drop become long.

Even so, it is instructive to examine uncomputability more careful as it motivates future directions. Suppose we had an oracle Ω that would determine whether or not a program $\boldsymbol{a}$ is capable of reproducing a mapping $(\boldsymbol{x}, \boldsymbol{y})$ in a finite amount of time

$$\Omega(\boldsymbol{\varphi}, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \text{true} & \boldsymbol{y} = \boldsymbol{\varphi}(\boldsymbol{a}, \boldsymbol{x}) \\ \text{false} & \text{otherwise.} \end{cases}$$

This is a stronger criterion than the well-known halting problem. The existance of such an oracle would allow us to determine the Kolmogorov complexity by brute force, generating and checking programs in order of increasing length until the oracle returns true. Further, it would be a trivial matter to write another brute force subroutine to identify the first sequence $\boldsymbol{y}$ with Kolmogorov complexity above an arbitrarily high limit $\mathcal{K}_{\boldsymbol{\varphi}}(\emptyset, \boldsymbol{y}) > \tau$. By setting τ to exceed the combined lengths of the oracle and brute force subroutines, we would have succeded in writing a program that contradicts the Kolmogorov complexity.

The core problem with this thought experiment is the arbitrarily large amount of memory and elementary operations that would be required to run the program. Disregarding the halting problem, the brute force Kolmogorov complexity function alone needs to generate full programs in memory, but it only incurs the cost of programming a counter. We may conclude that problems associated with computability will be alleviated within this theory if we simply include memory operations, every symbol generated or transmitted between slow and fast levels of memory, in the definition of program *execution* length. This definition would articulate prior belief that prefers efficient algorithms, allowing us to restrict our attention to models that can be evaluated within a limited computational budget. Lempel and Ziv (1976) present a related framework to measure sequence production complexity as the minimum number of steps required to build a sequence from a production process to construct a heirarchy of subsequences. Speidel (2008) provides additional discussion of recent work by Titchener (1998).

In this view, prior belief becomes an expression for the degree of utility considering a model would contribute to obtaining feasible predictions. It is not useful to consider models that, in order to provide a substantial contribution to predictions, would require more evidence than we anticipate having. Likewise, it is not useful to consider models that would require either more computational power, or time to evaluate, than is practical to generate useable predictions. This simple modification provides a principled foundation to prefer algorithms that may be more difficult to program, but also achieve faster results. For example, randomized algorithms such as Randomized QR with Column Pivoting (RQRCP) (Duersch and Gu, 2020) would gain plausibility by having reduced slow communication bottlenecks, despite incurring a controllable increase in the uncertainty of conditions that are used to make decisions. In order for machine learning to be capable of providing discoverable, computationally feasible, and useful models, we cannot avoid limiting our attention accordingly.

5.3 Relationship to other Bayesian methods

Our hyperprior formulation provides a principled foundation to derive results that are similar in function to several well-known methods for specific Bayesian inference problems. Notable comparisons include sparsity inducing priors, like Automatic Relevance Determination, for regression problems with continuous coefficients. The ARD prior is a hyperprior over parameters which is intended to identify critical parameters and drive remaining parameters towards zero. The ARD prior is implemented as:

$$\mathbf{p}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid 0, \boldsymbol{\sigma})$$

where we need to specify $\mathbf{p}(\boldsymbol{\sigma})$. In the original work that introduced the ARD prior, $\mathbf{p}(\boldsymbol{\sigma})$ was taken to be a gamma distribution in the precision $\tau = \sigma^{-2}$ with a small shape parameter. This closely corresponds to the improper Jeffrey's prior $\mathbf{p}(\boldsymbol{\sigma}) \propto \frac{1}{\boldsymbol{\sigma}}$, often used in practice for unknown scalar covariances. If we partition the potential values of $\boldsymbol{\sigma}$ into intervals $0 < a < b < \infty$, where a and b are any positive real numbers, the cumulative probability diverges for values less than a and values greater than b, dominating over the finite contribution within the interval $[a, b]$. It follows that sampling the Jeffrey's prior would yield outcomes either very close to zero or diverging towards infinity. Within this formulation, if $\boldsymbol{\theta}$ has little relevance to the likelihood, then probability is maximized when $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ approach zero. Otherwise, a large enough $\boldsymbol{\sigma}$ will be found to allow $\boldsymbol{\theta}$ to take moderate nonzero values with the Jeffrey's prior introducing only a slight penalty as $\boldsymbol{\sigma}$ increases. Therefore, it can be interpreted as making a binary choice between very large or very small σ. More generally, a gamma distribution allows, indeed requires, the relative probably of these two outcomes to be tuned.

If we uniformly discretize the possible values of $\boldsymbol{\sigma}$ as $\boldsymbol{\sigma}_i = i\boldsymbol{\sigma}_1$ and assign them probabilities according to $\mathbf{p}(\boldsymbol{\sigma}_i) \propto \frac{1}{\boldsymbol{\sigma}_i}$, we can equate this prior with the complexity prior $\mathbf{p}(\sigma_i) = 2^{-\ell(\sigma_i)}$ to obtain

$$\ell(\boldsymbol{\sigma}_i) = \log_2(\boldsymbol{\sigma}_i) + c = \log_2(i) + \log_2(\boldsymbol{\sigma}_1) + c = \log_2(i) + \ell(\boldsymbol{\sigma}_1),$$

where the constant c corresponds to the normalization. The complexity of $\boldsymbol{\sigma}_i$ increases logarithmically. Thus, the Jeffrey's prior is the continuous limit of the number of bits required to express an integer number of fixed increments in $\boldsymbol{\sigma}$, i.e. uniformly within the exponential term.

While sparsity is a useful notion of complexity for many problems, it is not universal. Sparsity either regards a continuous parameter as either complex (nonzero) or not complex (zero). While sparsity-inducing priors, like ARD, can compel continuous parameters to zero if they do not provide enough benefit to predictions, they have no affordance to suppress other forms of complexity. For example, there is no compelling notion of sparsity within the construction of decision trees. Moreover, when we need to encode constants within prior descriptions, our theory supports consistent distinctions in complexity among potential constants.

5.4 Relationship to Minimum Description Length

Rissanen's Minimum Description Length (MDL) shares many similarities with our theory, but it is not motivated by the philosophical foundations of reason and learning that drive the Bayesian perspective. Rather, MDL views inference as finding an optimal compressed representation of the data and probability as a way of developing efficient codes. MDL representations contain both the model used to construct an efficient code and the compressed form of the data that follows. The length of the data representation $\ell(\mathcal{D})$ is the sum of the number of bits needed to describe the model $\ell(\boldsymbol{a})$ and the number of bits needed to describe the residual data $\ell(\mathcal{D} \mid \boldsymbol{a})$. In its simplest form, the inference problem for identifying a hypothesis or program $\boldsymbol{a} \in \mathcal{A}$ is

$$\ell(\mathcal{D}) = \min_{\boldsymbol{a} \in \mathcal{A}} \quad \ell(\boldsymbol{a}) + \ell(\mathcal{D} \mid \boldsymbol{a}),$$

requiring a specific discretization and encoding for a hypothesis space $\mathcal{A}$. To address the arbitrary task of designing the encoding of a hypothesis space, MDL proposes using minimax optimal universal codes that minimize the worst-case regret of using that encoder to encode arbitrary data. This simple form of MDL can also be refined to compare and optimize hypothesis classes instead of individual hypotheses, which corresponds to the Bayesian model-class selection problem.

While under certain conditions there is an equivalence between MDL and a Bayesian formulation this is not true under all circumstances. Particular similarity is observed with objective Bayesian formulations such as Jeffreys' priors. The key difference is the necessity of the prior from the Bayesian perspective. A prior represents our beliefs and it is updated as new data becomes available. Within MDL the encoding of the hypothesis may be thought of as encoding a prior, but this encoding is not defined with respect to any prior beliefs wither subjective or otherwise. Within the minimax framework, it is a property of the model class and the possible data. The Minimum Message Length (MML) framework is a similar Bayesian framework to MDL, where the encoding can be specified to capture prior belief. We hold with the Bayesian view that prior information is a critical aspect of inference. Our theory provides a mechanism for integrating prior belief but also optimizing the prior representation. This means that simple priors that are consistent with our beliefs are preferred.

Further, we highlight the difference in the representation and optimization that our theory provides compared to MDL and MML. Optimization is fundamentally inconsistent with Bayesian probability theory. Inference updates belief from a state of prior belief to posterior belief, it does explicitly provide the optimal model but instead expresses beliefs about possible models. This is why in our theory we distinguish rational choice from rational belief. We update rational beliefs using the full formalism of Bayesian inference but recognize that ultimately a choice must be made to simplify this process so that problems can be solved on machines with finite computational resources. Rational choices require a utility function and are informed by our rational belief. Building on Bernardo's work (Bernardo, 1979), Corollary 21, Corollary 22, Corollary 10, and Corollary 11 show the variety of circumstances in which information serves to guide well-posed optimization of belief. The rational choice we make becomes the representation we use to approximate posterior belief for future predictions. A rational choice could be finding a single model that best captures the posterior or the hyperparameters, as in MDL and MML. However, other representations are likely to have greater utility and provide better prediction uncertainty quantification.

5.5 Summary and Conclusion

We proposed Parismonious Inference, a complete theory of learning based on an information-theoretic formulation of Bayesian inference that quantifies and suppresses a general notion of explanatory complexity. We showed how our information-theoretic objective allows us to understand the relationship between model complexity and increased agreement between corresponding predictions and data labels.

Within the Bayesian perspective, once the prior, the likelihood, and the data are specified, the posterior inexorably follows. Yet, when we consider the infinit varieties of algorithms that may be developed in machine learning, we find that any universal prior that reserves some degree of plausibility for an arbitrary algorithm becomes uncomputable in practice. Our framework allows us to resolve the imperative of utility by allowing us to quantify the value of a choice, a restricted manifold of belief in which we only consider computationally feasible models to obtain well-justified predictions within a practical computational budget. By accounting for model complexity from first principles, we may evaluate the utility of various restrictions of prior belief, as well as feasible posterior approximations, within a single framework.

A central aspect of our framework is the distinction between the intrinsic meaning of a potential state of belief from the description needed to distinguish that state from other possibilities. Encoding complexity provides a critical missing component that is needed to measure the complexity of arbitrary inference architectures and naturally associate complexity with plausibility. Our formulation of generalized length allows us to assign length to a wide variety of codes, beyond binary codes that are typically associated with program length. We examined some elementary codes to express integers and fractions on the open interval, which can be mapped to a broad class of numbers that may prove useful to represent prior beliefs.

We showed how even feasibility-constrained optimizers satisfy quantifiable memorization bounds in comparison to models that may produce better adherence to training data, but at the cost of increased description length, increased inference information, or information generated by an approximating distribution proposed to generate predictions. Our experimental results show how our hyperposterior ensembles avoid developing artifacts that artificially hew to seen data within the predictive structure. Moreover, accounting for multiple explanations by hyperposterior sampling allows us to compute extrapolation uncertainty from first principels as the input domain deviates from that of past observations. These experimental results demonstrate how our theory allows us to obtain predictions from extremely small datasets without cross-validation.

Our theory solves critical challenges in understanding how we may efficiently learn from data, obtain well-grounded justification for uncertainty in predictions, and anticipate extrapolation regimes where additional data would prove most beneficial, thus opening a new domain of predictive capabilities. Further, this work provides a principled foundation to address the challenge of feasible learning and model discovery, from limited data, in the face of high dimensionality.

Chapter 3

Randomized Projection for Rank-Revealing Factorizations

Rank-revealing matrix decompositions provide an essential tool in spectral analysis of matrices, including the Singular Value Decomposition (SVD) and related low-rank approximation techniques. QR with Column Pivoting (QRCP) is usually suitable for these purposes, but it can be much slower than the unpivoted QR algorithm. For large matrices, the difference in performance is due to increased communication between the processor and slow memory, which QRCP needs in order to choose pivots during decomposition. Our main algorithm, Randomized QR with Column Pivoting (RQRCP), uses randomized projection to make pivot decisions from a much smaller sample matrix, which we can construct to reside in a faster level of memory than the original matrix. This technique may be understood as trading vastly reduced communication for a controlled increase in uncertainty during the decision process. For rank-revealing purposes, the selection mechanism in RQRCP produces results that are the same quality as the standard algorithm, but with performance near that of unpivoted QR (often an order of magnitude faster for large matrices). We also propose two formulas that facilitate further performance improvements. The first efficiently updates sample matrices to avoid computing new randomized projections. The second avoids large trailing updates during the decomposition in truncated low-rank approximations. Our truncated version of RQRCP also provides a key initial step in our truncated SVD approximation, TUXV. These advances open up a new performance domain for large matrix factorizations that will support efficient problem-solving techniques for challenging applications in science, engineering, and data analysis.

1. Introduction

QR with Column Pivoting (QRCP) is a fundamental kernel in numerical linear algebra that broadly supports scientific analysis. As a rank-revealing matrix factorization, QRCP provides the first step in efficient implementations of spectral methods such as the eigenvalue decomposition and Principal Component Analysis (PCA), also called the Singular Value Decomposition (SVD) (Higham, 2000). QRCP also plays a key role in least-squares approximation (Chan and Hansen, 1992) and stable basis extraction (Stathopoulos and Wu, 2002; Hetmaniuk and Lehoucq, 2006; Duersch et al., 2018) for other important algorithms. These methods allow us to form compressed representations of linear operators by truncation while retaining dominant features that facilitate analytic capabilities that would otherwise be impractical for large matrices. In the field of data science, PCA and its generalizations (Vidal et al., 2005) support unsupervised machine learning techniques to extract salient features in two-dimensional numerical arrays. Randomized methods have been extended further to support analysis of multidimensional data using tensor decompositions (Battaglino et al., 2018; Hong et al., 2020).

QRCP builds on the QR decomposition, which expresses a matrix $\boldsymbol{A}$ as the product of an orthogonal matrix $\boldsymbol{Q}$ and a right-triangular factor $\boldsymbol{R}$ as $\boldsymbol{A} = \boldsymbol{QR}$. Unlike the LU decomposition, or Gaussian elimination, QR always exists and may be stably computed regardless of the conditioning of $\boldsymbol{A}$. Furthermore, finely tuned library implementations use a blocked algorithm that operates with the communication efficiency of matrix-matrix multiply, or level-3 kernels in the Basic Linear Algebra Subprograms (BLAS-3). The standard QR decomposition is not, however, suitable for rank detection or low-rank approximations. These applications require a column permutation scheme to process more representative columns of $\boldsymbol{A}$ earlier in the decomposition (Chan, 1987; Bischof and Hansen, 1991). A permutation matrix $\boldsymbol{P}$ encodes these pivoting decisions so that the decomposition becomes $\boldsymbol{AP} = \boldsymbol{QR}$.

The basic QRCP approach selects an unfactorized column with a maximal 2-norm of components that do not reside within the span of previously factorized columns. This heuristic is a greedy algorithm attempting to maximize the sequence of partial determinants in $\boldsymbol{R}$, which is typically adequate for rank detection with a few notable rare exceptions, such as the Kahan matrix (Golub and Van Loan, 2013). The critical drawback, however, is that these computations suffer a substantial increase in communication between slow (large) and fast (small) levels of memory, especially for large matrices. Each pivot decision requires at least one matrix-vector multiply in series, thus limiting overall efficiency to that of level-2 kernels in the Basic Linear Algebra Subprograms (BLAS-2).

Our primary motivation to examine efficient rank-revealing algorithms is that, for large matrices or algorithms that require frequent use of QRCP, the communication bottleneck becomes prohibitively expensive and impedes utilization of this important set of analytic tools.

1.1 Our Contributions

Randomized projection allows us to reduce communication complexity at the cost of increasing uncertainty regarding latent 2-norms (and inner products) among columns in a large matrix. Randomized QR with Column Pivoting (RQRCP), shown in algorithm 6, harnesses this trade-off to obtain full blocks of pivot decisions that can be applied to $\boldsymbol{A}$ all at once. We do this by drawing a Gaussian Independent Identically Distributed (GIID) matrix, $\boldsymbol{\Omega}$, and compressing columns of $\boldsymbol{A}$ into a sample matrix, $\boldsymbol{B} = \boldsymbol{\Omega A}$. The sample matrix retains enough information about the original matrix for us to obtain a full block of pivoting decisions, while requiring far less communication to do so. Because we construct the sample to contain far fewer rows than the original, it resides in a faster level of memory than the full matrix. Having a full block of pivots allows us to update the matrix that becomes $\boldsymbol{R}$ with blocked BLAS-3 operations in the same fashion as unpivoted QR, thus entirely eliminating the communication bottleneck of BLAS-2 operations encountered in the standard algorithm.

We show how rank-revealing decompositions are well suited to this approach because each pivot decision must merely avoid selecting a column containing a relatively small component orthogonal to the span of previous pivots. As many columns are often suitable for this purpose, the use of precise 2-norm computations in standard QRCP is unnecessarily. Our approach has been adopted in subsequent work in computing matrix factorization-based low-rank approximations on sequential and parallel platforms (Erichson et al., 2019; Martinsson et al., 2019, 2017; Martinsson and Tropp, 2020; Xiao et al., 2017).

We also propose a sample update formula that reduces the number of BLAS-3 operations required to process a full matrix. Given a block of pivots, updating $\boldsymbol{R}$ with blocked Householder reflections requires two matrix-matrix multiplications. Without a formula to update the sample, we would need a new sample matrix after each block and one additional matrix-matrix multiply. Instead, we utilize computations that are needed to update $\boldsymbol{R}$ to also update $\boldsymbol{B}$ into a suitable sample for the next decision block.

RQRCP naturally extends to a truncated formulation, described in algorithm 7, that further reduces communication by avoiding trailing updates. For low-rank matrix approximations, this algorithm requires only one block matrix-matrix multiply per update. We accomplish this by storing reflector information in the compact WY notation (Puglisi, 1992), which allows us to construct each block of $\boldsymbol{R}$ without intermediate updates. Moreover, this algorithm serves as a key initial step in our approximation of the truncated SVD described in algorithm 8, which is a variant of Stewart's QLP algorithm (Stewart, 1999).

Section 2 will discuss the nature of the communication bottleneck and related approaches to address it. Section 3 will analyze sample-based pivoting as the maximization of expected utility and derive our main result, RQRCP. Section 4 will derive and explain our truncated algorithms for low-rank approximation. Section 5 will provide numerical experiments that explore the performance and decomposition quality of these approaches. Section 6 will offer concluding remarks.

2. Related Work

In order to understand how our algorithms improve performance, we first review the reasons why additional communication could not be avoided with previous approaches. Given a large matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, the QRCP decomposition can be computed using algorithm 3. This algorithm is composed of a sequence of Householder reflections. To review, a reflector $\boldsymbol{y}$ is formed from a particular column, say $\boldsymbol{a}$, by subtracting the desired reflection (which must have the same 2-norm) from the current form, such as $\boldsymbol{y} = \boldsymbol{a} - (-\text{sign}(\boldsymbol{a}_1) \left\|\boldsymbol{a}\right\|_2 \boldsymbol{e}_1)$. The negative sign of the leading element of $\boldsymbol{a}$ is used to ensure that the reflector satisfies $\left\|\boldsymbol{y}\right\|_2 \geq \left\|\boldsymbol{a}\right\|_2$ for numerical stability. The corresponding reflection coefficient is $\tau = 2/\boldsymbol{y}^T\boldsymbol{y}$, which yields the Householder reflection $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{y}\tau\boldsymbol{y}^T$.

In the algorithms that follow, an intermediate state of an array or operator at the end of iteration j is denoted by superscript (j) to emphasize when updates overwrite a previous state. In contrast, an element computed on iteration j that remains accessible in some form is denoted with a simple subscript.

Algorithm 3 QRCP with BLAS-2 Householder reflections.

```
Require:
    $\boldsymbol{A}$ is $m \times n$.
Ensure:
    $\boldsymbol{Q}$ is an $m \times m$ orthogonal matrix.
    $\boldsymbol{R}$ is an $m \times n$ right triangular matrix, magnitude of diagonals nonincreasing.
    $\boldsymbol{P}$ is an $n \times n$ permutation matrix such that $\boldsymbol{AP} = \boldsymbol{QR}$.
 1: function $[\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{P}](\mathrm{QRCP})\boldsymbol{A}$
 2:     Compute initial column 2-norms which become trailing column norms.
 3:     for $j = 1, 2, \ldots, k$, where $k = \min(m, n)$ do
 4:         Find index $p_j$ of the column with maximum trailing 2-norm.
 5:         Swap columns $j$ and $p_j$ with permutation $\boldsymbol{P}_j$.
 6:         Form Householder reflection $\boldsymbol{H}_j = \boldsymbol{I} - \boldsymbol{y}_j \tau_j \boldsymbol{y}_j^T$ from new column.
 7:         Apply reflection $\boldsymbol{A}^{(j)} = \boldsymbol{H}_j(\boldsymbol{A}^{(j-1)}\boldsymbol{P}_j)$.
 8:         Update trailing column norms by removing the contribution of row $j$.
 9:     end for
10:     $\boldsymbol{Q} = \boldsymbol{H}_1 \boldsymbol{H}_2 \ldots \boldsymbol{H}_k$ is the product of all reflections.
11:     $\boldsymbol{R} = \boldsymbol{A}^{(k)}$.
12:     $\boldsymbol{P} = \boldsymbol{P}_1 \boldsymbol{P}_2 \ldots \boldsymbol{P}_k$ is the aggregate column permutation.
13: end function
```

At the end of iteration j we can represent the matrix $\boldsymbol{A}$ as a partial factorization using the permutation $\boldsymbol{P}^{(j)} = \boldsymbol{P}_1 \ldots \boldsymbol{P}_j$, the composition of column swaps so far, and the analogous composition of Householder reflections, $\boldsymbol{Q}^{(j)} = \boldsymbol{H}_1 \ldots \boldsymbol{H}_j$, to obtain

$$\boldsymbol{AP}^{(j)} = \boldsymbol{Q}^{(j)} \begin{bmatrix} \boldsymbol{R}_{11}^{(j)} & \boldsymbol{R}_{12}^{(j)} \\ 0 & \hat{\boldsymbol{A}}^{(j)} \end{bmatrix}.$$

The leading $j - 1$ entries of the vector $\boldsymbol{y}_j$ in line 6 of Algorithm 3 are 0. The upper-left submatrix $\boldsymbol{R}_{11}^{(j)} \in \mathbb{R}^{j \times j}$ is right-triangular. Likewise, $\boldsymbol{R}_{12}^{(j)} \in \mathbb{R}^{j \times j}$ completes the leading j rows of $\boldsymbol{R}$. It only remains to process the trailing matrix $\hat{\boldsymbol{A}}^{(j)}$. On the next iteration, the 2-norm of the selected column within $\hat{\boldsymbol{A}}^{(j)}$ becomes the magnitude of the next diagonal element in $\boldsymbol{R}_{11}^{(j+1)}$. We may understand QRCP as a greedy procedure intended to maximize the magnitude of the determinant of $\boldsymbol{R}_{11}^{(j+1)}$. The new determinant magnitude is $|\det \boldsymbol{R}_{11}^{(j+1)}| = |\det \boldsymbol{R}_{11}^{(j)}| \left\| \hat{\boldsymbol{A}}^{(j)}(:, p_{j+1}) \right\|_2$, so this scheme has selected the pivot that multiplies the previous determinant by the largest factor. Note that true determinant maximization would require exchanging prior columns and adjusting the factorization accordingly (Gu and Eisenstat, 1996).

Early implementations of QR also relied on BLAS-2 kernels and, thus, gave similar performance results until reflector blocking was employed in QR (Bischof and Van Loan, 1987; Schreiber and Van Loan, 1989). To process b columns of an $m \times n$ matrix with BLAS-2 kernels, $O(bmn)$ elements must pass from slow to fast memory. Blocking improves performance by reducing communication between these layers of memory using matrix-matrix multiply.

Instead of updating the entire matrix with each Householder reflection, transformations are collected into a matrix representation that can be applied using two BLAS-3 matrix-matrix multiplies, which reduces communication complexity to $O(bmn/M^{3/2})$, where M is the size of fast memory.

In order to produce a correct pivot decision at iteration $j+1$ in QRCP, however, trailing column norms must be updated to remove the contribution of row j, which depends on the Householder transformation $\boldsymbol{H}_j$. At first glance, this update appears to require two BLAS-2 operations on the trailing matrix per iteration. The first operation computes scaled inner products $\boldsymbol{w}_j^T = \tau_j \boldsymbol{y}_j^T \boldsymbol{A}^{(j-1)} \boldsymbol{P}_j$ and the second operation modifies the trailing matrix with the rank-1 update $\boldsymbol{A}^{(j)} = \boldsymbol{A}^{(j-1)}\boldsymbol{P}_j - \boldsymbol{y}_j \boldsymbol{w}_j^T$. These two operations cause the communication bottleneck in QRCP.

2.1 Attempts to Achieve BLAS-3 Performance

Quintana-Ortí, Sun, and Bischof (Quintana-Ortí et al., 1998) were able to halve BLAS-2 operations with the insight that the trailing norm update does not require completing the full rank-1 update on each iteration. Instead, reflections can be gathered into blocks, as in QR. This method appears in algorithm 4.

At the end of iteration j, the algorithm has collected a block of reflectors $\boldsymbol{Y}^{(j)}$. Reflector $\boldsymbol{y}_i$, for $i \leq j$, appears in column i from the diagonal down. This forms a block reflection $\boldsymbol{Q}^{(j)} = \boldsymbol{I} - \boldsymbol{Y}^{(j)}\boldsymbol{T}^{(j)}\boldsymbol{Y}^{(j)T}$, where $\boldsymbol{T}^{(j)}$ is an upper triangular $j \times j$ connection matrix that can be solved from $\boldsymbol{Y}^{(j)}$ so that $\boldsymbol{Q}^{(j)}$ is orthogonal. This algorithm must also collect each corresponding scaled inner product, $\boldsymbol{w}_i^T$, which appears as row i in a matrix $\boldsymbol{W}^{(j)T}$. In the compact WY notation, this block of inner products is $\boldsymbol{W}^{(j)T} = \boldsymbol{T}^{(j)T}\boldsymbol{Y}^{(j)T}\boldsymbol{A}\boldsymbol{P}^{(j)}$, which provides enough information to update row j alone and adjust trailing column norms to prepare for the next pivot selection

$$\boldsymbol{A}^{(j)}(j,:) = \boldsymbol{A}^{(j-1)}(j,:)\boldsymbol{P}_j - \boldsymbol{Y}^{(j)}(j,:)\boldsymbol{W}^{(j)T}.$$

Note, however, that this construction complicates reflector formation. As before, the next pivot index p_{j+1} is selected and swapped into column $j+1$. Let this new column be $\boldsymbol{a}_{j+1}$. From rows $j+1$ down, elements of $\boldsymbol{a}_{j+1}$ have not been updated with the current block of reflectors. Before we can form $\boldsymbol{y}_{j+1}$, prior transformations must be applied to these rows from the formula $\hat{\boldsymbol{a}}_{j+1} = \boldsymbol{a}_{j+1} - \boldsymbol{Y}^{(j)}\boldsymbol{W}^{(j)T}(:, p_{j+1})$. An additional step is also required to compute reflector inner products because they must account for reflections that have not been applied to the trailing matrix. The adjusted formula for these inner products is

$$\boldsymbol{w}_{j+1}^T = \tau_{j+1}\left(\boldsymbol{y}_{j+1}^T\boldsymbol{A}^{(j)} - (\boldsymbol{y}_{j+1}^T\boldsymbol{Y}^{(j)})\boldsymbol{W}^{(j)T}\right)\boldsymbol{P}_{j+1}.$$

The reflector and inner product blocks are then updated:

$$\boldsymbol{Y}^{(j+1)} = \begin{bmatrix}\boldsymbol{Y}^{(j)} & \boldsymbol{y}_{j+1}\end{bmatrix} \quad \text{and} \quad \boldsymbol{W}^{(j+1)T} = \begin{bmatrix}\boldsymbol{W}^{(j)T}\boldsymbol{P}_{j+1} \\ \boldsymbol{w}_{j+1}^T\end{bmatrix}.$$

Unfortunately, the remaining BLAS-2 operations $\boldsymbol{y}_{j+1}^T\boldsymbol{A}^{(j)}$ and $\boldsymbol{y}_{j+1}^T\boldsymbol{Y}^{(j)}$ in the inner product computation still dominate slow communication complexity for large matrices. The entire trailing matrix must pass from slow to fast memory once per iteration. Consequently, even heavily optimized implementations of blocked QRCP still run substantially slower than blocked QR on both sequential and parallel architectures.

Algorithm 4 QRCP with BLAS-3 reflection blocking.

```
Require:
    $\boldsymbol{A}$ is $m \times n$.
Ensure:
    $\boldsymbol{Q}$ is an $m \times m$ orthogonal matrix.
    $\boldsymbol{R}$ is an $m \times n$ right triangular matrix, diagonals in nonincreasing magnitude order.
    $\boldsymbol{P}$ is an $n \times n$ permutation matrix such that $\boldsymbol{AP} = \boldsymbol{QR}$.
function $[\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{P}] = \mathrm{QRCP}(\boldsymbol{A})$
    Compute initial column 2-norms, which will become trailing column norms.
    for $i = 0, b, 2b \ldots$, where $b$ is block size. do
        for $j = i+1, i+2, \ldots \min(i+b, k)$, where $k = \min(m, n)$. do
            Find index $p_j$ of the column with maximum trailing 2-norm.
            Apply permutation $\boldsymbol{P}_j$ swapping column $j$ with $p_j$.
            Update column $j$ with prior reflections in this block.
            Form reflector $\boldsymbol{y}_j$ and $\tau_j$ from new column $j$.
            Compute adjusted reflector inner products $\boldsymbol{w}_j^T$.
            Update row $j$ with all reflections in this block.
            Update trailing column norms by removing the contribution of row $j$.
        end for
        Apply block reflection to trailing matrix.
    end for
    $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{Y}_k \boldsymbol{T}_k \boldsymbol{Y}_k^T$ where $\boldsymbol{T}_k$ can be recovered from $\boldsymbol{Y}_k$ and $\tau_1, \ldots, \tau_k$.
    $\boldsymbol{R} = \boldsymbol{A}^{(k)}$.
    $\boldsymbol{P} = \boldsymbol{P}_1 \boldsymbol{P}_2 \ldots \boldsymbol{P}_k$ is the aggregate column permutation.
end function
```

2.2 Communication Avoiding Rank-Revealing QR

Several mechanisms have been put forward to avoid repeating full passes over the trailing matrix on each iteration. Bischof (Bischof, 1991) proposed pivoting restricted to local blocks and Demmel, et al. (Demmel et al., 2012, 2015) propose a procedure called Communication Avoiding Rank-Revealing QR (CARRQR). CARRQR proceeds by partitioning the trailing matrix into $\mathcal{P}$ subsets of columns that are processed independently and possibly simultaneously. From within each column subset, b candidate pivots are selected using QRCP. Adjacent subsets of candidates are then combined to form $\frac{1}{2}\mathcal{P}$ subsets of $2b$ candidates. This procedure continues, using QRCP to filter b candidates per subset followed by merging results into $\frac{1}{4}\mathcal{P}$ subsets and so on, until only one subset of b candidates remains. The trailing matrix is then updated as before, with blocked reflections.

We now examine several practical constraints in implementing CARRQR. First, the reflectors $\boldsymbol{Y}$, inner products $\boldsymbol{W}^T$, and leading rows of $\boldsymbol{R}$ must be stored separately from the original matrix for each independently processed subset of columns. Furthermore, one must employ a version of QRCP that avoids the trailing update, because the final reflectors are unknown until the last selection stage. Any intermediate changes to the original columns would have to be undone before the final transformations can be correctly processed. In contrast, QRCP can be written to convert columns into reflectors, storing the results in the same array as the input on the strictly lower triangle portion of the matrix. Likewise, $\boldsymbol{R}$ can be stored in place of the input on the upper triangle.

Depending on the initial column partition, CARRQR performs up to 2 times as many inner products as QRCP per block iteration. Note that as the reflector index j increases, the total number of inner products of the form $\boldsymbol{y}_{j+1}^T\boldsymbol{y}_{j+1}$, $\boldsymbol{y}_{j+1}^T\boldsymbol{Y}^{(j)}$ and $\boldsymbol{y}_{j+1}^T\boldsymbol{A}^{(j)}$ remains constant. Therefore, if the ith subset contains n_i columns, bn_i inner products will be required to produce b candidates. Letting $n_1 + n_2 + \cdots + n_{\mathcal{P}} = n$ on the first stage of refinement, summing over all of the subsets gives bn inner products to produce $\mathcal{P}$ subsets of b candidates. QRCP requires the same computational complexity to produce b final pivots. Assuming that the number of candidates is at least halved for each subsequent stage of refinement in CARRQR, it easily follows that no more than $2bn$ inner products will be computed in total.

Despite increased computational complexity, CARRQR is intended to benefit from better memory utilization and better parallel scalability. If each column subset is thin enough to fit in fast memory, then slow communication is eliminated between iterations of j. The only remaining slow communication transmits pivot candidates between stages of refinement.

Unfortunately, writing and tuning CARRQR is nontrivial. We implemented this algorithm and found that it ran slightly slower than the LAPACK implementation of algorithm 4, called DGEQP3, on a shared-memory parallel machine. We believe this was mainly due to inefficient parallelization in the final stages of refinement. We assigned each column subset to a different processor, which then worked independently to produce candidates. This approach was attractive because it did not require communication between processors during each filtration stage. However, despite communication efficiency, this technique can only engage as many processors as there are column subsets. Most processors are left idle during the final stages of refinement. A second problem with this approach occurres when the matrix is too tall. In such cases, it is not possible to select column subsets that are thin enough to fit into fast memory. An efficient implementation would need alternative or additional workload-splitting tactics to use all processors at every stage of refinement.

As we will discuss in the next section, the method we propose also gathers pivots into blocks, which are then applied to the trailing matrix. Our method, however, improves performance by reducing both the communication and computational complexity needed to form a block of pivots.

3. Randomized Projection for Sample Pivoting

Sampling via randomized projection has proven to be beneficial for a variety of applications in numerical linear algebra (Liberty et al., 2007; Woolfe et al., 2008; Rokhlin et al., 2010; Halko et al., 2011; Mahoney et al., 2011; Martinsson et al., 2011). Sampling reduces communication complexity via dimensional reduction, while simultaneously maintaining a safe degree of uncertainty in the approximations that follow. This technique is typically framed using the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984). Let $\boldsymbol{a}_j$ represent the jth column of $\boldsymbol{A}$ for $j = 1, 2, \ldots, n$. There exists a distribution over $\ell \times m$ matrices such that a randomly drawn matrix $\boldsymbol{\Omega}$ yields a lower dimensional embedding $\boldsymbol{b}_j = \boldsymbol{\Omega}\boldsymbol{a}_j$, with high probability of preserving distances within a relative error ε. Because we would like to use the sample to obtain a block of b pivots while controlling the degree of uncertainty in the true norms of $\boldsymbol{A}$, we include padding p in the sample rank so that $\ell = b + p$. When $\boldsymbol{\Omega}$ is GIID, the expected 2-norms and variance are

$$\mathbb{E}_{\boldsymbol{\Omega}}\left[\|\boldsymbol{b}_j\|_2^2\right] = \ell\,\|\boldsymbol{a}_j\|_2^2 \quad\text{and}\quad \mathrm{Var}_{\boldsymbol{\Omega}}\left[\|\boldsymbol{b}_j\|_2^2\right] = 2\ell\,\|\boldsymbol{a}_j\|_2^4 .$$

If we let $\boldsymbol{a}_0 = 0$ and $\boldsymbol{b}_0 = 0$, then we can express the probability of satisfying the relative error bounds as

$$P(\left|\frac{\|\boldsymbol{b}_j - \boldsymbol{b}_i\|_2^2}{\ell\,\|\boldsymbol{a}_j - \boldsymbol{a}_i\|_2^2} - 1\right| \leq \varepsilon) \geq 1 - 2\exp\left(\frac{-\ell\varepsilon^2}{4}(1-\varepsilon)\right),$$

where $0 < \varepsilon < \frac{1}{2}$ and $i, j = 0, 1, \ldots, n$. Capturing the norm of the difference between columns also implies coherence among inner product approximations by taking $(\boldsymbol{b}_1 - \boldsymbol{b}_2)^T(\boldsymbol{b}_1 - \boldsymbol{b}_2) = \|\boldsymbol{b}_1\|_2^2 - 2\boldsymbol{b}_1^T\boldsymbol{b}_2 + \|\boldsymbol{b}_2\|_2^2$. Thus, each component of $\boldsymbol{B}$ within a subspace defined by a few of its columns approximates the corresponding component of $\boldsymbol{A}$.

3.1 Bayesian Analysis

Bayesian inference is not often used in numerical linear algebra. In this case, it allows us to obtain an exact expression for the uncertainty in column norms, which rigorously frames the amount of padding p that is needed to satisfy a relative error bound with a desired probability of success.

We begin by expressing a single column $\boldsymbol{a}$ as its 2-norm multiplied by a unit vector $\boldsymbol{q}$, so that $\boldsymbol{a} = \|\boldsymbol{a}\|_2\,\boldsymbol{q}$. Because a GIID matrix is invariant in distribution under independent orthogonal transformations, we can construct an orthogonal matrix $\boldsymbol{Q} = [\boldsymbol{q}\ \boldsymbol{Q}_\perp]$ and write $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \left[\begin{array}{cc} \hat{\boldsymbol{\omega}} & \hat{\boldsymbol{\Omega}} \end{array}\right] \left[\begin{array}{c} \boldsymbol{q}^T \\ \boldsymbol{Q}_\perp^T \end{array}\right],$$

where both the leading column $\hat{\boldsymbol{\omega}}$ and remaining columns $\hat{\boldsymbol{\Omega}}$ are GIID. It easily follows that each element of the sample column, $\boldsymbol{b} = \|\boldsymbol{a}\|_2\,\hat{\boldsymbol{\omega}}$, is normally distributed with mean 0 and latent variance $\|\boldsymbol{a}\|_2^2$.

Inferring variance from a normal distribution with a known mean is a standard problem in Bayesian statistics. The likelihood probability distribution is written $\mathbf{p}(\boldsymbol{b} \mid \|\boldsymbol{a}\|_2^2) \equiv \mathcal{N}(\boldsymbol{b} \mid 0, \|\boldsymbol{a}\|_2^2\,\boldsymbol{I})$ where $\boldsymbol{I}$ is the $\ell \times \ell$ identity. Jeffreys proposed the maximally uninformative prior for an unknown variance, $\mathbf{p}(\|\boldsymbol{a}\|_2^2) \equiv \|\boldsymbol{a}\|_2^{-2}$, which is invariant under scaling and power transformations (Kass and Wasserman, 1996). We apply Bayes' theorem, $\mathbf{p}(\|\boldsymbol{a}\|_2^2 \mid \boldsymbol{b}) \propto \mathbf{p}(\boldsymbol{b} \mid \|\boldsymbol{a}\|_2^2)\mathbf{p}(\|\boldsymbol{a}\|_2^2)$, to obtain the posterior distribution. Normalization results in the inverse gamma distribution

$$\mathbf{p}(\|\boldsymbol{a}\|_2^2 \mid \|\boldsymbol{b}\|_2^2) \equiv \frac{\left(\frac{\|\boldsymbol{b}\|_2^2}{2}\right)^{\ell/2}}{\Gamma(\frac{\ell}{2})\,\|\boldsymbol{a}\|_2^{-\ell-2}} \exp\left(\frac{-\|\boldsymbol{b}\|_2^2}{2\,\|\boldsymbol{a}\|_2^2}\right).$$

Consequently, we can cast each pivoting decision as the maximizer of expected utility, where utility is taken to be the latent 2-norm squared

$$\mathbb{E}_{\mathbf{p}(\|\boldsymbol{a}\|_2^2|\|\boldsymbol{b}\|_2^2)}\left[\|\boldsymbol{a}\|_2^2\right] = \frac{\|\boldsymbol{b}\|_2^2}{\ell - 2}.$$

Since expected utility is monotonic in the sample column norms, we simply choose the maximum as in QRCP.

More importantly, the posterior distribution clearly relates sample rank ℓ to uncertainty. In order to capture the probability distribution of the latent relative error, we can change variables to express the latent column norm as a fraction ϕ of the expectation, $\|\boldsymbol{a}\|_2^2 = \frac{\phi}{\ell-2}\|\boldsymbol{b}\|_2^2$. This gives analytic expressions for both the probability distribution function and the cumulative distribution function of the relative scaling

$$\mathbf{p}(\phi \mid \ell) \equiv \frac{\left(\frac{\ell-2}{2}\right)^{\ell/2}}{\Gamma(\frac{\ell}{2})\phi^{\ell/2+1}} \exp\left(\frac{-(\ell-2)}{2\phi}\right) \quad \text{and} \quad P(\phi < \tau) = \frac{\Gamma(\frac{\ell}{2}, \frac{\ell-2}{2\tau})}{\Gamma(\frac{\ell}{2})},$$

respectively. The numerator of the cumulative distribution is the upper incomplete gamma function and normalization results in the regularized gamma function.

Rank-revealing decompositions must avoid selecting columns that are already well approximated by components in the span of previous pivots, i.e columns with small trailing norms. As such, we only care about the probability that a sample column radically overestimates the true column norm. The CDF plotted in fig. 3.1 shows the probability that the relative scaling ϕ falls below a specified upper bound τ for several choices of ℓ. Our experiments provide good results using padding $p = 8$ with a block size $b = 32$ so that the effective sample rank satisfies $8 < \ell \leq 40$ for each pivot decision. Note that this analysis also holds for linear combinations of columns in $\boldsymbol{A}$ and $\boldsymbol{B}$, provided that those linear combinations are independent of $\boldsymbol{B}$. An in-depth reliability and probability analysis has appeared in Xiao, Gu, and Langou (Xiao et al., 2017). In particular, they show that in the case of decaying singular values in the matrix $\boldsymbol{A}$, a nearly optimal low-rank approximation can be computed with the QR factorization with a slight modification in the strategy used to choose $\boldsymbol{P}$.

3.2 Sample QRCP Distribution Updates

We now show how the sample matrix $\boldsymbol{B} = \boldsymbol{\Omega}\boldsymbol{A}$ can be used to select a full block of b pivots. If QRCP is performed on the sample matrix, then at iteration j we can examine $\boldsymbol{B}$ as a partial factorization. We let $\boldsymbol{P}^{(j)}$ be the aggregate permutation so far and represent the accumulated orthogonal transformations applied to $\boldsymbol{B}$ as $\boldsymbol{U}^{(j)}$, with corresponding intermediate triangular factor $\boldsymbol{S}^{(j)}$, as shown below. We also consider a partial factorization of $\boldsymbol{A}$ using the same pivots that were applied to $\boldsymbol{B}$. The corresponding factors of $\boldsymbol{A}$ are $\boldsymbol{Q}^{(j)}$ and $\boldsymbol{R}$.

$$\boldsymbol{B}\boldsymbol{P}^{(j)} = \boldsymbol{U}^{(j)}\begin{bmatrix} \boldsymbol{S}_{11}^{(j)} & \boldsymbol{S}_{12}^{(j)} \\ 0 & \boldsymbol{S}_{22}^{(j)} \end{bmatrix} \quad \text{and} \quad \boldsymbol{A}\boldsymbol{P}^{(j)} = \boldsymbol{Q}^{(j)}\begin{bmatrix} \boldsymbol{R}_{11}^{(j)} & \boldsymbol{R}_{12}^{(j)} \\ 0 & \boldsymbol{R}_{22}^{(j)} \end{bmatrix}.$$

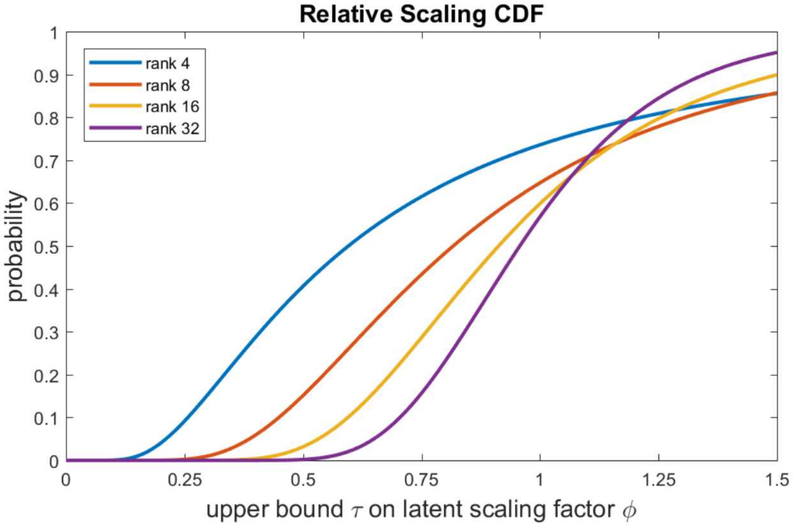

Figure 3.1: Cumulative distribution function for latent relative scaling factor. For sample rank 4, the probability that the latent 2-norm squared is less than 1/8 of the expectation is 0.30%. For sample rank 8, the probability that it is less than 1/4 is 0.23%. For sample rank 32, the probability that it is less than 1/2 is 0.20%.

Both $\boldsymbol{S}_{11}^{(j)}$ and $\boldsymbol{R}_{11}^{(j)}$ are upper triangular. $\boldsymbol{\Omega}$ can then be expressed as elements $\hat{\boldsymbol{\Omega}}$ in the bases given by $\boldsymbol{U}^{(j)}$ and $\boldsymbol{Q}^{(j)}$:

$$\boldsymbol{\Omega} = \boldsymbol{U}^{(j)} \begin{bmatrix} \hat{\boldsymbol{\Omega}}_{11}^{(j)} & \hat{\boldsymbol{\Omega}}_{12}^{(j)} \\ \hat{\boldsymbol{\Omega}}_{21}^{(j)} & \hat{\boldsymbol{\Omega}}_{22}^{(j)} \end{bmatrix} \boldsymbol{Q}^{(j)T}.$$

Noting that $\boldsymbol{B}\boldsymbol{P}^{(j)} = \boldsymbol{\Omega}\boldsymbol{A}\boldsymbol{P}^{(j)}$, we have

$$\begin{bmatrix} \boldsymbol{S}_{11}^{(j)} & \boldsymbol{S}_{12}^{(j)} \\ 0 & \boldsymbol{S}_{22}^{(j)} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\Omega}}_{11}^{(j)}\boldsymbol{R}_{11}^{(j)} & \hat{\boldsymbol{\Omega}}_{11}^{(j)}\boldsymbol{R}_{12}^{(j)} + \hat{\boldsymbol{\Omega}}_{12}^{(j)}\boldsymbol{R}_{22}^{(j)} \\ \hat{\boldsymbol{\Omega}}_{21}^{(j)}\boldsymbol{R}_{11}^{(j)} & \hat{\boldsymbol{\Omega}}_{21}^{(j)}\boldsymbol{R}_{12}^{(j)} + \hat{\boldsymbol{\Omega}}_{22}^{(j)}\boldsymbol{R}_{22}^{(j)} \end{bmatrix}. \tag{3.1}$$

If $\boldsymbol{S}_{11}^{(j)}$ is nonsingular, then both $\hat{\boldsymbol{\Omega}}_{11}^{(j)}$ and $\boldsymbol{R}_{11}^{(j)}$ are also nonsingular. It follows that $\hat{\boldsymbol{\Omega}}_{11}^{(j)} = \boldsymbol{S}_{11}^{(j)}\boldsymbol{R}_{11}^{(j)-1}$ is upper triangular and $\hat{\boldsymbol{\Omega}}_{21}^{(j)} = 0$. In other words, we have implicitly formed a QR factorization of $\boldsymbol{\Omega}\boldsymbol{Q}^{(j)}$ using the same orthogonal matrix $\boldsymbol{U}^{(j)}$. Finally, the trailing matrix in the sample simplifies to $\boldsymbol{S}_{22}^{(j)} = \hat{\boldsymbol{\Omega}}_{22}^{(j)}\boldsymbol{R}_{22}^{(j)}$, which is a sample of the trailing matrix $\boldsymbol{R}_{22}^{(j)}$ using the compression matrix $\hat{\boldsymbol{\Omega}}_{22}^{(j)}$. If the permutation $\boldsymbol{P}^{(j)}$ were independent of the sample, then $\boldsymbol{Q}^{(j)}$ would be formed from $\boldsymbol{A}$, independent of $\boldsymbol{\Omega}$. Likewise, $\boldsymbol{U}^{(j)}$ only depends on the leading j columns of the sample, which are independent of the trailing columns. Thus $\hat{\boldsymbol{\Omega}}_{22}^{(j)}$ would be GIID. As such, we may use column norms of $\boldsymbol{S}_{22}^{(j)}$ to approximate column norms of $\boldsymbol{R}_{22}^{(j)}$ when we select the $(j+1)$st pivot, so that a full block of pivots can be selected without interleaving any references to $\boldsymbol{A}$ or $\boldsymbol{R}$ memory. We note, however, that the permutation may exhibit a subtle dependence on the pivots, which we briefly discuss in section 3.3.

Once b pivots have been selected from the sample matrix $\boldsymbol{B}$, the corresponding columns of $\boldsymbol{A}$ are permuted and processed all at once, as is done in BLAS-3 QR, thus reducing both the communication and computational complexity associated with selecting a block of b pivots by a factor of ℓ/m. More significantly, if the sample matrix $\boldsymbol{B}$ fits in fast memory, then slow communication between consecutive pivot decisions is eliminated within each block iteration. As in BLAS-3 QR, the remaining communication costs are due to the matrix-matrix multiplications needed to perform block reflections. As such, RQRCP satisfies the BLAS-3 performance standard.

Algorithm 5 outlines the full procedure for a sample-based block permutation. It is acceptable for very low-rank approximations, wherein the required sample is small enough to maintain communication efficiency. That is, when the desired approximation rank k is small enough to be a single block, $b = k$. For larger approximations we will resort to a more comprehensive algorithm that includes a sample update formulation that subsumes this version. Since the single-sample algorithm illuminates the performance advantage gained from this approach, we examine it first.

Algorithm 5 Single-Sample Randomized QRCP.

```
Require:
    $\boldsymbol{A}$ is $m \times n$.
    $k$ is the desired approximation rank. $k \ll \min(m, n)$.
Ensure:
    $\boldsymbol{Q}$ is an $m \times m$ orthogonal matrix in the form of $k$ reflectors.
    $\boldsymbol{R}$ is a $k \times n$ truncated upper trapezoidal matrix.
    $\boldsymbol{P}$ is an $n \times n$ permutation matrix such that $\boldsymbol{AP} \approx \boldsymbol{Q}(:, 1 : k)\boldsymbol{R}$.
function $[\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{P}] = \textsc{SingleSampleRQRCP}(\boldsymbol{A}, k)$
    Set sample rank $l = k + p$ as needed for acceptable uncertainty.
    Generate random $l \times m$ matrix $\boldsymbol{\Omega}$.
    Form the sample $\boldsymbol{B} = \boldsymbol{\Omega A}$.
    Get $k$ column pivots from sample $[\cdot, \cdot, \boldsymbol{P}] = \texttt{QRCP}(\boldsymbol{B})$.
    Apply permutation $\boldsymbol{A}^{(1)} = \boldsymbol{AP}$.
    Construct $k$ reflectors from new leading columns $[\boldsymbol{Q}, \boldsymbol{R}_{11}] = \texttt{QR}(\boldsymbol{A}^{(1)}\texttt{(:,1:k)})$.
    Finish $k$ rows of $\boldsymbol{R}$ in remaining columns $\boldsymbol{R}_{12} = \boldsymbol{Q}\texttt{(:,1:k)}^T \boldsymbol{A}^{(1)}\texttt{(:,k+1:n)}$.
end function
```

3.3 Sample Bias

The bias of an estimator is the difference between the expected value of the estimator and the true value of the quantity being estimated. In this case, sample column norms are used to estimate true column norms. As illustrated in the following thought experiment, a subtle form of bias occurs when we use the maximum to select a pivot.

Suppose Jack and Jill roll one six-sided die each. The expected outcome for each of them is 3.5. We then learn that Jill rolled 5, which was greater than Jack's roll. This new information reduces Jack's expectation to 2.5, because it is no longer possible for him to have rolled 5 or 6. Similarly, the act of selecting the largest sample norm creates a dependency with remaining samples by truncating their plausible outcomes.

This effect becomes less pronounced, however, if the decision is more likely to be determined by the true value being estimated rather than a chance sample outcome. For example, now suppose that Jill rolls three dice and Jack rolls one. If we only know each person's sum, we can still infer the number of dice each person rolled and select the expected maximum as before. In 90.7% of trials, Jill's sum will be 7 or greater, driven by the fact that three dice were rolled. In most cases, this leaves Jack's potential outcomes unconstrained and his expectation unbiased after we observe the maximum. Analogously, when a selected column exhibits a norm that is an order of magnitude greater than others, the bias effect is negligible and we may proceed as though the pivot decision was independent of the sample.

The original version of this paper included analysis of this distribution truncation effect (Duersch and Gu, 2017). For our purposes, we proceed as though the progression of sample updates is independent of the pivot decisions, which is also the approach Xiao, et al. (Xiao et al., 2017) take. This assumption is potentially problematic when multiple trailing columns have similar norms, but any such column provides a suitable pivot in that scenario.

3.4 Blocked Sample Updates

The sample matrix in algorithm 5 is formulated to have rank $\ell = k + p$, where k was both the desired approximation rank and the permutation block size, b. As k increases, however, it becomes inefficient to simply increase the sample rank ℓ. In the extreme case, a full decomposition would require a sample just as big as the original matrix. If we require a decomposition with a larger rank than that which can be efficiently sampled and blocked, that is if the sample rank cannot exceed $\ell = b + p$, but we require $k > b$, then we need to update the sample matrix after each block. Martinsson (Martinsson and Voronin, 2016) developed an approach in which one simply processes each subsequent block by drawing a new random matrix $\boldsymbol{\Omega}$ and applying it to each new trailing matrix. We propose a sample update formulation that does not require multiplying the trailing matrix by a new compression matrix and reduces BLAS-3 communication in the overall factorization by at least one third.

The update formula we derive is an extension of the implicit update mechanism described in the previous section. Both algorithm 6 and algorithm 7, the truncated variation, will proceed in blocks of pivots. Bracket superscripts denote the results of a computation that occurred on the indicated block-iteration. At entry to the first block-iteration, the sample is $\boldsymbol{B}^{[0]} = \boldsymbol{\Omega}^{[0]}\boldsymbol{A}^{[0]}$, where $\boldsymbol{A}^{[0]}$ is the original matrix. At the end of block-iteration J, the sample will be in the transformed state

$$\begin{bmatrix} \boldsymbol{S}_{11}^{[J]} & \boldsymbol{S}_{12}^{[J]} \\ 0 & \boldsymbol{S}_{22}^{[J]} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\Omega}}_{11}^{[J]} & \hat{\boldsymbol{\Omega}}_{12}^{[J]} \\ 0 & \hat{\boldsymbol{\Omega}}_{22}^{[J]} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{11}^{[J]} & \boldsymbol{R}_{12}^{[J]} \\ 0 & \boldsymbol{A}^{[J]} \end{bmatrix}, \tag{3.2}$$

just as in eq. (3.1). $\boldsymbol{S}_{11}^{[J]}$ is the leading upper triangle from the partial factorization of the sample and $\boldsymbol{S}_{22}^{[J]}$ gives the trailing sample columns. Likewise $\boldsymbol{R}_{11}^{[J]}$ and $\boldsymbol{A}^{[J]}$, respectively, give the leading upper triangle and trailing columns that would be obtained by factorizing the original matrix with the same pivots. By absorbing the transformations $\boldsymbol{U}^{[J]T}$ and

$\boldsymbol{Q}^{[J]}$ into $\boldsymbol{\Omega}^{[J]}$, we obtained an effective compression matrix $\hat{\boldsymbol{\Omega}}_{22}^{[J]}$, which had already been implicitly applied to the trailing columns: $\boldsymbol{S}_{22}^{[J]} = \hat{\boldsymbol{\Omega}}_{22}^{[J]}\boldsymbol{A}^{[J]}$. The difficulty is $\boldsymbol{S}_{22}^{[J]}$ only has rank p. In order to construct a rank $\ell = b + p$ sample of the trailing matrix $\boldsymbol{A}^{[J]}$, we need to include $\hat{\boldsymbol{\Omega}}_{12}^{[J]}$ in the updated compression matrix

$$\boldsymbol{\Omega}^{[J]} = \begin{bmatrix} \hat{\boldsymbol{\Omega}}_{12}^{[J]} \\ \hat{\boldsymbol{\Omega}}_{22}^{[J]} \end{bmatrix} \quad \text{giving} \quad \boldsymbol{B}^{[J]} = \boldsymbol{\Omega}^{[J]}\boldsymbol{A}^{[J]} = \begin{bmatrix} \boldsymbol{S}_{12}^{[J]} - \hat{\boldsymbol{\Omega}}_{11}^{[J]}\boldsymbol{R}_{12}^{[J]} \\ \boldsymbol{S}_{22}^{[J]} \end{bmatrix}.$$

In other words, the new compression matrix $\boldsymbol{\Omega}^{[J]}$ is simply $\boldsymbol{U}^{[J]T}\boldsymbol{\Omega}^{[J-1]}\boldsymbol{Q}^{[J]}$, with the leading b columns removed. This new compression matrix does not need to be explicitly formed or applied to the trailing columns $\boldsymbol{A}^{[J]}$. Instead, we form the result implicitly by removing $\hat{\boldsymbol{\Omega}}_{11}^{[J]}\boldsymbol{R}_{12}^{[J]}$ from $\boldsymbol{S}_{12}^{[J]}$. Both $\boldsymbol{R}_{11}^{[J]}$ and $\boldsymbol{R}_{12}^{[J]}$ will be computed in blocked matrix multiply operations using the previous b pivots of $\boldsymbol{A}$. Since $\hat{\boldsymbol{\Omega}}_{11}^{[J]}$ can then be recovered from $\boldsymbol{S}_{11}^{[J]}$, we can avoid any direct computations on $\boldsymbol{\Omega}$. We only need to update the first b rows of $\boldsymbol{B}$, which gives us the sample update formula

$$\begin{bmatrix} \boldsymbol{B}_1^{[J]} \\ \boldsymbol{B}_2^{[J]} \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_{12}^{[J]} - \boldsymbol{S}_{11}^{[J]}\boldsymbol{R}_{11}^{[J]-1}\boldsymbol{R}_{12}^{[J]} \\ \boldsymbol{S}_{22}^{[J]} \end{bmatrix}. \tag{3.3}$$

Full RQRCP, described in algorithm 6, can be structured as a modification to blocked BLAS-3 QR. The algorithm must simply interleave processing blocks of reflectors with permutations obtained from each sample matrix, then update the sample as above. To obtain a modified version that employs repeated sampling simply replace the sample update with a new sample of the trailing matrix, $\boldsymbol{B}^{[J]} = \boldsymbol{\Omega}^{[J]}\boldsymbol{A}^{[J]}$.

When QRCP is applied to the sample matrix $\boldsymbol{B}$, only a partial decomposition is necessary. The second argument b in the subroutine call QRCP($\boldsymbol{B}^{[J]}, b$) indicates that only b column permutations are required. It is relatively simple to modify the QR algorithm to avoid any unnecessary computation. After sample pivots have been applied to the array containing both $\boldsymbol{A}$ and $\boldsymbol{R}$, we perform QR factorization on the new leading b columns of the trailing matrix. Although this is stated as returning $\boldsymbol{Q}^{[J]}$ for convenience, it can be implemented efficiently with the blocked Householder reflections described in section 2.1. We can then apply reflectors to the trailing matrix and form the sample update $\boldsymbol{B}^{[J]}$ to prepare for the next iteration.

4. Truncated Factorizations for Low-Rank Approximations

The trailing matrix is unnecessary for low-rank applications. We can reformulate RQRCP to avoid the trailing update, rather than computing it and discarding it. Provided the approximation rank is small ($k \ll \min(m, n)$), the truncated reformulation (TRQRCP) reduces large matrix multiplications by half and completes in roughly half the time.

4.1 Truncated RQRCP

Our technique is analogous to the method Quintana-Ortí, Sun, and Bischof used to halve BLAS-2 operations in QRCP. In their version of QRCP, all reflector inner products are

Algorithm 6 Randomized QR with Column Pivoting, RQRCP

```
Require:
    $\boldsymbol{A}$ is $m \times n$.
    $k$ is the desired factorization rank. $k \leq \min(m, n)$.
Ensure:
    $\boldsymbol{Q}$ is an $m \times m$ orthogonal matrix in the form of $k$ reflectors.
    $\boldsymbol{R}$ is a $k \times n$ upper trapezoidal (or triangular) matrix.
    $\boldsymbol{P}$ is an $n \times n$ permutation matrix such that $\boldsymbol{AP} \approx \boldsymbol{Q}(\texttt{:,1:k})\boldsymbol{R}$.
function $[\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{P}] = \mathrm{RQRCP}(\boldsymbol{A}, k)$
    Set sample rank $\ell = b + p$ as needed for acceptable uncertainty.
    Generate random $\ell \times m$ matrix $\boldsymbol{\Omega}^{[0]}$.
    Form the initial sample $\boldsymbol{B}^{[0]} = \boldsymbol{\Omega}^{[0]}\boldsymbol{A}^{[0]}$.
    for $J = 1, 2, \ldots, \frac{k}{b}$ do
        Get $b$ column pivots from sample $[\boldsymbol{U}^{[J]}, \boldsymbol{S}^{[J]}, \boldsymbol{P}^{[J]}] = \texttt{QRCP}(\boldsymbol{B}^{[J-1]}, b)$.
        Permute $\boldsymbol{A}^{[J-1]}$ and completed rows in $\boldsymbol{R}$ with $\boldsymbol{P}^{[J]}$.
        Construct $b$ reflectors $[\boldsymbol{Q}^{[J]}, \boldsymbol{R}_{11}^{[J]}] = \texttt{QR}(\boldsymbol{A}^{[J-1]}(\texttt{:,1:b}))$.
        Finish $b$ rows $\boldsymbol{R}_{12}^{[J]} = \boldsymbol{Q}^{[J]}(\texttt{:,1:b})^T\boldsymbol{A}^{[J-1]}(\texttt{:,b+1:end})$.
        Update the trailing matrix $\boldsymbol{A}^{[J]} = \boldsymbol{Q}^{[J]}(\texttt{:,b+1:end})^T\boldsymbol{A}^{[J-1]}(\texttt{:,b+1:end})$.
        Update the sample $\boldsymbol{B}_1^{[J]} = \boldsymbol{S}_{12}^{[J]} - \boldsymbol{S}_{11}^{[J]}\boldsymbol{R}_{11}^{[J]-1}\boldsymbol{R}_{12}^{[J]}$ and $\boldsymbol{B}_2^{[J]} = \boldsymbol{S}_{22}^{[J]}$.
    end for
    $\boldsymbol{Q} = \boldsymbol{Q}^{[1]} \begin{bmatrix} \boldsymbol{I}_b & \\ & \boldsymbol{Q}^{[2]} \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{I}_{([k/b]-1)\,b} & \\ & \boldsymbol{Q}^{[k/b]} \end{bmatrix}$.
    $\boldsymbol{P} = \boldsymbol{P}^{[1]} \begin{bmatrix} \boldsymbol{I}_b & \\ & \boldsymbol{P}^{[2]} \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{I}_{([k/b]-1)\,b} & \\ & \boldsymbol{P}^{[k/b]} \end{bmatrix}$.
end function
```

computed, but rows and columns are only updated as needed. In order to compute correct reflector inner products, without having updated the trailing matrix, we need to formulate blocked reflector compositions

$$(\boldsymbol{I} - \boldsymbol{Y}_1\boldsymbol{T}_1\boldsymbol{Y}_1^T)(\boldsymbol{I} - \boldsymbol{Y}_2\boldsymbol{T}_2\boldsymbol{Y}_2^T) = \boldsymbol{I} - \boldsymbol{YTY}^T,$$

$$\text{where} \quad \boldsymbol{Y} = \begin{bmatrix}\boldsymbol{Y}_1 & \boldsymbol{Y}_2\end{bmatrix} \quad \text{and} \quad \boldsymbol{T} = \begin{bmatrix}\boldsymbol{T}_1 & -\boldsymbol{T}_1\boldsymbol{Y}_1^T\boldsymbol{Y}_2\boldsymbol{T}_2 \\ 0 & \boldsymbol{T}_2\end{bmatrix}.$$

The corresponding reflector inner products $\boldsymbol{W}^T = \boldsymbol{T}^T\boldsymbol{Y}^T\boldsymbol{A}$ become

$$\boldsymbol{W}^T = \begin{bmatrix}\boldsymbol{W}_1^T \\ \boldsymbol{W}_2^T\end{bmatrix} \quad \text{with} \quad \boldsymbol{W}_1^T = \boldsymbol{T}_1^T\boldsymbol{Y}_1^T\boldsymbol{A} \quad \text{and} \quad \boldsymbol{W}_2^T = \boldsymbol{T}_2^T\left(\boldsymbol{Y}_2^T\boldsymbol{A} - (\boldsymbol{Y}_2^T\boldsymbol{Y}_1)\boldsymbol{W}_1^T\right).$$

If we store these reflector inner products, then we can construct any submatrix of the accumulated transformation $\hat{\boldsymbol{A}}^{[J]} = \boldsymbol{A} - \boldsymbol{Y}^{[J]}\boldsymbol{W}^{[J]T}$ as needed. Columns that are selected by sample pivots are constructed just before becoming the next reflectors and corresponding rows of $\boldsymbol{R}$ are constructed just before being used to update the sample, as outlined in algorithm 7.

Algorithm 7 Truncated RQRCP without trailing update

```
Require:
    $\boldsymbol{A}$ is $m \times n$.
    $k$ is the approximation rank. $k \ll \min(m, n)$.
Ensure:
    $\boldsymbol{Q}$ is an $m \times m$ orthogonal matrix in the form of $k$ reflectors.
    $\boldsymbol{R}$ is a $k \times n$ upper trapezoidal matrix.
    $\boldsymbol{P}$ is an $n \times n$ permutation matrix such that $\boldsymbol{AP} \approx \boldsymbol{Q}(\texttt{:,1:k})\boldsymbol{R}$.
function $[\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{P}]$ TruncatedRQRCP$(\boldsymbol{A}, k)$
    Set the sample rank $\ell = b + p$ as needed for acceptable uncertainty.
    Generate $\ell \times m$ random matrix $\boldsymbol{\Omega}^{[0]}$ and sample $\boldsymbol{B}^{[0]} = \boldsymbol{\Omega}^{[0]}\boldsymbol{A}[0]$.
    for $J = 1, 2, \ldots, \frac{k}{b}$ do
        Obtain $b$ pivots $[\boldsymbol{U}^{[J]}, \boldsymbol{S}^{[J]}, \boldsymbol{P}^{[J]}] = \texttt{QRCP}(\boldsymbol{B}^{[J]}, b)$.
        Permute $\boldsymbol{A}^{[J]} = \boldsymbol{A}^{[J-1]}\boldsymbol{P}^{[J]}$, $\boldsymbol{W}_1^{[J]T} = \boldsymbol{W}^{[J-1]T}\boldsymbol{P}^{[J]}$, and leading $\boldsymbol{R}$ rows.
        Construct selected columns $\hat{\boldsymbol{A}}_J$ from $\boldsymbol{A}^{[J]} - \boldsymbol{Y}^{[J-1]}\boldsymbol{W}_1^{[J]T}$.
        Form reflectors $\boldsymbol{Y}_2^{[J]}$ using $[\boldsymbol{Q}^{[J]}, \boldsymbol{R}_{11}^{[J]}] = \texttt{QR}(\hat{\boldsymbol{A}}_J)$.
        Form inner products $\boldsymbol{W}_2^{[J]} = \boldsymbol{T}_2^{[J]T}(\boldsymbol{Y}_2^{[J]T}\boldsymbol{A}^{[J]} - (\boldsymbol{Y}_2^{[J]T}\boldsymbol{Y}^{[J-1]})\boldsymbol{W}_1^{[J]T})$.
        Augment $\boldsymbol{Y}^{[J]} = [\boldsymbol{Y}^{[J-1]} \; \boldsymbol{Y}_2^{[J]}]$ and $\boldsymbol{W}^{[J]} = [\boldsymbol{W}_1^{[J]} \; \boldsymbol{W}_2^{[J]}]$.
        Construct new rows of $\boldsymbol{R}$ from $\boldsymbol{A}^{[J]} - \boldsymbol{Y}^{[J]}\boldsymbol{W}^{[J]T}$.
        Update the sample $\boldsymbol{B}_1^{[J]} = \boldsymbol{S}_{12}^{[J]} - \boldsymbol{S}_{11}^{[J]}\boldsymbol{R}_{11}^{[J]-1}\boldsymbol{R}_{12}^{[J]}$ and $\boldsymbol{B}_2^{[J]} = \boldsymbol{S}_{22}^{[J]}$.
    end for
    $\boldsymbol{Q} = \boldsymbol{Q}^{[1]} \begin{bmatrix} \boldsymbol{I}_b & \\ & \boldsymbol{Q}^{[2]} \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{I}_{([k/b]-1)\,b} & \\ & \boldsymbol{Q}^{[k/b]} \end{bmatrix}$.
    $\boldsymbol{P} = \boldsymbol{P}^{[1]} \begin{bmatrix} \boldsymbol{I}_b & \\ & \boldsymbol{P}^{[2]} \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{I}_{([k/b]-1)\,b} & \\ & \boldsymbol{P}^{[k/b]} \end{bmatrix}$.
end function
```

4.2 Truncated SVD Approximation

TRQRCP naturally extends to an approximation of the truncated SVD by following the QLP method proposed by Stewart. The QLP decomposition proceeds by first applying QRCP to obtain $\boldsymbol{AP}_0 = \boldsymbol{Q}_0\boldsymbol{R}$. Then the right triangular matrix $\boldsymbol{R}$ is factored again using an LQ factorization $\boldsymbol{P}_1\boldsymbol{R} = \boldsymbol{LQ}_1$, where row-pivoting is an optional safeguard (otherwise $\boldsymbol{P}_1 = \boldsymbol{I}$), giving the factorization $\boldsymbol{A} = (\boldsymbol{Q}_0\boldsymbol{P}_1^T)\boldsymbol{L}(\boldsymbol{Q}_1\boldsymbol{P}_0^T)$. The diagonal elements of $\boldsymbol{L}$ approximate the singular values of $\boldsymbol{A}$. Huckaby and Chan (Huckaby and Chan, 2003) provide convergence analysis.

The approximate truncated SVD we propose (TUXV) simply adapts low-rank versions of the steps in QLP. The rank-k approximation that results is exactly the same as the truncated approximation that would be obtained if QLP had been processed to completion using RQRCP, without secondary row-pivoting, and then truncated to a rank-k approximation.

We begin by using TRQRCP to produce k left reflectors, which defines the initial left orthogonal matrix $\boldsymbol{U}^{(0)}$. Superscript (0) refers to the initial state of an array upon entry to the first iteration of the main loop. We can compare our results to what would have been obtained from full RQRCP-based QLP:

$$\boldsymbol{A}\boldsymbol{P}^{(0)} \approx \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0)}\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_{11}^{(0)} & \boldsymbol{R}_{12}^{(0)} \\ 0 & 0\end{bmatrix} \quad \text{vs.} \quad \boldsymbol{A}\boldsymbol{P}^{(0*)} = \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0*)}\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_{11}^{(0)} & \boldsymbol{R}_{12}^{(0*)} \\ 0 & \boldsymbol{R}_{22}^{(0*)}\end{bmatrix}.$$

The asterisk denotes additional pivoting produced from the full factorization. The first k pivots in $\boldsymbol{P}^{(0)}$ and corresponding reflectors in $\boldsymbol{U}^{(0)}$ are the same, as are corresponding rows in $\boldsymbol{R}^{(0)}$ modulo additional column permutations. We reverse these permutations to construct the $k \times n$ matrix $\boldsymbol{Z}^{(0)} = \boldsymbol{R}^{(0)}\boldsymbol{P}^{(0)T}$ so that

$$\boldsymbol{A} \approx \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0)}\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}_{11}^{(0)} & \boldsymbol{Z}_{12}^{(0)} \\ 0 & 0\end{bmatrix} \quad \text{versus} \quad \boldsymbol{A} = \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0*)}\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}_{11}^{(0)} & \boldsymbol{Z}_{12}^{(0)} \\ \boldsymbol{Z}_{21}^{(0*)} & \boldsymbol{Z}_{22}^{(0*)}\end{bmatrix}.$$

Taking the LQ factorization from $\boldsymbol{Z}^{(0)}$, so that $\boldsymbol{L}^{(1)}\boldsymbol{V}^{(1)T} = \boldsymbol{Z}^{(0)}$, instead of from $\boldsymbol{R}^{(0)}$ simply absorbs the permutation $\boldsymbol{P}^{(0)T}$ into the definition of $\boldsymbol{V}^{(1)T}$ so that

$$\begin{aligned} & \boldsymbol{A} \approx \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0)}\end{bmatrix}\begin{bmatrix}\boldsymbol{L}_{11}^{(1)} & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_1^{(1)T} \\ \boldsymbol{V}_2^{(1)T}\end{bmatrix} \\ \text{versus} \quad & \boldsymbol{A} = \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0*)}\end{bmatrix}\begin{bmatrix}\boldsymbol{L}_{11}^{(1)} & 0 \\ \boldsymbol{L}_{21}^{(1*)} & \boldsymbol{L}_{22}^{(1*)}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_1^{(1)T} \\ \boldsymbol{V}_2^{(1*)T}\end{bmatrix}. \end{aligned}$$

For consistency with the following algorithm, we label the $k \times k$ connecting matrix $\boldsymbol{L}_{11}^{(1)}$ as $\boldsymbol{X}^{(0)}$. We will discuss the connecting matrix further after explaining the rest of the algorithm. Provided that no secondary row-pivoting is considered, the leading k reflectors in $\boldsymbol{V}^{(1)}$ and $\boldsymbol{V}^{(1*)}$ are identical because they are only computed from the leading k rows of $\boldsymbol{Z}^{(0)}$. At this point, the rank-k approximation of RQRCP-based QLP would require $\boldsymbol{L}_{21}^{(1*)}$, which is unknown. Fortunately, the leading k columns of $\boldsymbol{U}^{(0*)}\boldsymbol{L}^{(1*)}$ can be reconstructed with one matrix multiply. We label this $m \times k$ matrix $\boldsymbol{Z}^{(1)}$, which is

$$\boldsymbol{Z}^{(1)} = \boldsymbol{A}\boldsymbol{V}_1^{(1)} = \begin{bmatrix}\boldsymbol{U}_1^{(0)} & \boldsymbol{U}_2^{(0*)}\end{bmatrix}\begin{bmatrix}\boldsymbol{L}_{11}^{(1)} \\ \boldsymbol{L}_{21}^{(1*)}\end{bmatrix}$$

in both cases. We can then take the QR-factorization $\boldsymbol{U}^{(1)}\boldsymbol{X}^{(1)} = \boldsymbol{Z}^{(1)}$ to produce the approximation

$$\boldsymbol{A} \approx \boldsymbol{U}^{(1)}\begin{bmatrix}\boldsymbol{X}^{(1)} & 0 \\ 0 & 0\end{bmatrix}\boldsymbol{V}^{(1)^T}. \tag{3.4}$$

Further iterations can be computed to produce subsequent $k \times k$ connection matrices $\boldsymbol{X}^{(2)}$, $\boldsymbol{X}^{(3)}$, etc. which would flip-flop between upper triangular and lower triangular forms. To do this, we simply multiply the leading rows of $\boldsymbol{U}^T$ or columns of $\boldsymbol{V}$ on the left and right of $\boldsymbol{A}$, respectively, as outlined in algorithm 8. The leading singular values of $\boldsymbol{A}$ are approximated on the diagonals of $\boldsymbol{X}^{(j)}$. Since the connection matrix is small, however, it is feasible to obtain slightly better approximations by taking the SVD of $\boldsymbol{X}^{(j)}$. One could also insert mechanisms to iterate until a desired level of convergence is obtained, but Stewart observed that only one QRCP-LQ iteration is needed to produce a reasonable approximation of the SVD. One subtle point of possible confusion is that by setting $j_{\max} = 1$ our algorithm might appear to produce a truncated approximation from the sequence RQRCP-LQ-QR. It is true that the diagonal elements in $\boldsymbol{X}$ correspond to that sequence; however, the resulting factorization is equivalent to that which would be obtained by keeping only the leading columns of $\boldsymbol{L}$ after RQRCP-LQ. The final QR factorization simply extracts an orthogonal basis $\boldsymbol{U}$. In the next section, we test the performance of TUXV with $j_{\max} = 1$ for both timing and quality experiments.

Algorithm 8 TUXV approximation of the truncated SVD

```
Require:
    $\boldsymbol{A}$ is an $m \times n$ matrix to approximate.
    $k$ is the approximation rank. $k \ll \min(m, n)$.
    $j_{\max}$ is the number of LQ-QR iterations. We set $j_{\max} = 1$.
Ensure:
    $\boldsymbol{U}$ is an orthogonal $m \times m$ matrix.
    $\boldsymbol{V}$ is an orthogonal $n \times n$ matrix.
    $\boldsymbol{X}$ is a $k \times k$ upper or lower triangular matrix.
    $\boldsymbol{A} \approx \boldsymbol{U}(:, 1:k)\boldsymbol{X}\boldsymbol{V}(:, 1:k)^T$.
function $[\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{V}] = \mathrm{TUXV}(\boldsymbol{A}, k, \tau, j_{\max})$
    TRQRCP-Factorize $[\boldsymbol{U}^{(0)}, \boldsymbol{R}^{(0)}, \boldsymbol{P}^{(0)}] = \texttt{TRQRCP}(\boldsymbol{A}, k)$.
    Restore original column order $\boldsymbol{Z}^{(0)} = \boldsymbol{R}^{(0)}\boldsymbol{P}^{(0)T}$.
    LQ-Factorize $[\boldsymbol{V}^{(1)}, \boldsymbol{X}^{(0)T}] = \texttt{QR}(\boldsymbol{Z}^{(0)T})$.
    for $j = 1, 3, 5, \ldots$ do
        $\boldsymbol{Z}^{(j)} = \boldsymbol{A}\boldsymbol{V}^{(j)}(\texttt{:,1:k})$.
        QR-Factorize $[\boldsymbol{U}^{(j+1)}, \boldsymbol{X}^{(j)}] = \texttt{QR}(\boldsymbol{Z}^{(j)})$.
        If $j = j_{\max}$, then break.
        $\boldsymbol{Z}^{(j+1)} = \boldsymbol{U}^{(j+1)}(\texttt{:,1:k})^T\boldsymbol{A}$.
        LQ-Factorize $[\boldsymbol{V}^{(j+2)}, \boldsymbol{X}^{(j+1)T}] = \texttt{QR}(\boldsymbol{Z}^{(j+1)T})$.
        If $j + 1 = j_{\max}$, then break.
    end for
end function
```

5. Experiments

Our first Fortran version of RQRCP used simple calls to BLAS and LAPACK subroutines without directly managing workloads among available cores. Library implementations of BLAS and LAPACK subroutines automatically distribute the computation to available cores using OpenMP. Although we knew RQRCP should have nearly the same communication complexity as blocked QR, that version did not compete well with library calls to the LAPACK subroutine DGEQRF, the BLAS-3 QR factorization. We believe this was due to poor automatic memory coordination between large blocked matrix operations. In order to provide a convincing demonstration of the efficiency of RQRCP, we had to carefully manage workloads using OpenMP within each phase of the main algorithm. The following experiments show that our RQRCP and TRQRCP subroutines can be written to require substantially less computation time than the optimized QRCP implementation DGEQP3 available through Intel's Math Kernel Library.

5.1 Full Factorization Time

These tests examine how factorization times scale with various problem dimensions for several full matrix decompositions. Since we wanted to understand how different sizes and shapes of matrices could affect performance, we separately tested order scaling, row scaling, and column scaling. In order scaling, we vary the number of rows and columns together so that the matrix remains square. We also wanted to see how the performance of each algorithm scales as we increase the number of cores engaged. Unless the experiment specifies otherwise, each matrix has 12000 rows, 12000 columns, and the algorithm engages 24 cores. The algorithms we tested are listed in section 5.1. fig. 3.2 shows performance results.

Table 3.1: Full decomposition experiments compare these algorithms. Rank-revealing subroutines are DGESVD, DGEQP3, RSRQRCP, and RQRCP. DGEQRF demonstrates the limit of performance without pivoting. DGEQR2 to shows the historical evolution of these algorithms.

Subroutine	Description
`DGEQR2`	LAPACK BLAS-2 implementation of QR
`DGESVD`	LAPACK singular value decomposition
`DGEQP3`	LAPACK competing implementation of QRCP
`RSRQRCP`	algorithm 6 modified for repeated-sampling
`RQRCP`	algorithm 6 with sample update (unmodified)
`DGEQRF`	LAPACK BLAS-3 implementation of QR

These tests show that RQRCP performs nearly as well as DGEQRF, the LAPACK implementation of BLAS-3 QR without pivoting. We further note that our implementation even outperforms DGEQRF in some column scaling cases. These experiments were run on a single node of the NERSC machine Edison. Each node has two 12-core Intel processors. Subroutines were linked with Intel's Math Kernel Library. Each test matrix was randomly generated and the same matrix was submitted to each algorithm.

5.2 Truncated Decomposition

These experiments compare truncated approximation performance. In addition to probing how the matrix shape affects execution time, we examine the effect of varying the truncation rank. As before, we test parallelization by varying the number of cores engaged. Unless specified otherwise, each algorithm engages 24 cores, operates on a 12000 × 12000 matrix, and truncates to rank 1200. Since the proprietary optimized implementations of LAPACK functions were unavailable for modification, we rewrote and adjusted each algorithm to halt at the desired rank. Section 5.2 lists the algorithms we tested. Again, the same random matrix is submitted to each algorithm. Figure 3.3 shows results.

Since TRQRCP uses the sample update formula and avoids the trailing update, it is nearly always fastest. Our TUXV experiments use $j_{\max} = 1$. As such, TUXV performs just one additional matrix multiply. These results show that TUXV requires only a modest

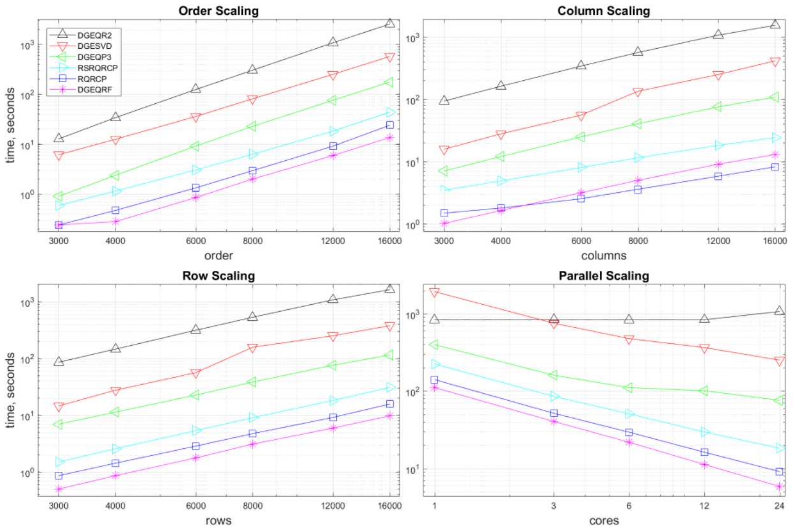

Figure 3.2: Full decomposition benchmarks. Top-left: 24 cores, rows and columns scaled equally. Top-right: 24 cores, 12000 rows, columns scaled. Bottom-left: 24 cores, rows scaled, 12000 columns. Bottom-right: cores scaled, 12000 rows and columns. RQRCP consistently performs almost as well as DGEQRF, QR without pivoting.

Table 3.2: These algorithms are compared in truncated decomposition scaling experiments. Comparing RSRQRCP with RQRCP reveals the cost of repeated sampling. Comparing RQRCP with QR reveals the cost of pivot selection from samples. Comparing RQRCP with TRQRCP reveals the cost of the trailing matrix update. This version of QR is identical to RQRCP after eliminating sample operations and pivoting.

Subroutine	Description
`QRCP`	algorithm 4, QRCP with trailing update
`RSRQRCP`	algorithm 6 modified for repeated-sampling
`TUXV`	algorithm 8, approximation of truncated SVD
`RQRCP`	algorithm 6 with sample update and trailing update (unmodified)
`QR`	QR (no pivoting) with trailing update
`TRQRCP`	algorithm 7 with sample update and no trailing update

increase in processing time over optimized truncated QR. Furthermore, the next set of experiments shows that TUXV gains a significant improvement in approximation quality over both QRCP and RQRCP.

5.3 Decomposition quality

The pivots that result from randomized sampling are not the same as those obtained from QRCP. In order to compare factorization quality, we construct sequences of partial factor-

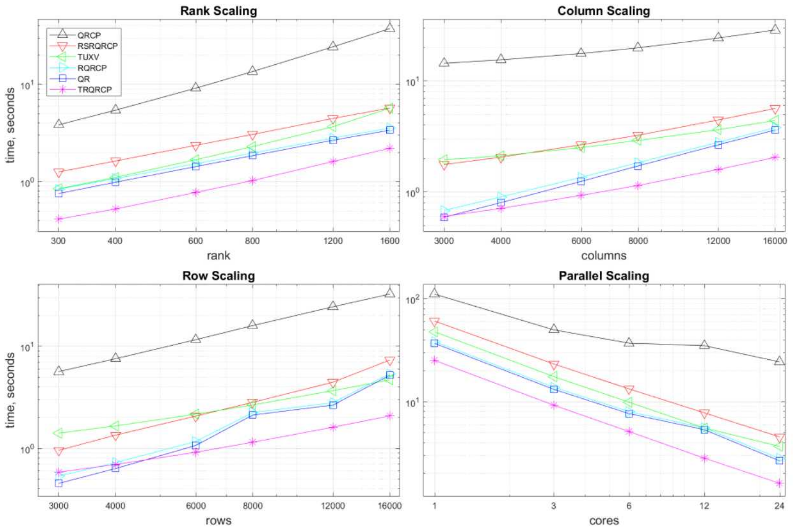

Figure 3.3: Truncated decomposition benchmarks. Top-left: 24 cores, 12000 rows and columns, truncation rank scaled. Top-right: 24 cores, 12000 rows, columns scaled, rank 1200. Bottom-left: 24 cores, rows scaled, 12000 columns, rank 1200. Bottom-right: cores scaled, 12000 rows and columns, rank 1200. RQRCP and QR perform similarly. TRQRCP is fastest. TUXV requires only modest additional cost.

izations and compute the corresponding low rank approximations. The resulting relative error in the Frobenius norm is plotted in fig. 3.4 against the corresponding approximation rank. For both RQRCP and TUXV, we perform 100 runs and plot the median, as well as both the minimum and maximum error bounds. Random samples used padding size $p = 8$ and block size $b = 32$. Matrix decomposition quality is compared for the proposed algorithms using test cases from the San Jose State University Singular Matrix Database: `FIDAP/ex33`, `HB/lock2232`, and `LPnetlib/lpi_gran`. We also test a matrix corresponding to a gray-scale image of a differential gear (image credit: Alex Kovach (Kovach, 2016)). Plot axes have been chosen to magnify the differences among these algorithms.

At the top of each plot we have QR without pivoting. In order to produce competitive results, QR was applied after presorting columns in order of descending 2-norms. Despite this modification, QR performs poorly (as expected) with approximation error dropping much more slowly than the other approximations, thus demonstrating why pivoting is necessary to prioritize representative columns in the truncated decomposition. Below QR we have RQRCP, which achieves results that are consistent with QRCP in each case. The QRCP-like low-rank approximations are further improved by TUXV and the exact truncated SVD, which is theoretically optimal. In each case, TUXV produces approximation error closer to the SVD, with very low variation among the 100 runs.

In fig. 3.5, we compare approximation quality by reconstructing the image of the differential gear using low-rank approximations. The original image is 2442 × 3888 and we truncate to rank 80. Again, truncated QR shows the poorest reconstruction quality, de-

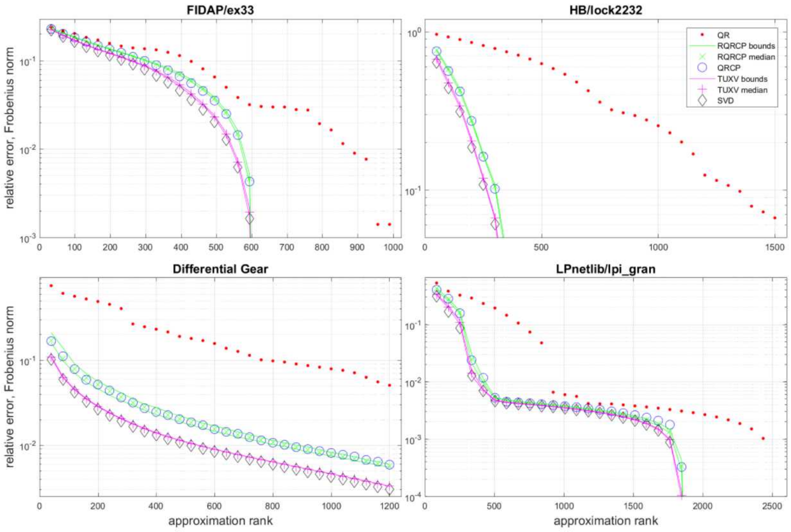

Figure 3.4: Truncated decomposition quality experiments. Top-left: `FIDAP/ex33` (1733 × 1733). Top-right: `HB/lock2232` (2232 × 2232). Bottom-left: Differential gear (2442 × 3888). Bottom-right: `LPnetlib/lpi_gran` (2658 × 2525). Both RQRPC and TUXV show the minimum and maximum truncation errors over 100 runs. The range of outcomes for RQRCP is narrow and holds to QRCP. TUXV outcomes are even narrower and slightly above the truncated SVD.

spite presorting. Truncated RQRCP produces better results, but close inspection shows fine defects. Reconstruction using TUXV is nearly indistinguishable from the original.

6. Conclusion

We have shown that RQRCP achieves strong parallel scalability and the pivoting quality of QRCP at BLAS-3 performance, often an order of magnitude faster than the standard approach. By using randomized projection to construct a small sample matrix $\boldsymbol{B}$ from a much larger original matrix $\boldsymbol{A}$, it becomes possible to substantially reduce the communication complexity associated with a series of algorithmic decisions. Our analysis of latent column norms, inferred from the sample, justifies the selection computations that we use to obtain full blocks of column pivots. Having blocks of pivots allows our algorithm to factorize $\boldsymbol{A}$ using matrix-matrix multiplications instead of interleaving multiple series of matrix-vector operations, making RQRCP the algorithm of choice for numerical rank determination.

Critically, we have shown how to leverage intermediate block transformations of $\boldsymbol{A}$ to update $\boldsymbol{B}$ with a sample update formula. This technique allows us to avoid computing a new randomized projection for each block operation, thus substantially reducing the matrix-matrix multiplication work needed to process each block.

We have extended this method of factorization to produce truncated low-rank approximations. TRQRCP, the truncated formulation of RQRCP, avoids block updates to the

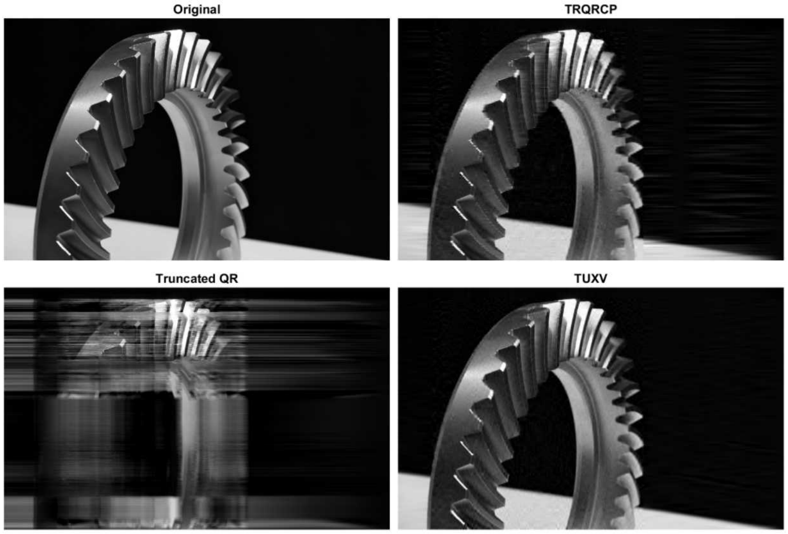

Figure 3.5: Low-rank image reconstruction comparison. Reconstructions are computed from a 2442 × 3888 grayscale matrix. Approximations are reconstructed with rank 80. Truncated QR is computed with columns presorted by descending 2-norm. TRQRCP dramatically improves visual approximation quality. Likewise, TUXV further improves fine details.

trailing matrix during factorization, which reduces the leading contribution to communication for low-rank approximations. Moreover, TRQRCP provides an efficient initial operation in our approximation of the truncated SVD, TUXV.

Our algorithms are implemented in Fortran with OpenMP. Numerical experiments compare performance with LAPACK subroutines, linked with the Intel Math Kernel Library, using a 24-core system. These experiments demonstrate that the computation time of RQRCP is nearly as short as that of unpivoted QR and substantially shorter than QRCP. Problems that have been too large to process with QRCP-dependent subroutines may now become feasible. Other applications that had to settle for QR, due to performance constraints, may find improved numerical stability at little cost by switching to RQRCP. For low-rank approximations, TRQRCP offers an additional performance advantage and TUXV improves approximation quality at a modest additional cost. These algorithms open a new performance domain for large matrix factorizations that we believe will be useful in science, engineering, and data analysis.

Randomized methods harness the fact that it is possible to make good algorithmic decisions in the face of uncertainty. By permitting controlled uncertainty, we can substantially improve algorithm performance. Future work will address how we may understand the foundations of credible uncertainty in predictions. Improving this understanding will facilitate the development of efficient predictive algorithms and support our ability to make good decisions from limited information.

Acknowledgements

We would like to extend our earnest appreciation to Michael Bierma, Andrew Charman, James Demmel, Laura Grigori, Christopher Harrison, Steven Holtzen, Justin Jacobs, Philip Kegelmeyer, Chris Melgaard, Jaideep Ray, and Jean-Paul Watson for several helpful discussions on these topics. We also sincerely thank anonymous reviewers and referees for providing expert feedback to improve this work. We also deeply appreciate editing feedback by Jacquilyn Weeks that improved the clarity of presentation of these ideas.

Funding: This work was funded by the U.S. Department of Energy. Original work was funded, in part, by NSF award 1319312 and by the Department of Homeland Security under DOE Contract Number DE-NA0003525 for the management and operation of Sandia National Laboratories. Revised work was supported by NSF award 1760316 and by the Laboratory Directed Research and Development program at Sandia National Laboratories.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Appendix A. Proof of Principal Information Theory

Lemma 1 *Let $g(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}$ be a function such that $g(x_1 x_2) = g(x_1) + g(x_2)$ for all $x_1, x_2 > 0$. It follows that $g(x) = a \log(x)$ where a is a constant.*

Proofs of Lemma 1 given by Erdös (Erdös, 1946), Fadeev (Fadeev, 1957), Rényi (Rényi, 1961).

Lemma 2 *Let $f(\cdot,\cdot,\cdot) : \mathbb{R}_+^3 \mapsto \mathbb{R}$ be a function such that $f(r_1 r_2, q_1 q_2, p_1 p_2) = f(r_1, q_1, p_1) + f(r_2, q_2, p_2)$ for all $r_1, q_1, p_1, r_2, q_2, p_2 > 0$. It follows that $f(r, q, p) = \log\left(r^\gamma q^\alpha p^\beta\right)$.*

Proof of Lemma 2 We begin by defining $g_1(x) \equiv f(x^{-1}, x, x)$, $g_2(y) \equiv f(y, y^{-1}, y)$, and $g_3(z) \equiv f(z, z, z^{-1})$. It follows that $g_1(x_1 x_2) = g(x_1) + g(x_2)$. From Lemma 1, we have $g_1(x) = a \log(x)$ for some constant a. Similarly, $g_2(y) = b \log(y)$ and $g_3(z) = c \log(z)$ for constants b and c. We may now construct positive quantities $x = \sqrt{qp}$, $y = \sqrt{pr}$, and $z = \sqrt{rq}$ and observe $f(r, q, p) = f(x^{-1}yz, xy^{-1}z, xyz^{-1}) = f(x^{-1}, x, x) + f(y, y^{-1}, y) + f(z, z, z^{-1}) =$

$g_1(x) + g_2(y) + g_1(z)$. The desired result follows by identifying constants $\alpha = (c+a)/2$, $\beta = (a+b)/2$, and $\gamma = (b+c)/2$ ■

Proof of Theorem 1 We proceed by combining Postulate 1 with Postulate 2, which gives

$$\begin{aligned}
&\mathbb{I}_{\mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w})}[\,\mathbf{q}_1(\boldsymbol{z})\mathbf{q}_1(\boldsymbol{w}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\mathbf{q}_0(\boldsymbol{w})\,] \\
&= \int d\boldsymbol{z}\, d\boldsymbol{w}\, \mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w}) f(\mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w}), \mathbf{q}_1(\boldsymbol{z})\mathbf{q}_1(\boldsymbol{w}), \mathbf{q}_0(\boldsymbol{z})\mathbf{q}_0(\boldsymbol{w})) \\
&= \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] + \mathbb{I}_{\mathbf{r}(\boldsymbol{w})}[\,\mathbf{q}_1(\boldsymbol{w}) \,\|\, \mathbf{q}_0(\boldsymbol{w})\,] \\
&= \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) f(\mathbf{r}(\boldsymbol{z}), \mathbf{q}_1(\boldsymbol{z}), \mathbf{q}_0(\boldsymbol{z})) + \int d\boldsymbol{w}\, \mathbf{r}(\boldsymbol{w}) f(\mathbf{r}(\boldsymbol{w}), \mathbf{q}_1(\boldsymbol{w}), \mathbf{q}_0(\boldsymbol{w})) \\
&= \int d\boldsymbol{z}\, d\boldsymbol{w}\, \mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w}) \left[f(\mathbf{r}(\boldsymbol{z}), \mathbf{q}_1(\boldsymbol{z}), \mathbf{q}_0(\boldsymbol{z})) + f(\mathbf{r}(\boldsymbol{w}), \mathbf{q}_1(\boldsymbol{w}), \mathbf{q}_0(\boldsymbol{w})) \right].
\end{aligned}$$

The last line follows by multiplying each term in the previous line by $\int d\boldsymbol{w}\, \mathbf{r}(\boldsymbol{w}) = 1$ and $\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) = 1$, respectively. Since this must hold for arbitrary $\mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w})$, this implies

$$f(\mathbf{r}(\boldsymbol{z})\mathbf{r}(\boldsymbol{w}), \mathbf{q}_1(\boldsymbol{z})\mathbf{q}_1(\boldsymbol{w}), \mathbf{q}_0(\boldsymbol{z})\mathbf{q}_0(\boldsymbol{w})) = f(\mathbf{r}(\boldsymbol{z}), \mathbf{q}_1(\boldsymbol{z}), \mathbf{q}_0(\boldsymbol{z})) + f(\mathbf{r}(\boldsymbol{w}), \mathbf{q}_1(\boldsymbol{w}), \mathbf{q}_0(\boldsymbol{w})).$$

By Lemma 2, we have $f\left(\mathbf{r}(\boldsymbol{z}), \mathbf{q}_1(\boldsymbol{z}), \mathbf{q}_0(\boldsymbol{z})\right) = \log\left(\mathbf{r}(\boldsymbol{z})^{\gamma}\mathbf{q}_1(\boldsymbol{z})^{\alpha}\mathbf{q}_0(\boldsymbol{z})^{\beta}\right)$. From Postulate 3, we require

$$\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\mathbf{r}(\boldsymbol{z})^{\gamma}\mathbf{q}_0(\boldsymbol{z})^{\alpha+\beta}\right) = \gamma \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log \mathbf{r}(\boldsymbol{z}) + (\alpha+\beta) \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log \mathbf{q}_0(\boldsymbol{z}) = 0.$$

Since this must hold for arbitrary $\mathbf{r}(\boldsymbol{z})$ and $\mathbf{q}_0(\boldsymbol{z})$, this implies $\gamma = 0$ and $\beta = -\alpha$. Thus we see that $\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{r}(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \alpha D_{KL}[\,\mathbf{r}(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]$, which is a constant α times the Kullback–Leibler divergence (Kullback and Leibler, 1951). Jensen's inequality easily shows nonnegativity of the form

$$\begin{aligned}
&\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{r}(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) = -\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_0(\boldsymbol{z})}{\mathbf{r}(\boldsymbol{z})}\right) \\
&\geq -\log\left(\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \frac{\mathbf{q}_0(\boldsymbol{z})}{\mathbf{r}(\boldsymbol{z})}\right) = -\log(1) = 0.
\end{aligned}$$

It follows from Postulate 4 that $\alpha > 0$. As Shannon notes, the scale is arbitrary and simply defines the unit of measure. ■

Appendix B. Proofs of Information Corollaries

Proof of triangle inequality for Definition 2 To simplify notation, we use the following definitions

$$\begin{aligned}
a(\boldsymbol{z}) &= \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right), \quad \alpha = \mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \left(\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z})\, |a(\boldsymbol{z})|^p\right)^{1/p}, \\
b(\boldsymbol{z}) &= \log\left(\frac{\mathbf{q}_2(\boldsymbol{z})}{\mathbf{q}_1(\boldsymbol{z})}\right), \quad \beta = \mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_1(\boldsymbol{z})\,] = \left(\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z})\, |b(\boldsymbol{z})|^p\right)^{1/p}.
\end{aligned}$$

Applying homogeneity followed by Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{L}^p_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] = \left(\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z})\, |a(\boldsymbol{z}) + b(\boldsymbol{z})|^p\right)^{1/p} \\
&= (\alpha+\beta)\left(\int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \left| \left(\frac{\alpha}{\alpha+\beta}\right) \frac{a(\boldsymbol{z})}{\alpha} + \left(\frac{\beta}{\alpha+\beta}\right) \frac{b(\boldsymbol{z})}{\beta} \right|^p \right)^{1/p} \\
&\leq (\alpha+\beta)\left(\left(\frac{\alpha}{\alpha+\beta}\right) \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \left|\frac{a(\boldsymbol{z})}{\alpha}\right|^p + \left(\frac{\beta}{\alpha+\beta}\right) \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \left|\frac{b(\boldsymbol{z})}{\beta}\right|^p \right)^{1/p} \\
&= (\alpha+\beta)\left(\frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}\right)^{1/p} \\
&= \alpha+\beta.
\end{aligned}$$

■

Proof of Corollary 1 We unpack Theorem 1 and write joint distributions as the marginalization times the corresponding conditional distribution such as $\mathbf{r}(\boldsymbol{z}_1, \boldsymbol{z}_2) \equiv \mathbf{r}(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)\mathbf{r}(\boldsymbol{z}_1)$. This gives

$$\begin{aligned}
&\mathbb{I}_{\mathbf{r}(\boldsymbol{z}_1,\boldsymbol{z}_2)}[\,\mathbf{q}_1(\boldsymbol{z}_1, \boldsymbol{z}_2) \,\|\, \mathbf{q}_0(\boldsymbol{z}_1, \boldsymbol{z}_2)\,] \\
&= \int d\boldsymbol{z}_1\, d\boldsymbol{z}_2\, \mathbf{r}(\boldsymbol{z}_1, \boldsymbol{z}_2) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z}_1, \boldsymbol{z}_2)}{\mathbf{q}_0(\boldsymbol{z}_1, \boldsymbol{z}_2)}\right) \\
&= \int d\boldsymbol{z}_1\, \mathbf{r}(\boldsymbol{z}_1) \int d\boldsymbol{z}_2\, \mathbf{r}(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z}_1)\mathbf{q}_1(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)}{\mathbf{q}_0(\boldsymbol{z}_1)\mathbf{q}_0(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)}\right) \\
&= \int d\boldsymbol{z}_1\, \mathbf{r}(\boldsymbol{z}_1) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z}_1)}{\mathbf{q}_0(\boldsymbol{z}_1)}\right) \\
&\quad + \int d\boldsymbol{z}_1\, \mathbf{r}(\boldsymbol{z}_1) \int d\boldsymbol{z}_2\, \mathbf{r}(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)}{\mathbf{q}_0(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)}\right) \\
&= \mathbb{I}_{\mathbf{r}(\boldsymbol{z}_1)}[\,\mathbf{q}_1(\boldsymbol{z}_1) \,\|\, \mathbf{q}_0(\boldsymbol{z}_1)\,] + \mathbb{E}_{\mathbf{r}(\boldsymbol{z}_1)}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{z}_2|\boldsymbol{z}_1)}[\,\mathbf{q}_1(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \,\|\, \mathbf{q}_0(\boldsymbol{z}_2 \mid \boldsymbol{z}_1)\,].
\end{aligned}$$

■

Proof of Corollary 2 Again, we simply unpack Theorem 1 and apply the product property of the logarithm as

$$\begin{aligned}
\mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] &= \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_2(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) \\
&= \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_2(\boldsymbol{z})\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_1(\boldsymbol{z})\mathbf{q}_0(\boldsymbol{z})}\right) \\
&= \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_2(\boldsymbol{z})}{\mathbf{q}_1(\boldsymbol{z})}\right) + \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) \\
&= \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_2(\boldsymbol{z}) \,\|\, \mathbf{q}_1(\boldsymbol{z})\,] + \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,].
\end{aligned}$$

■

Proof of Corollary 3 Swapping $\mathbf{q}_0(\boldsymbol{z})$ and $\mathbf{q}_1(\boldsymbol{z})$ reciprocates the argument of the logarithm in Theorem 1, which gives the negative of the original ordering. ■

Proof of Corollary 4 After realization $\check{\boldsymbol{z}} = \boldsymbol{z}_j$, probability is distributed as $\mathbf{r}(\boldsymbol{z} = \boldsymbol{z}_i \mid \boldsymbol{z}_j) = \delta_{ij}$. Restricting support to $\boldsymbol{z} = \check{\boldsymbol{z}}$ gives

$$\mathbb{I}_{\mathbf{r}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\,\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}}) \,\|\, \mathbf{q}(\boldsymbol{z})\,] = \log\left(\frac{1}{\mathbf{q}(\boldsymbol{z} = \check{\boldsymbol{z}})}\right) = \mathbb{D}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z} = \check{\boldsymbol{z}})\,]\,.$$

■

Proof of Corollary 5 Computing the expectation value as given easily reconstructs the standard formulation of cross entropy

$$\begin{aligned} S_{\mathbf{r}(\boldsymbol{z})}[\,\mathbf{q}(\boldsymbol{z})\,] &= \mathbb{E}_{\mathbf{r}(\check{\boldsymbol{z}})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\,\mathbf{r}(\boldsymbol{z} \mid \check{\boldsymbol{z}}) \,\|\, \mathbf{q}(\boldsymbol{z})\,] \\ &= \int d\boldsymbol{z}\,\mathbf{r}(\boldsymbol{z}) \log\left(\frac{1}{\mathbf{q}(\boldsymbol{z})}\right) \\ &= \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\,1 \,\|\, \mathbf{q}(\boldsymbol{z})\,]\,. \end{aligned}$$

■

Proof of Corollary 6 This follows by simply replacing $\mathbf{r}(\boldsymbol{z})$ with $\mathbf{q}(\boldsymbol{z})$ in cross entropy. ■

Proof of Corollary 7 Plausible joint values of $\boldsymbol{z}$ and $\boldsymbol{w}$ are $\mathbf{p}(\boldsymbol{z}, \boldsymbol{w}) = \mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w})\mathbf{p}(\boldsymbol{w})$. Marginalizing over $\boldsymbol{w}$ recovers present belief $\int d\boldsymbol{w}\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w})\mathbf{p}(\boldsymbol{w}) = \mathbf{p}(\boldsymbol{z})$. It follows

$$\begin{aligned} \mathbb{E}_{\mathbf{p}(\boldsymbol{w})}\,\mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{w})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,] &= \int d\boldsymbol{w}\,\mathbf{p}(\boldsymbol{w}) \int d\boldsymbol{z}\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) \\ &= \int d\boldsymbol{z}\,\mathbf{p}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) \\ &= \mathbb{I}_{\mathbf{p}(\boldsymbol{z})}[\,\mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z})\,]\,. \end{aligned}$$

■

Proof of Corollary 8 We compute the expectation value stated and simply rewrite the product of the marginalization and conditional distribution as the joint distribution $\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w})\mathbf{p}(\boldsymbol{w}) \equiv \mathbf{p}(\boldsymbol{z}, \boldsymbol{w})$. This gives

$$\begin{aligned} &\mathbb{E}_{\mathbf{p}(\boldsymbol{w})}\,\mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\boldsymbol{w})}[\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w}) \,\|\, \mathbf{p}(\boldsymbol{z})\,] \\ &= \int d\boldsymbol{w}\,\mathbf{p}(\boldsymbol{w}) \int d\boldsymbol{z}\,\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w}) \log\left(\frac{\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{w})\mathbf{p}(\boldsymbol{w})}{\mathbf{p}(\boldsymbol{w})\mathbf{p}(\boldsymbol{z})}\right) \\ &= \int d\boldsymbol{w}\,d\boldsymbol{z}\,\mathbf{p}(\boldsymbol{z}, \boldsymbol{w}) \log\left(\frac{\mathbf{p}(\boldsymbol{z}, \boldsymbol{w})}{\mathbf{p}(\boldsymbol{z})\mathbf{p}(\boldsymbol{w})}\right) \\ &= \mathbb{I}_{\mathbf{p}(\boldsymbol{z},\boldsymbol{w})}[\,\mathbf{p}(\boldsymbol{z}, \boldsymbol{w}) \,\|\, \mathbf{p}(\boldsymbol{z})\mathbf{p}(\boldsymbol{w})\,]\,. \end{aligned}$$

■

Proof of Corollary 9 The limit of increasing precision yields the Dirac delta function $\mathbf{p}(\boldsymbol{z} \mid \check{\boldsymbol{z}}) \equiv \boldsymbol{\delta}(\boldsymbol{z} - \check{\boldsymbol{z}})$. It follows

$$\mathbb{I}_{\mathbf{p}(\boldsymbol{z}|\check{\boldsymbol{z}})}[\, \mathbf{q}_1(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] = \int d\boldsymbol{z}\, \boldsymbol{\delta}(\boldsymbol{z} - \check{\boldsymbol{z}}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})} \right) = \log\left(\frac{\mathbf{q}_1(\boldsymbol{z} = \check{\boldsymbol{z}})}{\mathbf{q}_0(\boldsymbol{z} = \check{\boldsymbol{z}})} \right).$$

■

Proof of Corollary 10 We consider differential variations at the optimizer $\mathbf{q}_1(\boldsymbol{z}) = \mathbf{q}^*(\boldsymbol{z}) + \varepsilon\boldsymbol{\eta}(\boldsymbol{z})$ where variations $\boldsymbol{\eta}(\boldsymbol{z})$ must maintain normalization $\int d\boldsymbol{z}\, \boldsymbol{\eta}(\boldsymbol{z}) = 0$ and are otherwise arbitrary. Taking the Gâteaux derivative (with respect to the differential element ε) and applying the variational principle gives

$$0 = \int d\boldsymbol{z}\, \boldsymbol{\eta}(\boldsymbol{z}) \left[\frac{\mathbf{p}(\boldsymbol{z} \mid \boldsymbol{x})}{\mathbf{q}^*(\boldsymbol{z})} \right].$$

To satisfy the normalization constraint for otherwise arbitrary $\boldsymbol{\eta}(\boldsymbol{z})$, the term in brackets must be constant. The stated result immediately follows. ■

Proof of Corollary 11 Let measurable disjoint subsets of outcomes be $\Omega_> = \{\boldsymbol{z} \,|\, \mathbf{r}(\boldsymbol{z}) > \mathbf{q}_1(\boldsymbol{z}) > 0\}$ and $\Omega_< = \{\boldsymbol{z} \,|\, \mathbf{r}(\boldsymbol{z}) < \mathbf{q}_1(\boldsymbol{z})\}$. If $\boldsymbol{\eta}(\boldsymbol{z})$ drives belief toward $\mathbf{r}(\boldsymbol{z})$ on all measurable subsets then $\boldsymbol{\eta}(\boldsymbol{z}) \geq 0$ for $\boldsymbol{z}$ almost everywhere in $\Omega_>$. Likewise, $\boldsymbol{\eta}(\boldsymbol{z}) \leq 0$ almost everywhere in $\Omega_<$. Finally, $\boldsymbol{\eta}(\boldsymbol{z}) = 0$ almost everywhere on the complement $\Omega_{\boldsymbol{z}} \setminus (\Omega_> \cup \Omega_<)$. In order to retain normalization, we note that $\int d\boldsymbol{z}\, \boldsymbol{\eta}(\boldsymbol{z}) = 0$. If information is finite, then we may express $\mathbf{r}(\boldsymbol{z}) = \mathbf{q}_1(\boldsymbol{z})(1 + \boldsymbol{\delta}(\boldsymbol{z}))$ almost everywhere (except an immeasurable subset that is not contained in $\Omega_> \cup \Omega_<$ for which we could have $\mathbf{q}_1(\boldsymbol{z}) = 0$ and $\mathbf{r}(\boldsymbol{z}) > 0$) and observe that $\boldsymbol{\delta}(\boldsymbol{z}) > 0$ for $\boldsymbol{z} \in \Omega_>$ just as $\boldsymbol{\delta}(\boldsymbol{z}) < 0$ for $\boldsymbol{z} \in \Omega_<$. Since $\boldsymbol{\eta}(\boldsymbol{z})$ has the same sign as $\boldsymbol{\delta}(\boldsymbol{z})$ almost everywhere, it follows

$$\begin{aligned}
&\lim_{\varepsilon \to 0} \frac{\partial}{\partial \varepsilon} \mathbb{I}_{\mathbf{r}(\boldsymbol{z})}[\, \mathbf{q}_1(\boldsymbol{z}) + \varepsilon\boldsymbol{\eta}(\boldsymbol{z}) \,\|\, \mathbf{q}_0(\boldsymbol{z}) \,] \\
&= \lim_{\varepsilon \to 0} \frac{\partial}{\partial \varepsilon} \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \log\left(\frac{\mathbf{q}_1(\boldsymbol{z}) + \varepsilon\boldsymbol{\eta}(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})} \right) \\
&= \int d\boldsymbol{z}\, \mathbf{r}(\boldsymbol{z}) \frac{\boldsymbol{\eta}(\boldsymbol{z})}{\mathbf{q}_1(\boldsymbol{z})} \\
&= \int d\boldsymbol{z}\, \left(1 + \boldsymbol{\delta}(\boldsymbol{z})\right) \boldsymbol{\eta}(\boldsymbol{z}) \\
&= \int d\boldsymbol{z}\, \boldsymbol{\delta}(\boldsymbol{z}) \boldsymbol{\eta}(\boldsymbol{z}) \\
&> 0.
\end{aligned}$$

■

Proof of Corollary 12 We proceed by constructing the Lagrangian

$$\mathcal{L}[\,\mathbf{r}(\boldsymbol{z}),\lambda\,] = \int d\boldsymbol{z}\,\mathbf{r}(\boldsymbol{z})\left(\log\!\left(\frac{\mathbf{r}(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) - \sum_{i=1}^{n}\lambda_i\,(f_i(\boldsymbol{z})-\varphi_i)\right).$$

We consider differential variations at the optimizer $\mathbf{r}(\boldsymbol{z}) = \mathbf{r}^{*}(\boldsymbol{z}) + \varepsilon\boldsymbol{\eta}(\boldsymbol{z})$ where variations $\boldsymbol{\eta}(\boldsymbol{z})$ must maintain normalization $\int d\boldsymbol{z}\,\boldsymbol{\eta}(\boldsymbol{z}) = 0$ and are otherwise arbitrary. Taking the Gâteaux derivative and applying the variational principle gives

$$0 = \int d\boldsymbol{z}\,\boldsymbol{\eta}(\boldsymbol{z})\left[\log\!\left(\frac{\mathbf{r}^{*}(\boldsymbol{z})}{\mathbf{q}_0(\boldsymbol{z})}\right) + 1 - \sum_{i=1}^{n}\lambda_i\,(f_i(\boldsymbol{z})-\varphi_i)\right].$$

To satisfy the normalization constraint for otherwise arbitrary $\boldsymbol{\eta}(\boldsymbol{z})$, the variational principle requires the term in brackets to be constant. The stated result immediately follows. ■

Proof of Corollary 13 We unpack Theorem 1 and apply the local conditionality property to write $\mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) \equiv \mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})$. This gives

$$\begin{aligned}
&\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\,] \\
&= \int d\boldsymbol{\theta}_1\,d\boldsymbol{\theta}_2\,\mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})\log\!\left(\frac{\mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}\right) \\
&= \int d\boldsymbol{\theta}_2\,\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})\log\!\left(\frac{\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\mathbf{p}(\boldsymbol{\theta}_1)}\right) \\
&= \int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})\log\!\left(\frac{\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_1)}\right) \\
&= \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1)\,].
\end{aligned}$$

■

Proof of Corollary 14 As a consequence of local conditional dependence, we observe that $\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) = \int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})$. Then we apply Jensen's inequality followed by Bayes' Theorem to obtain

$$\begin{aligned}
&\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_2|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_2)\,] \\
&= \int d\boldsymbol{\theta}_2\left[\int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})\right]\log\!\left(\frac{\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_2)}\right) \\
&\leq \int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})\log\!\left(\int d\boldsymbol{\theta}_2\,\mathbf{p}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\frac{\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_2)}\right) \\
&= \int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})\log\!\left(\int d\boldsymbol{\theta}_2\,\frac{\mathbf{p}(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_1)}\right) \\
&= \int d\boldsymbol{\theta}_1\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})\left[\log\!\left(\frac{\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_1)}\right) + \log\!\left(\int d\boldsymbol{\theta}_2\,\frac{\mathbf{p}(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}})}\right)\right].
\end{aligned}$$

The first term provides the upper bound we seek. It remains to show that the second term is bound from above by zero, which follows from a second application of Jensen's inequality

$$\begin{aligned}
&\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_2|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_2)\,] \\
&\leq \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1)\,] + \log\left(\int d\boldsymbol{\theta}_1\, d\boldsymbol{\theta}_2\, \mathbf{p}(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)\mathbf{p}(\boldsymbol{\theta}_2 \mid \check{\boldsymbol{y}})\right) \\
&= \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}_1|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta}_1 \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta}_1)\,]\,.
\end{aligned}$$

■

Proof of Corollary 15 The denominator of the first log argument is model evidence and we apply Bayes' Theorem to the second log argument. Denominators of log arguments cancel and Jensen's inequality implies that the first term must be greater than the second.

$$\begin{aligned}
&\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{q}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{y})\,] - \mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{\theta})\,] \\
&= \int d\boldsymbol{y}\, \boldsymbol{\delta}(\boldsymbol{y} - \check{\boldsymbol{y}}) \log\left(\frac{\int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})}{\int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta})}\right) - \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) \log\left(\frac{\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta})}\right) \\
&= \log\left(\frac{\int d\boldsymbol{\theta}\, \mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})}{\mathbf{p}(\check{\boldsymbol{y}})}\right) - \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) \log\left(\frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})}{\mathbf{p}(\check{\boldsymbol{y}})}\right) \\
&= \log\left(\int d\boldsymbol{\theta}\, \mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})\right) - \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}) \log(\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})) \geq 0.
\end{aligned}$$

■

Proof of Corollary 16 The stated result immediately follows by taking $f_i(\boldsymbol{\theta}) \equiv \log\left(\frac{\mathbf{q}_i(\boldsymbol{\theta})}{\mathbf{q}_0(\boldsymbol{\theta})}\right)$ and applying Corollary 12. ■

Proof of Corollary 17 We take $n = 1$, $\mathbf{q}_0(\boldsymbol{\theta}) \equiv \mathbf{p}(\boldsymbol{\theta})$, and $\mathbf{q}_1(\boldsymbol{\theta}) \equiv \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})$ and apply Corollary 16 with Bayes' theorem to obtain

$$\mathbf{r}(\boldsymbol{\theta}) \propto \mathbf{p}(\boldsymbol{\theta}) \left(\frac{\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}})}{\mathbf{p}(\boldsymbol{\theta})}\right)^{\boldsymbol{\lambda}} = \mathbf{p}(\boldsymbol{\theta}) \left(\frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})}{\mathbf{p}(\check{\boldsymbol{y}})}\right)^{\boldsymbol{\lambda}}.$$

■

Appendix C. Information computations in experiment

C.1 Inference

Prior belief is $\mathbf{p}(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\theta} \mid 0, \boldsymbol{A})$ and $\mathbf{p}(\boldsymbol{x}^{(j)}) \equiv \mathcal{N}(\boldsymbol{x}^{(j)} \mid 0, \boldsymbol{\Sigma})$. Since $\boldsymbol{y}^{(j)} = \boldsymbol{\theta} + \boldsymbol{x}^{(j)}$, we have $\mathbf{p}(\boldsymbol{y}^{(j)} \mid \boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{y}^{(j)} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})$. If we let n samples have an average $\bar{\boldsymbol{y}}$, it easily follows that $\mathbf{p}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta}) \equiv \mathcal{N}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta}, \frac{1}{n}\boldsymbol{\Sigma})$. Bayes' rule gives $\mathbf{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{y}}) \propto \mathbf{p}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta})$ or

$$\begin{aligned}\mathbf{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{y}}) &\propto \exp\left[\frac{-1}{2}\boldsymbol{\theta}^T\boldsymbol{A}^{-1}\boldsymbol{\theta} - \frac{n}{2}(\boldsymbol{\theta} - \bar{\boldsymbol{y}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{\boldsymbol{y}})\right] \\ &\propto \exp\left[\frac{-1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T\boldsymbol{B}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\end{aligned}$$

where $\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1} + n\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\mu} = n\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}}$. Normalization yields $\mathbf{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{y}}) \equiv \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{B})$.

C.2 Mutual information

We marginalize over plausible $\boldsymbol{\theta}$ to obtain corresponding probability of observing $\bar{\boldsymbol{y}}$ as $\mathbf{p}(\bar{\boldsymbol{y}}) = \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{\theta})\mathbf{p}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta})$. This gives $\mathbf{p}(\bar{\boldsymbol{y}}) \equiv \mathcal{N}(\bar{\boldsymbol{y}} \mid 0, \boldsymbol{A} + \frac{1}{n}\boldsymbol{\Sigma})$. Mutual information, which is the expected information gained by observing $\bar{\boldsymbol{y}}$ according to present belief, is computed

$$\begin{aligned}&\int d\bar{\boldsymbol{y}}\, d\boldsymbol{\theta}\, \mathbf{p}(\bar{\boldsymbol{y}}, \boldsymbol{\theta}) \log\left(\frac{\mathbf{p}(\bar{\boldsymbol{y}}, \boldsymbol{\theta})}{\mathbf{p}(\bar{\boldsymbol{y}})\mathbf{p}(\boldsymbol{\theta})}\right) \\ &= \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{\theta}) \int d\bar{\boldsymbol{y}}\, \mathbf{p}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta}) \log\left(\frac{\mathbf{p}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta})}{\mathbf{p}(\bar{\boldsymbol{y}})}\right) \\ &= \int d\boldsymbol{\theta}\, \mathbf{p}(\boldsymbol{\theta}) \int d\bar{\boldsymbol{y}}\, \mathbf{p}(\bar{\boldsymbol{y}} \mid \boldsymbol{\theta}) \log\left(\frac{\left|2\pi\frac{1}{n}\boldsymbol{\Sigma}\right|^{-1/2} \exp\left(\frac{-n}{2}(\bar{\boldsymbol{y}} - \boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\theta})\right)}{\left|2\pi(\boldsymbol{A} + \frac{1}{n}\boldsymbol{\Sigma})\right|^{-1/2} \exp\left(\frac{-1}{2}\bar{\boldsymbol{y}}^T(\boldsymbol{A} + \frac{1}{n}\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{y}}\right)}\right) \\ &= \frac{1}{2}\log\det\left(n\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{I}\right)\end{aligned}$$

C.3 First inference information in a subsequent view

Let the state of belief before an experiment be $\mathbf{p}(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\theta} \mid 0, \boldsymbol{A})$. After observing $\bar{\boldsymbol{y}}^{(1)}$ inference gives $\mathbf{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{y}}^{(1)}) \equiv \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{B})$. Additional observations $\bar{\boldsymbol{y}}^{(2)}$ yield $\mathbf{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{y}}^{(1)}, \bar{\boldsymbol{y}}^{(2)}) \equiv \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\nu}, \boldsymbol{C})$. Information gained in the first observation in the view of inference following the second observation is computed

$$\begin{aligned}&\mathbb{I}_{\mathbf{p}(\boldsymbol{\theta}\mid\bar{\boldsymbol{y}}^{(1)},\bar{\boldsymbol{y}}^{(2)})}\left[\mathbf{p}(\boldsymbol{\theta} \mid \bar{\boldsymbol{y}}^{(1)}) \,\middle\|\, \mathbf{p}(\boldsymbol{\theta})\right] \\ &= \int d\boldsymbol{\theta}\, \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\nu}, \boldsymbol{C}) \log\left(\frac{|2\pi\boldsymbol{B}|^{-1/2} \exp\left(\frac{-1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T\boldsymbol{B}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)}{|2\pi\boldsymbol{A}|^{-1/2} \exp\left(\frac{-1}{2}\boldsymbol{\theta}^T\boldsymbol{A}^{-1}\boldsymbol{\theta}\right)}\right) \\ &= \frac{1}{2}\left(\log\det\left(\boldsymbol{A}\boldsymbol{B}^{-1}\right) + \operatorname{tr}\left((\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1})\boldsymbol{C}\right) + \boldsymbol{\nu}^T\boldsymbol{A}^{-1}\boldsymbol{\nu} - (\boldsymbol{\nu} - \boldsymbol{\mu})^T\boldsymbol{B}^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu})\right).\end{aligned}$$

Appendix D. Proofs of Parsimonious Inference Theorem and Corollaries

Proof of Corollary 18. The primary complication is that each alphabet $\boldsymbol{\Sigma}_{j+1}$ depends on previously realized symbols $(\boldsymbol{s})_1^j$. We proceed by induction. The induction hypothesis regarding a partial sequence $(\boldsymbol{s})_1^j$ for $j < n$ is

$$\mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^j)}\left[\sum_{i=1}^{j} \log_2\left(|\boldsymbol{\Sigma}_i|\right)\right] \geq \mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^j)}\left[\sum_{i=1}^{j} \log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1})}\right)\right].$$

The base case associated with the first symbol $\boldsymbol{s}_1$ easily follows from Jensen's inequality as

$$\begin{aligned}\mathbb{E}_{\mathbf{p}(\boldsymbol{s}_1)} \log_2\left(|\boldsymbol{\Sigma}_1|\right) &= \log_2\left(|\boldsymbol{\Sigma}_1|\right) = \log_2\left(\sum_{\boldsymbol{s}_1\in\boldsymbol{\Sigma}_1} 1\right)\\ &\geq \sum_{\boldsymbol{s}_1\in\boldsymbol{\Sigma}_1} \mathbf{p}(\boldsymbol{s}_1)\log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_1)}\right) = \mathbb{E}_{\mathbf{p}(\boldsymbol{s}_1)}\log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_1)}\right).\end{aligned}$$

The induction step is given by applying Jensen's inequality and the induction hypothesis

$$\begin{aligned}&\mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^{j+1})}\left[\sum_{i=1}^{j+1}\log_2\left(|\boldsymbol{\Sigma}_i|\right)\right]\\ &= \mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^{j})}\left[\sum_{i=1}^{j}\log_2\left(|\boldsymbol{\Sigma}_i|\right) + \mathbb{E}_{\mathbf{p}(\boldsymbol{s}_{j+1}|(\boldsymbol{s})_1^j)}\left[\log_2\left(|\boldsymbol{\Sigma}_{j+1}|\right)\right]\right]\\ &= \mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^{j})}\left[\sum_{i=1}^{j}\log_2\left(|\boldsymbol{\Sigma}_i|\right) + \log_2\left(|\boldsymbol{\Sigma}_{j+1}|\right)\right]\\ &\geq \mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^{j})}\left[\sum_{i=1}^{j}\log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1})}\right) + \mathbb{E}_{\mathbf{p}(\boldsymbol{s}_{j+1}|(\boldsymbol{s})_1^j)}\left[\log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_{j+1} \mid (\boldsymbol{s})_1^{j})}\right)\right]\right]\\ &= \mathbb{E}_{\mathbf{p}((\boldsymbol{s})_1^{j+1})}\left[\sum_{i=1}^{j+1}\log_2\left(\frac{1}{\mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1})}\right)\right].\end{aligned}$$

The claim follows by noting that $\mathbf{p}(\boldsymbol{a}) = \mathbf{p}((\boldsymbol{s})_1^n) = \prod_{i=1}^n \mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1})$. ■

Proof of Corollary 19. If the encoding is consistent with object probability and the encoding also maximizes entropy then

$$\mathbf{p}(\boldsymbol{a}) = \prod_{i=1}^{n}\mathbf{p}(\boldsymbol{s}_i \mid (\boldsymbol{s})_1^{i-1}) = \prod_{i=1}^{n}\frac{1}{|\boldsymbol{\Sigma}_i|} = 2^{-\ell(\boldsymbol{a})}.$$

■

Proof of Corollary 19, recursive consistency alternative. We also obtain the same prior by combining a counting argument with consistent prior belief in length descriptions. In order to distinguish a complete short sequence from merely a subsequence of a longer description, we need some indication of the complete sequence length. The following argument holds for descriptions $\boldsymbol{\psi}$ that are capable of being partitioned into a component $\boldsymbol{\psi_l}$ that determines the sequence length and the unconstrained complement $\boldsymbol{\psi_c}$ so that $\boldsymbol{\psi} = (\boldsymbol{\psi_l}, \boldsymbol{\psi_c})$ and $\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi_l}) + \ell(\boldsymbol{\psi_c})$. Once $\boldsymbol{\psi_l}$ is known, we can easily identify $\boldsymbol{\psi_c}$, regardless of its content. Yet, the same problem arises with knowing when we have a complete length description $\boldsymbol{\psi_l}$, which can be resolved with recursive partitions $\boldsymbol{\psi_l} = (\boldsymbol{\psi_{ll}}, \boldsymbol{\psi_{lc}})$ and so on. For the description to be finite, this recursion must end implicitly with only one possible outcome $\boldsymbol{\psi_{l\ldots ll}} = \emptyset$.

We define $\ell(\boldsymbol{\psi_c})$ so that the complement allows for $2^{\ell(\boldsymbol{\psi}_c)}$ outcomes. Binary sequences provide useful intuition for this construction. If we believe all descriptions of a given length have the same prior probability, then

$$1 = \int d\boldsymbol{\psi_c}\, \mathbf{p}(\boldsymbol{\psi_c} \mid \ell(\boldsymbol{\psi})) = 2^{\ell(\boldsymbol{\psi}_c)} \mathbf{p}(\boldsymbol{\psi_c} \mid \ell(\boldsymbol{\psi})) \quad \text{so that} \quad \mathbf{p}(\boldsymbol{\psi_c} \mid \ell(\boldsymbol{\psi})) = 2^{\ell(\boldsymbol{\psi})-\ell(\boldsymbol{\psi}_l)}.$$

Thus, $\mathbf{p}(\boldsymbol{\psi}) = 2^{\ell(\boldsymbol{\psi})-\ell(\boldsymbol{\psi}_l)}\mathbf{p}(\boldsymbol{\psi_l})$. But if the length description satisfies a consistent prior to determine our belief in the length of the seqence, we must also have $\mathbf{p}(\boldsymbol{\psi_l}) = 2^{\ell(\boldsymbol{\psi}_l)-\ell(\boldsymbol{\psi}_{ll})}\mathbf{p}(\boldsymbol{\psi_{ll}})$. This gives

$$\mathbf{p}(\boldsymbol{\psi}) = 2^{\ell(\boldsymbol{\psi})-\ell(\boldsymbol{\psi}_{ll})}\mathbf{p}(\boldsymbol{\psi_{ll}}).$$

Since the recursion ends with only one outcome, we have $\ell(\boldsymbol{\psi_{l\ldots ll}}) = 0$ and $\mathbf{p}(\boldsymbol{\psi_{l\ldots ll}}) = 1$. Thus, $\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi_c}) + \ell(\boldsymbol{\psi_{lc}}) + \cdots + \ell(\boldsymbol{\psi_{l\ldots lc}})$ and $\mathbf{p}(\boldsymbol{\psi}) = 2^{-\ell(\boldsymbol{\psi})}$. ■

Proof of Corollary 20. Let $\mathbf{r}^*(\boldsymbol{a})$ be the optimizer. We express arbitrary infinitesimal belief perturbations in the vicinity of the optimizer as $\mathbf{r}(\boldsymbol{a}) = \mathbf{r}^*(\boldsymbol{a}) + \varepsilon\boldsymbol{\eta}(\boldsymbol{a})$ where ε is a scalar differential element and $\boldsymbol{\eta}(\boldsymbol{a})$ is an arbitrary perturbation, so long as $\mathbf{r}^*(\boldsymbol{a}) > 0$, so that

$$\sum_{\boldsymbol{a}\in\mathcal{B}} \mathbf{r}(\boldsymbol{a}) = \sum_{\boldsymbol{a}\in\mathcal{B}} \mathbf{r}^*(\boldsymbol{a}) = 1 \quad \text{and} \quad \sum_{\boldsymbol{a}\in\mathcal{B}} \boldsymbol{\eta}(\boldsymbol{a}) = 0.$$

The information gained from prior belief to $\mathbf{r}(\boldsymbol{a})$ is

$$D_{KL}[\,\mathbf{r}(\boldsymbol{a}) \,\|\, \mathbf{p}(\boldsymbol{a})\,] = \sum_{\boldsymbol{a}\in\mathcal{B}} \mathbf{r}(\boldsymbol{a}) \log_2\!\left(\frac{\mathbf{r}(\boldsymbol{a})}{2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}}\right).$$

Differentiating with respect to ε, evaluating at $\varepsilon = 0$, and applying the variational principle yields

$$0 = \left[\frac{\partial}{\partial\varepsilon}\mathbb{I}_{\mathbf{r}(\boldsymbol{a})}[\,\mathbf{r}(\boldsymbol{a}) \,\|\, \mathbf{p}(\boldsymbol{a})\,]\right]_{\varepsilon=0} = \sum_{\boldsymbol{a}\in\mathcal{B}} \boldsymbol{\eta}(\boldsymbol{a}) \left(\log_2\!\left(\frac{\mathbf{r}^*(\boldsymbol{a})}{2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}}\right) + 1\right).$$

As this must hold for arbitrary $\boldsymbol{\eta}(\boldsymbol{a})$, the factor in parenthesis must be constant, provided $\mathbf{r}^*(\boldsymbol{a}) > 0$. Solving for $\mathbf{r}^*(\boldsymbol{a})$ shows that the stated distribution is the unique critical point. It remains to show that this distribution achieves the global minimum. Applying Jensen's inequality to the information gained from any feasible state gives

$$\begin{aligned} D_{KL}[\,\mathbf{r}(\boldsymbol{a}) \,\|\, \mathbf{p}(\boldsymbol{a})\,] &= -\sum_{\boldsymbol{a}\in\mathcal{B}} \mathbf{r}(\boldsymbol{a}) \log_2\!\left(\frac{2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}}{\mathbf{r}(\boldsymbol{a})}\right) \\ &\geq -\log_2\!\left(\sum_{\boldsymbol{a}\in\mathcal{B}} 2^{-\ell(\boldsymbol{\psi}(\boldsymbol{a}))}\right) = \log_2\!\left(\frac{1}{\mathbf{p}(\mathcal{B})}\right) = D_{KL}[\,\mathbf{r}^*(\boldsymbol{a}) \,\|\, \mathbf{p}(\boldsymbol{a})\,]. \end{aligned}$$

■

Proof of Theorem 2. The chain rule of conditional dependence, Corollary 1, allows us to express the information gained as

$$\begin{aligned}&D_{KL}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})\,]\\
&= \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi})}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})\,]\\
&= \mathbb{I}_{\mathbf{r}(\boldsymbol{\psi})}[\,\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\psi})\,] + \mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})}[\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\,]\\
&\quad + \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\,]\,.\end{aligned}$$

Combining properties of additivity over belief sequences, Corollary 2, with antisymmetry, Corollary 3, we express the argument of expectation in the last term as

$$\begin{aligned}&\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\,]\\
&\quad = \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,]\end{aligned}$$

where $\boldsymbol{\theta}_0$ is any fixed model that serves as a convenient baseline for measuring predictive information with regard to training labels. Once the data are observed, the term $\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,]$ is a constant, independent of choices $\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi})$. Thus we have

$$\begin{aligned}\omega &= D_{KL}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - D_{KL}[\,\mathbf{r}(\boldsymbol{y} \mid \check{\boldsymbol{y}})\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})\,]\\
&= \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})\mathbf{r}(\boldsymbol{\psi})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - \mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})}\,D_{KL}[\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\,]\\
&\quad - D_{KL}[\,\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\psi})\,]\,.\end{aligned}$$

We apply Corollary 19 as a prior over descriptions and unpack the KL divergence to obtain

$$-D_{KL}[\,\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\psi})\,] = \int d\boldsymbol{\psi}\,\mathbf{r}(\boldsymbol{\psi})\log_2\left(\frac{2^{-\ell(\boldsymbol{\psi})}}{\mathbf{r}(\boldsymbol{\psi})}\right) = S[\,\mathbf{r}(\boldsymbol{\psi})\,] - \mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})}\,\ell(\boldsymbol{\psi}).$$

■

Proof of Corollary 21. We unpack definitions, combine both terms, and apply Bayes' theorem to obtain

$$\begin{aligned}&\mathbb{E}_{\mathbf{r}(\boldsymbol{\theta}|\boldsymbol{\psi})}\,\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - D_{KL}[\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\,]\\
&= \int d\boldsymbol{\theta}\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\log_2\left(\frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta})\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})}\right)\\
&= \int d\boldsymbol{\theta}\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})\log_2\left(\frac{\mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}, \boldsymbol{\psi})\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi})}\right)\\
&= \log_2\left(\frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)}\right) - D_{KL}[\,\mathbf{r}(\boldsymbol{\theta} \mid \boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}, \boldsymbol{\psi})\,]\,.\end{aligned}$$

The second term is the negative Kullback-Leibler divergence, i.e. nonpositive. Therefore the objective is maximized when the second term vanishes, which occurs if and only if $\mathbf{r}^{*}(\boldsymbol{\theta} \mid \boldsymbol{\psi}) = \mathbf{p}(\boldsymbol{\theta} \mid \check{\boldsymbol{y}}, \boldsymbol{\psi})$. ■

Proof of Corollary 22. As in Corollary 21, we unpack definitions, combine both terms, and apply Bayes' theorem to obtain

$$\mathbb{E}_{\mathbf{r}(\boldsymbol{\psi})} \log_2\left(\frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)}\right) - D_{KL}[\,\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\psi})\,] = \sum_{\boldsymbol{\psi}} \mathbf{r}(\boldsymbol{\psi}) \log_2\left(\frac{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\psi})\mathbf{p}(\boldsymbol{\psi})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)\mathbf{r}(\boldsymbol{\psi})}\right)$$
$$= \sum_{\boldsymbol{\psi}} \mathbf{r}(\boldsymbol{\psi}) \log_2\left(\frac{\mathbf{p}(\boldsymbol{\psi} \mid \check{\boldsymbol{y}})\mathbf{p}(\check{\boldsymbol{y}})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)\mathbf{r}(\boldsymbol{\psi})}\right) = \log_2\left(\frac{\mathbf{p}(\check{\boldsymbol{y}})}{\mathbf{p}(\check{\boldsymbol{y}} \mid \boldsymbol{\theta}_0)}\right) - D_{KL}[\,\mathbf{r}(\boldsymbol{\psi}) \,\|\, \mathbf{p}(\boldsymbol{\psi} \mid \check{\boldsymbol{y}})\,]\,.$$

The objective is maximized if and only if the second term vanishes, thus $\mathbf{r}^{*}(\boldsymbol{\psi}) = \mathbf{p}(\boldsymbol{\psi} \mid \check{\boldsymbol{y}})$. ■

Proof of Corollary 23. Applying Jensen's inequality to expected prediction information from the optimizer gives

$$\begin{aligned}
&\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}^{*}(\boldsymbol{y}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - \boldsymbol{\chi}[\mathbf{r}^{*}(\boldsymbol{\psi}, \boldsymbol{\theta})] \\
&\geq \mathbb{E}_{\mathbf{r}^{*}(\boldsymbol{\theta},\boldsymbol{\psi})}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - \boldsymbol{\chi}[\mathbf{r}^{*}(\boldsymbol{\psi}, \boldsymbol{\theta})] \\
&\geq \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta},\boldsymbol{\psi})}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] - \boldsymbol{\chi}[\mathbf{r}(\boldsymbol{\psi}, \boldsymbol{\theta})].
\end{aligned}$$

Since $\mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}^{*}(\boldsymbol{y}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,] = \mathbb{E}_{\mathbf{r}(\boldsymbol{\theta},\boldsymbol{\psi})}\, \mathbb{I}_{\mathbf{r}(\boldsymbol{y}|\check{\boldsymbol{y}})}[\,\mathbf{r}^{*}(\boldsymbol{y}) \,\|\, \mathbf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\,]$, we can apply antisymmetry and additivity within the expectations to arrive at the stated result. ■

Bibliography

E. W. Adams. On rational betting systems. *Archive for Mathematical Logic*, 6(1):7–29, 1962.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, dec 1974. doi: 10.1109/tac.1974.1100705. URL https://doi.org/10.1109/tac.1974.1100705.

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *technometrics*, 16(1):125–127, 1974.

G. Barnard. The theory of information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(1):46–64, 1951.

C. Battaglino, G. Ballard, and T. G. Kolda. A practical randomized CP tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, 2018.

J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, pages 686–690, 1979.

J. Bernoulli. *Ars conjectandi*. Impensis Thurnisiorum, fratrum, 1713.

C. Bischof and C. Van Loan. The WY representation for products of Householder matrices. *SIAM Journal on Scientific and Statistical Computing*, 8(1):s2–s13, 1987.

C. H. Bischof. A parallel QR factorization algorithm with controlled local pivoting. *SIAM Journal on Scientific and Statistical Computing*, 12(1):36–57, 1991.

C. H. Bischof and P. C. Hansen. Structure-preserving and rank-revealing QR-factorizations. *SIAM Journal on Scientific and Statistical Computing*, 12(6):1332–1350, 1991. URL https://doi.org/10.1137/0912073.

R. B. Brandt. The concept of rational belief. *The Monist*, 68(1):3–23, 1985.

K. P. Burnham and D. R. Anderson. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 28(2):111, 2001. doi: 10.1071/wr99107. URL https://doi.org/10.1071/wr99107.

T. F. Chan. Rank revealing QR factorizations. *Linear Algebra and its Applications*, 88-89:67–82, Apr. 1987. doi: 10.1016/0024-3795(87)90103-0. URL https://doi.org/10.1016/0024-3795(87)90103-0.

T. F. Chan and P. C. Hansen. Some applications of the rank revealing QR factorization. *SIAM Journal on Scientific and Statistical Computing*, 13(3):727–741, 1992.

R. T. Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

B. De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. Communication-optimal parallel and sequential QR and LU factorizations. *SIAM Journal on Scientific Computing*, 34(1): A206–A239, 2012.

J. W. Demmel, L. Grigori, M. Gu, and H. Xiang. Communication avoiding rank revealing QR factorization with column pivoting. *SIAM Journal on Matrix Analysis and Applications*, 36(1):55–89, 2015.

J. A. Duersch and T. A. Catanach. Generalizing information to the evolution of rational belief. *Entropy*, 22(1):108, 2020.

J. A. Duersch and M. Gu. Randomized QR with column pivoting. *SIAM Journal on Scientific Computing*, 39(4):C263–C291, 2017.

J. A. Duersch and M. Gu. Randomized projection for rank-revealing matrix factorizations and low-rank approximations. *SIAM Review*, 62(3):661–682, 2020.

J. A. Duersch, M. Shao, C. Yang, and M. Gu. A robust and efficient implementation of LOBPCG. *SIAM Journal on Scientific Computing*, 40(5):C655–C676, Jan. 2018. doi: 10.1137/17m1129830. URL `https://doi.org/10.1137/17m1129830`.

N. Ebrahimi, E. S. Soofi, and H. Zahedi. Information properties of order statistics and spacings. *IEEE Transactions on Information Theory*, 50(1):177–183, 2004.

N. Ebrahimi, E. S. Soofi, and R. Soyer. Information measures in perspective. *International Statistical Review*, 78(3):383–412, 2010.

P. Erdös. On the distribution function of additive functions. *The Annals of Mathematics*, 47(1):1, Jan. 1946. doi: 10.2307/1969031.

N. B. Erichson, S. Voronin, S. L. Brunton, and J. N. Kutz. Randomized matrix decompositions using R. *Journal of Statistical Software*, 89(11):1–48, 2019.

R. A. Evans. The principle of minimum information. *IEEE Transactions on Reliability*, 18 (3):87–90, 1969.

D. Fadeev. Zum begriff der entropie einer endlichen wahrscheinlichkeitsschemas. *Arbeiten zur Informationstheorie I. Deutscher Verlag der Wissenschaften*, pages 85–90, 1957.

R. A. Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925.

D. A. Freedman and R. A. Purves. Bayes' method for bookies. *The Annals of Mathematical Statistics*, 40(4):1177–1186, 1969.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, Aug. 2013. doi: 10.1007/s11222-013-9416-2. URL `https://doi.org/10.1007/s11222-013-9416-2`.

G. H. Golub and C. F. Van Loan. *Matrix Computations*, volume 3. JHU Press, 2013.

I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934, 1963.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

A. Hájek. Dutch book arguments. *The handbook of rational and social choice*, pages 173–196, 2008.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

R. Hanel and S. Thurner. A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. *EPL (Europhysics Letters)*, 93(2):20006, 2011.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

U. Hetmaniuk and R. Lehoucq. Basis selection in LOBPCG. *Journal of Computational Physics*, 218(1):324–332, 2006.

N. J. Higham. QR factorization with complete pivoting and accurate computation of the SVD. *Linear Algebra and its Applications*, 309(1-3):153–174, 2000.

D. Hong, T. G. Kolda, and J. A. Duersch. Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163, 2020.

D. A. Huckaby and T. F. Chan. On the convergence of Stewart's QLP algorithm for approximating the SVD. *Numerical Algorithms*, 32(2-4):287–316, 2003.

D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

M. Hutter. On universal prediction and bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

V. M. Ilić and M. S. Stanković. Generalized shannon–khinchin axioms and uniqueness theorem for pseudo-additive entropies. *Physica A: Statistical Mechanics and its Applications*, 411:138–145, 2014.

E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

E. T. Jaynes. *Probability theory: The logic of science.* Cambridge university press, 2003.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007): 453–461, 1946.

H. Jeffreys. *The theory of probability.* OUP Oxford, 1998.

P. Jizba and T. Arimitsu. Generalized statistics: yet another generalization. *Physica A: Statistical Mechanics and its Applications*, 340(1-3):110–116, 2004.

P. Jizba and J. Korbel. Maximum entropy principle in statistical inference: case for nonshannonian entropies. *Physical review letters*, 122(12):120601, 2019.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

R. E. Kass and L. Wasserman. Formal rules for selecting prior distributions: A review and annotated bibliography. *Journal of the American Statistical Association*, 435:1343–1370, 1996.

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1):1–7, 1965.

A. Kovach. Differential gear, 2016. URL `https://commons.wikimedia.org/wiki/File:Diff_gear.jpg`.

S. Kullback. *Information theory and statistics.* Courier Corporation, 1997.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar. 1951. doi: 10.1214/aoms/1177729694.

P. S. LaPlace. Mémoire sur la probabilité des causes par les événements. *Mém. de math. et phys. présentés à lAcad. roy. des sci*, 6:621–656, 1774.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

R. S. Lehman. On confirmation and rational betting. *The Journal of Symbolic Logic*, 20 (3):251–262, 1955.

A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on information theory*, 22(1):75–81, 1976.

E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

D. V. Lindley. The Bayesian approach to statistics. Technical report, University of California, Berkeley, Operations Research Center, 1980.

D. J. MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

D. J. MacKay. Probable networks and plausible predictionsa review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3): 469–505, 1995.

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. ISBN 0521642981.

M. W. Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

P.-G. Martinsson and J. Tropp. Randomized numerical linear algebra: Foundations & algorithms. *Manuscript*, 2020. doi: https://arxiv.org/pdf/2002.01387.pdf.

P.-G. Martinsson and S. Voronin. A randomized blocked algorithm for efficiently computing rank-revealing factorizations of matrices. *SIAM Journal on Scientific Computing*, 38(5): S485–S507, 2016.

P.-G. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, Jan. 2011. doi: 10.1016/j.acha.2010.02.003. URL `https://doi.org/10.1016/j.acha.2010.02.003`.

P.-G. Martinsson, G. Q. Ortí, N. Heavner, and R. van de Geijn. Householder QR factorization with randomization for column pivoting (HQRRP). *SIAM J. Sci. Comput.*, 39(2): C96C115, 2017.

P.-G. Martinsson, G. Q. Ortí, and N. Heavner. randUTV: A blocked randomized algorithm for computing a rank-revealing UTV factorization. *ACM Transactions on Mathematical Software*, 45(1):1–26, 2019.

R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

O. Nikodym. Sur une généralisation des intégrales de MJ Radon. *Fundamenta Mathematicae*, 15(1):131–179, 1930.

H. Owhadi, C. Scovel, and T. Sullivan. On the brittleness of bayesian inference. *SIAM Review*, 57(4):566–582, Jan. 2015. doi: 10.1137/130938633. URL `https://doi.org/10.1137/130938633`.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

A. Potapov, A. Svitenkov, and Y. Vinogradov. Differences between kolmogorov complexity and solomonoff probability: consequences for agi. In *International Conference on Artificial General Intelligence*, pages 252–261. Springer, 2012.

C. Puglisi. Modification of the householder method based on the compact WY representation. *SIAM Journal on Scientific and Statistical Computing*, 13(3):723–726, 1992.

G. Quintana-Ortí, X. Sun, and C. H. Bischof. A BLAS-3 version of the QR factorization with column pivoting. *SIAM Journal on Scientific Computing*, 19(5):1486–1494, 1998.

F. P. Ramsey. Truth and probability. In *Readings in Formal Epistemology*, pages 21–45. Springer, 2016.

A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, Berkeley, Calif., 1961. University of California Press.

J. J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, pages 416–431, 1983.

J. J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, 30(4):629–636, 1984.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2010.

R. Schreiber and C. Van Loan. A storage-efficient WY representation for products of Householder transformations. *SIAM Journal on Scientific and Statistical Computing*, 10 (1):53–57, 1989.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, Mar. 1978. doi: 10.1214/aos/1176344136. URL `https://doi.org/10.1214/aos/1176344136`.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, July 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26 (1):26–37, 1980.

J. Shore and R. Johnson. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, 27(4):472–482, 1981.

B. Skyrms. Dynamic coherence and probability kinematics. *Philosophy of Science*, 54(1): 1–20, 1987.

R. J. Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964a.

R. J. Solomonoff. A formal theory of inductive inference. part ii. *Information and control*, 7(2):224–254, 1964b.

R. J. Solomonoff. Algorithmic probability: Theory and applications. In *Information theory and statistical learning*, pages 1–23. Springer, 2009.

E. S. Soofi. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89(428):1243–1254, 1994.

E. S. Soofi. Principal information theoretic approaches. *Journal of the American Statistical Association*, 95(452):1349–1353, 2000.

U. Speidel. On the bounds of the titchener t-complexity. In *2008 6th International Symposium on Communication Systems, Networks and Digital Signal Processing*, pages 321–325. IEEE, 2008.

A. Stathopoulos and K. Wu. A block orthogonalization procedure with constant synchronization requirements. *SIAM Journal on Scientific Computing*, 23(6):2165–2182, 2002.

G. Stewart. The QLP approximation to the singular value decomposition. *SIAM Journal on Scientific Computing*, 20(4):1336–1348, 1999.

P. Tempesta. Group entropies, correlation laws, and zeta functions. *Physical Review E*, 84 (2):021121, 2011.

N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

M. Titchener. Deterministic computation of string complexity, information and entropy. In *Inter. Symp. On Inform. Theory, Aug. 16-21, 1998, Boston*, 1998.

R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

J. Weisberg. Varieties of bayesianism. *Inductive Logic*, 10:477–551, 2011.

F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

J. Xiao, M. Gu, and J. Langou. Fast parallel randomized QR with column pivoting algorithms for reliable low-rank matrix approximations. In *2017 IEEE 24th International Conference on High Performance Computing (HiPC)*, pages 233–242. IEEE, 2017.

C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.