Bayesian Analysis (2020) **15**, Number 4, pp. 1299–1321

# An Explanatory Rationale for Priors Sharpened Into Occam's Razors

David R. Bickel[*]

**Abstract.** In Bayesian statistics, if the distribution of the data is unknown, then each plausible distribution of the data is indexed by a parameter value, and the prior distribution of the parameter is specified. To the extent that more complicated data distributions tend to require more coincidences for their construction than simpler data distributions, default prior distributions should be transformed to assign additional prior probability or probability density to the parameter values that refer to simpler data distributions. The proposed transformation of the prior distribution relies on the entropy of each data distribution as the relevant measure of complexity. The transformation is derived from a few first principles and extended to stochastic processes.

**Keywords:** explanatory coherence, foundations of Bayesian statistics, informative prior distribution, objective Bayes, objective prior distribution, Ockham's razor, simplicity postulate, sharpened prior distribution.

## 1 Introduction

The typical Bayesian data analysis involves specifying one or more default prior distributions, often called "objective priors" (Ghosh et al., 2006; Press, 2009). They are objective in the sense that they are automatically determined by the application of some algorithm as opposed to representing the beliefs of one or more people. The simplest case is the uniform prior distribution on a finite set of parameter values. In hypothesis testing, the assignment of equal prior probability to the null hypothesis and the alternative hypothesis is the most common default. In Bayesian model selection and Bayesian model averaging, the most common default is to assign each model equal prior probability. When the parameter value is continuous, more sophisticated procedures replace the assignment of equal probabilities (Kass and Wasserman, 1996).

The following toy models explain why default prior distributions may need to be modified to reflect the simplicity or complexity of each data distribution specified by a parameter value.

**Example 1.** The observable outcomes from a black box are independent and identically distributed (IID) integers between 1 and 20. Before observations are made, it is known that $n$ outcomes $x = (x_1, x_2, \ldots, x_n)$ will be generated by rolls of a fair die with a number on each face from 1 up to the number of sides of the die. The die is shaped like one of the five Platonic solids, which implies that the die has 4, 6, 8, 12, or 20 sides.

[*] Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology and Immunology, Department of Mathematics and Statistics, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, K1H 8M5, dbickel@uottawa.ca

https://doi.org/10.1214/19-BA1189

The die was constructed inside the box by an unknown mechanism that constructs shapes at random until it happens upon one that closely resembles a Platonic solid. To make Bayesian inferences about $\theta$, the number of sides of the die, we need the posterior probability that it has $\theta$ sides:

$$P\left(\theta|x\right) \propto P\left(\theta\right) \prod_{x=1}^{n} f_\theta\left(x_i\right), \tag{1}$$

where $P\left(\theta\right)$ is the prior probability that it has $\theta$ sides, and $f_\theta\left(x_i\right)$ is the probability that $x_i$ would be observed if it has $\theta$ sides. From the given information, each *data distribution* $f_\theta$ is a uniform distribution on $\{1, \ldots, \theta\}$, so that $f_\theta\left(i\right) = 1/\theta$ for $i = 1, \ldots, \theta$ but $f_\theta\left(i\right) = 0$ for $i = \theta + 1, \ldots, 20$. While it might be tempting to assign the uniform prior distribution such that $P\left(\theta\right) = 1/5$ for $\theta = 4, 6, 8, 12, 20$, that would not account for how many more coincidences it would take for the mechanism to generate a die with a higher number of sides than a lower number of sides. Incorporating that information means assigning more probability to simpler dice and less probability to more complex dice:

$$P\left(4\right) > P\left(6\right) > P\left(8\right) > P\left(12\right) > P\left(20\right).$$

Which prior distribution satisfying that constraint should be used? ▲

Such coincidences, occurrences of multiple improbable events without strong dependence on each other, are tacit in many other complex distributions (cf. White, 2005). A less Platonic example emphasizes the need to consider the simplicity of data distributions when assigning a prior.

**Example 2.** Inside a black box, an unknown mechanism randomly constructed one or more balls of different colors and placed them in an urn. The number of balls in the urn is the number that could be constructed within a short time window. If none were constructed within that time period, the process started over and continued until at least one ball was placed into the urn. The mechanism had access to a million colors. From the urn, $n$ balls will be drawn independently, with equal probability, and with replacement. The observer wants to make inferences about $\phi$, the set of the colors of the $|\phi|$ balls in the urn. That set differs from the *configuration* $\theta$, the $|\phi|$-tuple of the colors of the balls in the order in which they were placed into the urn, in that $\phi$ is an unordered set and $\theta$ is an ordered set or vector of the same number of colors. Since $\theta$ is a permutation of the members of $\phi$, the posterior probability that the set of colors of the balls in the urn is $\phi$ is the sum of each posterior probability that the configuration is $\theta$ over all permutations of the members of $\phi$. The latter posterior probability is given by (1) with $P\left(\theta\right)$ as the configuration is $\theta$ and with $f_\theta\left(x_i\right)$ as the probability that the $i$th ball drawn from the urn will be of color $x_i$, conditional on $\theta$ as the configuration. It may have been reasonable to assign a uniform prior distribution over $\Theta$, the configuration space, were it not for the information about how the urns were populated, information indicating the higher number of coincidences needed to populate an urn with more balls as opposed to fewer. That information, without being enough to determine the prior distribution, requires configurations with fewer balls to have higher prior probabilities than those with more balls: $P\left(\theta_1\right) > P\left(\theta_2\right)$ for all $\theta_1, \theta_2 \in \Theta$ such that $|\theta_1| < |\theta_2|$, where $|\theta|$ is the dimension of $\theta$. Subject to that constraint, what should the prior be? ▲

*Occam's razor* is the principle that simpler explanations are more credible than more complex explanations in the absence of evidence favoring more complex explanations. In a Bayesian framework with the number of free parameters in a model as the measure of complexity, that greater credibility may show up as a higher posterior probability (cf. Rosenkrantz, 1976) or, via the simplicity postulate (Jeffreys, 1948, pp. 100–101, 113, 222), as a higher prior probability (Jefferys and Berger, 1992). Multiple methodology researchers reached similar conclusions for other forms of complexity. Among others, Poston (2014) argues that complexity should constrain the prior distribution, with simpler explanations being at least as probable as more complex explanations. Explanations requiring more coincidences, while not impossible, tend to be less probable than those requiring fewer coincidences (Myrvold, 2017; Blanchard, 2018).

A special case of that type of constraint on prior distributions is seen in Examples 1 and 2. In both examples, each data distribution $f_\theta$ is uniform on some sample space $\mathcal{X}_\theta$ of a number of possible outcomes equal to $|\mathcal{X}_\theta| = \theta$ in Example 1 and $|\mathcal{X}_\theta| = |\theta|$ in Example 2. Also in both examples, the prior probability $P(\theta)$ decreases as a function of $|\mathcal{X}_\theta|$ since it reflects the number of coincidences that $f_\theta$ would require. Although $|\mathcal{X}_\theta|$ increases with the complexity of a uniform distribution, another measure of complexity is needed for other data distributions.

Entropy is a measure of complexity that generalizes the reasoning of Examples 1 and 2, for the entropy of a uniform distribution $f_\theta$ is $\log |\mathcal{X}_\theta|$. The conditions of Section 2 result in the constraint that parameter values corresponding to data distributions with lower entropy have higher prior probabilities than those of higher entropy. Although those conditions are not universally applicable, they provide the foundation for the more general methods of later sections.

Merely arranging parameter values in order of prior probability is not enough for Bayesian data analysis, as an ordering in itself does not determine a prior distribution. Starting with the ordering, Section 3 derives a method for transforming a preliminary prior distribution such as the uniform distributions of Examples 1 and 2 into a prior distribution informed by Occam's razor. The derivation is based on desirable properties of such a transformation.

That prior distribution, however, is only determined up to a parameter that controls the extent to which it differs from the preliminary prior. In applications requiring flexibility, the ability to set the parameter on a case-by-case basis may be desirable. In other applications, using a default value would save resources or reduce concerns about a potential conflict of interest. Section 4 derives such a value from an idealized model of constructing a data distribution, with more complex distributions being less probable because they require more coincidences to construct.

Because Shannon's entropy has a number of complexity-suitable properties that uniquely characterize it (e.g., Rényi, 1965), it is the measure of complexity emphasized in this paper. As an excursus, Section 5 explores alternative definitions of entropy as complexity. It notes that since all Rényi entropies are additive and have the property that the entropy of a uniform distribution $f_\theta$ is $\log |\mathcal{X}_\theta|$, any of them may replace the Shannon entropy.

Since most statistics applications involve probability distributions with infinite domains, the framework is generalized from finite parameter spaces to infinite parameter spaces in Section 6 and from finite sample spaces to infinite sample spaces in Section 7. The latter section applies the framework to null hypothesis significance testing.

Since the entropy of a whole sample is greater than that of a single observation, the relation to the sample size is specified in Section 8, which extends the framework to stochastic processes. In the usual case of a sample of $n$ IID observations from a distribution conditional on $\theta$, the sharpened prior probability is proportional to the unsharpened prior probability and inversely proportional to the exponential of the entropy or differential entropy of that distribution. Examples include both the IID processes of Examples 1–2 and some binomial and normal models that commonly occur in practice.

Section 9 closes with a discussion of which priors require simplicity adjustments.

## 2 Prior probabilities constrained by the simplicity of data distributions

The preliminary concepts of this section provide a foundation for generating prior distributions that satisfy a generalization of the simplicity conditions suggested by Examples 1 and 2. The subsequent sections build on this foundation.

The *entropy* of a probability mass function (PMF) $g$ on finite set $\mathcal{X}$ of possible observations is

$$S(g) = -\sum_{x\in\mathcal{X}} g(x)\ln g(x), \tag{2}$$

understood such that $0 \ln 0 = 0$. All data distributions are on the same sample space $\mathcal{X}$. That means the data distributions of Example 1, while uniform if restricted, are not uniform on $\mathcal{X} = \{1,\ldots,20\}$, with the exception of $f_{20}$, the distribution of outcomes of the 20-sided die. For reasons explained below, we also need a continuum between different uniform distributions on a finite sample space. For example, on the sample space $\mathcal{X} = \{1,2,3,4,5,6\}$, the PMF given by

$$g(x) = \begin{cases} 2/5 & \text{if } x = 1,2 \\ 1/5 & \text{if } x = 3 \\ 0 & \text{if } x = 4,5,6 \end{cases}, \tag{3}$$

is uniform on $\{1,2\}$ but with $x = 3$ having a probability mass between that of each $x$ in the main supported set $\{1,2\}$ and that of each $x$ in the non-supported set $\{4,5,6\}$:

$$g(4) = g(5) = g(6) = 0 < 1/5 = g(3) = 1/5 < 2/5 = g(1) = g(2). \tag{4}$$

To streamline development involving distributions like $g$, we need a term for them.

**Definition 1.** A PMF $g$ on $\mathcal{X}$ is called *partially uniform* if it meets these conditions:

1. It is uniform on $\mathcal{X}(g)$, a non-empty subset of $\mathcal{X}$. That is, there is a $g^{\max} > 0$ such that $g(x) = g^{\max}$ for all $x \in \mathcal{X}(g)$.

2. There is no more than one $y \in \mathcal{X}$ such that $0 < g(y) < g^{\max}$.

3. It otherwise has a value of 0. That is, $g(x) = 0$ for all $x \neq y$ in $\mathcal{X}$ but not in $\mathcal{X}(g)$ if $y$ exists; otherwise, $g(x) = 0$ for all $x \in \mathcal{X} \backslash \mathcal{X}(g)$.

Condition 3 requires $g$ to have a value of 0 outside its support. Loosely speaking, conditions 1–2 require $g$ to be as uniform as possible on its support.

More precisely, condition 2 requires that, outside of $\mathcal{X}(g)$, which is the uniform portion of $g$'s domain, $g$ is only allowed to have at most one component deviating from the uniformity specified by condition 1. It allows one component ("$y$") to have lower probability than the components on $\mathcal{X}(g)$, as seen in (4), in which $y = 3$. The reason to allow that departure from uniformity is to create a continuum of PMFs between the uniform distribution on the whole support and the uniform distribution on the support without $y$. That continuum is needed in the next paragraph to generalize the cardinality of a sample space's support from a counting number to a positive real number. Without the continuum and its real-valued generalization of cardinality, the measure of complexity to be derived would be unnecessarily burdened with discontinuities to keep track of. In short, a little inelegance in Definition 1 will pay off in the elegance of the result.

For a continuous functional $I$ defined as follows, the *intricacy* of a partially uniform $g$ is $I(g)$. If $y$ does not exist, then $I(g) = |\mathcal{X}(g)|$, the cardinality of $g$'s support. If $y$ exists, then $I(g) = I_{|\mathcal{X}(g)|}(g(y))$, where $I_{|\mathcal{X}(g)|}$ is a strictly monotonic increasing, continuous function on $]0,1[$ such that $\lim_{q \to 0} I_{|\mathcal{X}(g)|}(q) = |\mathcal{X}(g)|$ and $\lim_{q \to 1} I_{|\mathcal{X}(g)|}(q) = |\mathcal{X}(g)| + 1$. For example, the $g$ of (3) is of intricacy $I(g) = I_{|\{1,2\}|}(1/5) = I_2(1/5)$, which is a non-integer between 2 and 3. That is because $g(y) = g(3) = 1/5$ is between 0 and $g(1) = g(2) = 2/5$, as per (4). Were $g(1) = g(2) = 1/2$ and $g(3) = 0$, then the intricacy would instead be $|\{1,2\}| = 2$, whereas were $g(1) = g(2) = g(3) = 1/3$, then the intricacy would instead be $|\{1,2,3\}| = 3$. In that way, the intricacy of a partially uniform PMF generalizes the cardinality of its support to a continuum of non-integer values.

A *Bayesian model* is a pair $(\theta \mapsto f_\theta, P)$, where $\theta \mapsto f_\theta$, abbreviated as $f_\bullet$, is a function on $\Theta$ such that, for every $\theta \in \Theta$, $f_\theta$ is a (data) PMF on $\mathcal{X}$ and $P$ is a (prior) PMF $P$ on $\Theta$. Let $\mathcal{M}(\mathcal{X}, \Theta)$ denote the set of all Bayesian models with $\mathcal{X}$ as the sample space and $\Theta$ as the parameter space.

What it means for a prior distribution to be constrained by the simplicity of the sampling distributions uses entropy as a generally applicable measure of complexity and intricacy as a measure of complexity that only applies to partially uniform distributions.

**Definition 2.** Let $\mathcal{M}^S$ denote a subset of $\mathcal{M}(\mathcal{X}, \Theta)$. $\mathcal{M}^S$ is called a *set of Bayesian models with simplicity-constrained PMFs* if these conditions are satisfied:

1. There exists a function $p$ such that, for every $(f_\bullet, P) \in \mathcal{M}^S$,

$$P = p\left(S_{f_\bullet}\right), \tag{5}$$

where $S_{f_\bullet}$ is the function on $\Theta$ defined to satisfy $S_{f_\bullet}(\theta) = S(f_\theta)$ for all $\theta \in \Theta$. The function $p$ is called the *prior generator* for $\mathcal{M}^S$, and the function $S_{f_\bullet}$ is called the *entropy spectrum* of $f_\bullet$.

2. For every Bayesian model $(f_\bullet, P)$ such that $(f_\bullet, P) \in \mathcal{M}(\mathcal{X}, \Theta)$ and such that $f_\theta$ is partially uniform for every $\theta \in \Theta$,
$$P(\theta_1) \geq P(\theta_2) \iff I(f_{\theta_1}) \leq I(f_{\theta_2}) \tag{6}$$
for all $\theta_1, \theta_2 \in \Theta$ and $(f_\bullet, P) \in \mathcal{M}^S$.

While (5) says the prior distribution is a function of the entropy spectrum, (6) says parameter values labeling less intricate partially uniform distributions have higher prior probabilities. The rationale is that, in the absence of other information, uniform distributions on larger domains tend to require more coincidences and thus to be less probable than those on smaller domains, as seen in Examples 1 and 2.

The result is that simpler data PMFs tend to have higher prior probabilities.

**Lemma 1.** *If $\mathcal{M}^S$ is a set of Bayesian models with simplicity-constrained PMFs, then the prior generator for $\mathcal{M}^S$ is constrained such that each $(f_\bullet, P) \in \mathcal{M}^S$ satisfies*
$$P(\theta_1) \geq P(\theta_2) \iff S(f_{\theta_1}) \leq S(f_{\theta_2}) \tag{7}$$
*for all $\theta_1, \theta_2 \in \Theta$.*

*Proof.* If a PMF $g$ is partially uniform, then
$$\begin{aligned} S(g) &= \begin{cases} -\sum_{x \in \mathcal{X}(g)} \frac{1}{|\mathcal{X}(g)|} \ln \frac{1}{|\mathcal{X}(g)|} & \text{if } \forall x \in \mathcal{X}\, g(x) \in \left\{0, \frac{1}{|\mathcal{X}(g)|}\right\} \\ -|\mathcal{X}(g)|\, g^{\max} \ln g^{\max} - g(y) \ln g(y) & \text{if } \exists y \in \mathcal{X}\, 0 < g(y) < g^{\max} \end{cases} \\ &= \begin{cases} \ln |\mathcal{X}(g)| & \text{if } \forall x \in \mathcal{X}\, g(x) \in \left\{0, \frac{1}{|\mathcal{X}(g)|}\right\} \\ \ln \left( (g^{\max})^{|\mathcal{X}(g)| g^{\max}} \left(\frac{1}{g(y)}\right)^{g(y)} \right) & \text{if } \exists y \in \mathcal{X}\, 0 < g(y) < g^{\max} \end{cases} . \end{aligned} \tag{8}$$

Consider a Bayesian model $(f_\bullet, P) \in \mathcal{M}^S$ such that $f_\theta$ is partially uniform and $\forall x \in \mathcal{X}\, f_\theta(x) \in \{0, 1/|\mathcal{X}(f_\theta)|\}$ for every $\theta \in \Theta$. Since in that case $I(f_\theta) = |\mathcal{X}(f_\theta)|$ for every $\theta \in \Theta$, (6) and (8) imply that (7) holds for all $\theta_1, \theta_2 \in \Theta$.

Now consider instead a Bayesian model $(f_\bullet, P) \in \mathcal{M}^S$ such that $f_\theta$ is partially uniform and $\forall x \in \mathcal{X}\, f_\theta(x) \in \{0, 1/|\mathcal{X}(f_\theta)|\}$ for every $\theta \in \Theta$ except $\theta(y)$ for some $y \in \mathcal{X}$ such that $0 < f_{\theta(y)}(y) < f_{\theta(y)}^{\max}$, where $f_{\theta(y)}^{\max} = \max_{x \in \mathcal{X}} f_{\theta(y)}(x)$. Equations (6) and (8) would be satisfied if $I(f_\theta) = \exp(S(f_\theta))$ for every $\theta \in \Theta$. In fact, since such a Bayesian model is in $\mathcal{M}^S$ by part 2 of Definition 2 for every real value of $f_{\theta(y)}(y)$ strictly between 0 and 1, the functions $I$ and $S$ must be isomorphic on the domain of $I$ by its monotonicity and continuity properties. Thus, (5) constrains the prior generator $p$ such that its PMF assignments satisfy (7) holds for all $\theta_1, \theta_2 \in \Theta$.

Since every possible entropy spectrum $S_{f_\bullet}$ is achieved by some Bayesian model $(f_\bullet, P)$ such that $f_\theta$ is partially uniform for every $\theta \in \Theta$, and since every such model is in $\mathcal{M}^S$ (Definition 2, part 2), it follows that $p$ is completely determined by $I$ in such a way that (7) holds for all $\theta_1, \theta_2 \in \Theta$. That same $p$ is the prior generator not only for those Bayesian models but for all Bayesian models in $\mathcal{M}^S$ by part 1 of Definition 2. It follows that (7) holds just as generally. □

## 3 Adjusting prior probabilities for the simplicity of data distributions

The method of this section transforms a prior distribution that does not account for the simplicity of the data distributions into a prior that does. That prior satisfies the constraints of Section 2 in the special case that the pre-transformation prior is uniform. More generally, any prior on a finite parameter space may serve as the pre-transformation prior, regardless of how far it deviates from uniformity. That is how this section paves the way for the extensions to priors for more general parameter spaces in Section 6, including commonly used objective priors such as those of Section 8's Examples 6 and 7.

**Definition 3.** Let $\bullet^{\sharp}$ denote a function that transforms a PMF to another PMF on the same parameter space and that necessarily satisfies the following conditions for a prior generator $p$. That function is called a *sharpener*. Given any PMF $P$, its *sharpened counterpart* is $P^{\sharp}$, that is, $\bullet^{\sharp}$ evaluated at $P$. **Conditions:**

1. *Simplicity constraint.* The sharpened counterpart $P^{\sharp}$ of a uniform PMF $P$ is a simplicity-constrained PMF generated by $p$.
2. *Coherence.* The sharpened counterpart $P\left(\bullet|X=x\right)^{\sharp}$ of a posterior PMF $P\left(\bullet|x\right)$ based on a prior PMF $P$ is $P^{\sharp}\left(\bullet|X=x\right)$, the posterior distribution based on $P^{\sharp}$, the sharpened counterpart of $P$, where $x$ is an observed sample.
3. *Independence preservation.* Consider the finite parameter sets $\Theta$ and $\Phi$. Suppose that, for all $\theta\in\Theta$ and $\phi\in\Phi$, $X\sim f_{\theta}$ and $Y\sim g_{\phi}$ are independent random variables of joint PMF $h_{\theta,\phi}$ with values in $\mathcal{X}$ and $\mathcal{Y}$, where $f_{\theta}$ and $g_{\phi}$ are PMFs on $\mathcal{X}$ and $\mathcal{Y}$, respectively. If $P^{\sharp}$ is the sharpened counterpart of a PMF $\left(\theta,\phi\right)\mapsto P\left(\theta,\phi\right)=P_{1}\left(\theta\right)P_{2}\left(\phi\right)$ that is the joint PMF of independent parameters $\theta$ and $\phi$ that have prior PMFs $P_{1}$ and $P_{2}$ and their sharpened counterparts $P_{1}^{\sharp}$ and $P_{2}^{\sharp}$, respectively, then $P^{\sharp}\left(\theta,\phi\right)=P_{1}^{\sharp}\left(\theta\right)P_{2}^{\sharp}\left(\phi\right)$ for all $\theta\in\Theta$ and $\phi\in\Phi$.

The simplicity constraint (condition 1) builds Definition 3 on the foundation laid in Section 2. The coherence condition (condition 2) means considering simplicity commutes with conditioning on the observed data so that it does not matter which happens first. Independence preservation (condition 3) means that if two quantities have nothing to do with each other, then that should be reflected in their priors adjusted for simplicity.

**Theorem 1.** *Let $\left(f_{\bullet},P\right)$ denote a Bayesian model. If $P^{\sharp}$ is the sharpened counterpart of $P$ on a parameter set $\Theta$, then there is a $\kappa>0$ such that*

$$P^{\sharp}\left(\theta\right)=\frac{P\left(\theta\right)e^{-\kappa\,S\left(f_{\theta}\right)}}{\sum_{\theta'\in\Theta}P\left(\theta'\right)e^{-\kappa\,S\left(f_{\theta'}\right)}} \tag{9}$$

*for all $\theta\in\Theta$.*

*Proof.* Let $f_{\theta}$, $g_{\phi}$, $h_{\theta,\phi}$, $P_{1}$, $P_{2}$, and $P$ denote PMFs that satisfy the independence assumptions of Condition 3, and assume $P_{1}$ and $P_{2}$ are uniform. Since $P$, $P_{1}$, and $P_{2}$ are uniform, the simplicity constraint (Condition 1) requires that $P^{\sharp}$, $P_{1}^{\sharp}$, and $P_{2}^{\sharp}$ are simplicity-constrained PMFs generated by the same prior generator $p$. By (5), $P_1^\sharp = p\left(S_{f_\bullet}\right)$, $P_2^\sharp = p\left(S_{g_\bullet}\right)$, and $P^\sharp = p\left(S_{h_{\bullet,\bullet}}\right)$. According to Lemma 7, they satisfy 7 for the same prior generator $p$. Thus, there is a strictly monotonic decreasing function $q$ such that $P_1^\sharp(\theta) \propto q\left(S\left(f_\theta\right)\right)$, $P_2^\sharp(\phi) \propto q\left(S\left(g_\phi\right)\right)$, and $P^\sharp(\theta,\phi) \propto q\left(S\left(h_{\theta,\phi}\right)\right)$ for all $\theta \in \Theta$ and $\phi \in \Phi$. According to the independence preservation condition, $P^\sharp(\theta,\phi) = P_1^\sharp(\theta)\, P_2^\sharp(\phi)$ for all $\theta \in \Theta$ and $\phi \in \Phi$. Thus, there is a real number $c$ such that

$$\ln q\left(S\left(h_{\theta,\phi}\right)\right) = \ln q\left(S\left(f_\theta\right)\right) + \ln q\left(S\left(g_\phi\right)\right) + c,$$

which, with the property that $S\left(h_{\theta,\phi}\right) = S\left(f_\theta\right) + S\left(g_\phi\right)$ for independent random variables, implies that there are real numbers $a$ and $b$ such that $\ln q(\bullet) = a \times \bullet + b$, where $a < 0$ since $q$ is strictly monotonic decreasing, and $b$ may differ between $f_\theta$, $g_\phi$, and $h_{\theta,\phi}$.

It follows that, even when the independence assumptions are not satisfied, $\ln P(\theta|X=x)^\sharp = a\, S\left(f_\theta\right)$ up to a constant term for every uniform posterior PMF $P(\bullet|X=x)$ according to the simplicity constraint. Letting $\kappa = -|a|$,

$$P(\theta|X=x)^\sharp \propto e^{-\kappa\, S(f_\theta)}.$$

As a posterior distribution, $P(\theta|X=x) \propto P(\theta)\, f_\theta(x)$ for all $\theta \in \Theta$ by Bayes's theorem. Since $P(\bullet|X=x)$ is uniform, $P(\theta) \propto 1/f_\theta(x)$. Coherence (Condition 2) then gives

$$P^\sharp(\theta)\, f_\theta(x) \propto P^\sharp(\theta|X=x) = P(\theta|X=x)^\sharp \propto e^{-\kappa\, S(f_\theta)},$$

where the first proportionality results from another application of Bayes's theorem. Thus,

$$P^\sharp(\theta) \propto e^{-\kappa\, S(f_\theta)}/f_\theta(x) \propto P(\theta)\, e^{-\kappa\, S(f_\theta)},$$

which ensures that the more generally applicable sharpener $\bullet^\sharp$ has the form of (9). □

A prior would require sharpening whenever it neglects relevant information about the simplicity of the data distributions.

The value of $\kappa$ is called the *sharpness* of the sharpener, which may now be written as $\bullet^{\sharp\kappa}$ to distinguish it from other sharpeners. Each sharpener corresponds to a different sharpened prior distribution, as will be seen in Figure 1 of Example 3. The application to real data may suggest a way to specify the value of $\kappa$ in some cases. In other cases, a default value is needed.

# 4 How much should priors be adjusted for simplicity by default?

The method of Section 3 cannot be applied without somehow specifying a value of $\kappa$, the degree to which priors are adjusted for the simplicity of the data distributions. This section argues for a default setting of $\kappa = 1$.

**Definition 4.** Let $\bullet^{\sharp\kappa^\star}$ denote a sharpener of sharpness $\kappa^\star > 0$ with the following constraint. For any Bayesian model $\left(f_\bullet^\star, P^\star\right)$ such that $f_\theta^\star$ is partially uniform, $\forall x \in \mathcal{X}\; f_\theta^\star(x) \in \left\{0, 1/\left|\mathcal{X}\left(f_\theta^\star\right)\right|\right\}$ for every $\theta \in \Theta$, and there is an $x^\star \in \mathcal{X}$ such that $x^\star \in \mathcal{X}\left(f_\theta^\star\right)$

for every $\theta \in \Theta$ and such that the sharpened counterpart $P^{\star\sharp\kappa^{\star}}$ of $P^{\star}$ is the conditional PMF given by

$$P^{\star\sharp\kappa^{\star}}(\theta) = \mathrm{Prob}_{\vartheta\sim P^{\star}, X\sim f_{\vartheta}}(\vartheta = \theta | X = x^{\star}) \tag{10}$$

for all $\theta \in \Theta$. Then, given any Bayesian model $(f_{\bullet}, P)$, the sharpened counterpart $P^{\sharp\kappa^{\star}}$ of $P$ is *ideal*, whether or not $P$ meets the conditions for $P^{\star}$.

Thus, the universal ideal $\kappa^{\star}$ may be found by conditioning on successfully generating the correct realization, with the unsharpened PMF as the distribution of opportunities to attempt a correct realization. In Example 1, that means successfully constructing a die of a certain number of sides, whereas in Example 2, it means successfully constructing an urn with the specified configuration of colors.

**Theorem 2.** *For any Bayesian model $(f_{\bullet}, P)$, the ideal sharpened counterpart $P^{\sharp\kappa^{\star}}$ of $P$ satisfies*

$$P^{\sharp\kappa^{\star}}(\theta) = P^{\sharp 1}(\theta) = \frac{P(\theta)\, e^{-S(f_{\theta})}}{\sum_{\theta'\in\Theta} P(\theta')\, e^{-S(f_{\theta'})}} \tag{11}$$

*for all $\theta \in \Theta$.*

*Proof.* By (10) and the conditions on $f_{\bullet}^{\star}$,

$$\begin{aligned} P^{\star\sharp\kappa^{\star}}(\theta) &\propto \mathrm{Prob}_{\vartheta\sim P^{\star}}(\vartheta = \theta)\,\mathrm{Prob}_{\vartheta\sim P^{\star}}(X = x^{\star}|\vartheta = \theta) \\ &= P^{\star}(\theta)\, f_{\theta}^{\star}(x) = P^{\star}(\theta) / \left|\mathcal{X}(f_{\theta}^{\star})\right| \\ &= P^{\star}(\theta)\, e^{-\ln\left|\mathcal{X}(f_{\theta}^{\star})\right|} = P^{\star}(\theta)\, e^{-S(f_{\theta}^{\star})}. \end{aligned}$$

That only agrees with (9) if $\kappa^{\star} = 1$. Thus, $P^{\sharp\kappa^{\star}} = P^{\sharp 1}$ for any Bayesian model $(f_{\bullet}, P)$, and the right-hand-side of (11) results from the relevant special case of (9). □

Similar results may be derived from fewer assumptions using the concept of Rényi entropy (Section 5).

An alternative derivation of (11) appears in Bickel (2016). Instead of conditioning on the event that a data distribution is constructed correctly, it conditions on the event that a randomly typed computer program yields output representing the data distribution.

## 5 Excursus: Rényi entropy as a measure of complexity

For any $\alpha > 0$, the *$\alpha$-Rényi entropy* of a probability mass function (PMF) $g$ on finite set $\mathcal{X}$ of possible observations is

$$S_{\alpha}(g) = -\ln\left(\sum_{x\in\mathcal{X}} g(x)\, g^{\alpha-1}(x)\right)^{\frac{1}{\alpha-1}} \tag{12}$$

if $\alpha \neq 1$ and is the Shannon entropy given by (2) if $\alpha = 1$. Thus, if $g$ is uniform, then

$$S_{\alpha}(g) = -\ln\left(\frac{|\mathcal{X}|}{|\mathcal{X}|}\left(\frac{1}{|\mathcal{X}|}\right)^{\alpha-1}\right)^{\frac{1}{\alpha-1}} = \ln|\mathcal{X}|$$

for all $\alpha \neq 1$. With that property and additivity under independence, substituting $S_\alpha$ for $S_1$ throughout the paper would yield analogous results for any other Rényi entropy.

In Section 4, the Shannon entropy was derived as a component of the ideal sharpened prior given assumptions including one about the coincidences involved in constructing a data distribution. A Rényi entropy and a limiting case of Rényi entropy can be derived from fewer assumptions, as follows.

$S_2(g)$, the *quadratic entropy*, also called the "collision entropy" (Teixeira et al., 2012), is related to the prior distribution obtained if the difficulty of constructing a data distribution is modeled in terms of the coincidence that two independent realizations collide with each other.

**Definition 5.** Given any Bayesian model $(f_\bullet, P)$, *the collision prior PMF corresponding to $P$* is

$$P^{\mathrm{coll}}(\theta) = \mathrm{Prob}_{\vartheta\sim P, X, X'\sim f_\vartheta}(\vartheta = \theta | X = X'),$$

as a function of $\theta$, where $X$ and $X'$ are IID.

**Theorem 3.** *The collision prior PMF corresponding to $P$ satisfies*

$$P^{\mathrm{coll}}(\theta) = \frac{P(\theta)\, e^{-S_2(f_\theta)}}{\sum_{\theta'\in\Theta} P(\theta')\, e^{-S_2(f_{\theta'})}}$$

*for all $\theta \in \Theta$.*

*Proof.* By Bayes's theorem with $X'$ as data,

$$\begin{aligned} P^{\mathrm{coll}}(\theta) &\propto \mathrm{Prob}_{\vartheta\sim P}(\vartheta = \theta)\,\mathrm{Prob}_{\vartheta\sim P, X, X'\sim f_\vartheta}(X = X' | \vartheta = \theta) \\ &= P(\theta)\sum_{x\in\mathcal{X}} f_\theta(x)\, f_\theta(x) = P(\theta)\, e^{-S_2(f_\theta)}, \end{aligned}$$

where the substitution of $e^{-S_2(f_\theta)}$ for $\sum_{x\in\mathcal{X}} f_\theta(x)\, f_\theta(x)$ is sanctioned by (12) with $\alpha = 2$. □

A limiting case of Rényi entropy that is important in cryptography is the *min-entropy* (Teixeira et al., 2012),

$$S_\infty(g) = \lim_{\alpha\to\infty} S_\alpha(g) = -\ln \max_{x\in\mathcal{X}} g(x) = \min_{x\in\mathcal{X}} \ln \frac{1}{g(x)}. \tag{13}$$

It is related to the prior distribution obtained if the difficulty of constructing a data distribution is modeled in terms of the coincidence of a successful prediction using the best predicted observation given the data distribution.

**Definition 6.** Given any Bayesian model $(f_\bullet, P)$, *the prediction prior PMF corresponding to $P$* is

$$P^{\mathrm{pred}}(\theta) = \mathrm{Prob}_{\vartheta\sim P, X\sim f_\vartheta}\left(\vartheta = \theta | X = x_\theta^{\mathrm{pred}}\right),$$

as a function of $\theta$, where $x_\theta^{\mathrm{pred}} = \arg\max_{x\in\mathcal{X}} f_\theta$.

**Theorem 4.** *The prediction prior PMF corresponding to $P$ satisfies*

$$P^{\text{pred}}(\theta) = \frac{P(\theta)\, e^{-S_\infty(f_\theta)}}{\sum_{\theta' \in \Theta} P(\theta')\, e^{-S_\infty(f_{\theta'})}}$$

*for all $\theta \in \Theta$.*

*Proof.* By Bayes's theorem with $x_\theta^{\text{pred}}$ as data,

$$\begin{aligned} P^{\text{pred}}(\theta) &\propto \text{Prob}_{\vartheta \sim P}(\vartheta = \theta)\, \text{Prob}_{\vartheta \sim P, X \sim f_\vartheta}\left(X = x_\theta^{\text{pred}} | \vartheta = \theta\right) \\ &= P(\theta)\, f_\theta\left(x_\theta^{\text{pred}}\right) = P(\theta)\, e^{-S_\infty(f_\theta)}, \end{aligned}$$

where the substitution of $e^{-S_\infty(f_\theta)}$ for $f_\theta(x_\theta^{\text{pred}})$ is permitted by (13). □

## 6 Adjusting prior densities for the simplicity of data distributions

The extension of sharpened prior PMFs to general sharpened priors, while requiring more notation, is straightforward. It says sharpened prior probability distributions and sharpened probability density function (PDFs) project to sharpened prior PMFs on all finite partitions the parameter space, which need not be finite.

**Definition 7.** Given any $\kappa > 0$ and a probability measure $\Pi$ on a measure space $(\Theta, \mathfrak{F})$, the probability measure $\Pi^{\sharp\kappa}$ on $(\Theta, \mathfrak{F})$ is the *$\kappa$-sharpened counterpart* of $\Pi$ if

$$\forall \mathcal{N} \in \mathfrak{F}'\ \Pi^{\sharp\kappa}(\mathcal{N}) = P_{\Pi,\mathfrak{F}'}^{\sharp\kappa}(\mathcal{N}) \tag{14}$$

for every finite partition $\mathfrak{F}' \subset \mathfrak{F}$ of $\Theta$, where each $P_{\Pi,\mathfrak{F}'}$ is the PMF on $\mathfrak{F}'$ such that

$$\forall \mathcal{N} \in \mathfrak{F}'\ P_{\Pi,\mathfrak{F}'}(\mathcal{N}) = \Pi(\mathcal{N}). \tag{15}$$

For every measure $\nu$ that dominates $\Pi$, the PDF $d\Pi^{\sharp\kappa}/d\nu$ is the *$\kappa$-sharpened counterpart* of the PDF $d\Pi/d\nu$.

The differential element "$d\nu(\theta')$" in the next result may be read as "$d\theta'$" in the usual case that the dominating measure is uniform on the real line.

**Corollary 1.** *The $\kappa$-sharpened counterpart of any continuous PDF $\pi$ on $\Theta$ satisfies*

$$\pi^{\sharp\kappa}(\theta) = \frac{\pi(\theta)\, e^{-\kappa S(f_\theta)}}{\int \pi(\theta')\, e^{-\kappa S(f_{\theta'})} d\nu(\theta')}, \tag{16}$$

*where $\nu$ is the measure that defines $\pi$ as $d\Pi/d\nu$ for some probability measure $\Pi$ on $(\Theta, \mathfrak{F})$ that is dominated by $\nu$, almost surely with respect to $\Pi$.*

*Proof.* By the definition of the $\kappa$-sharpened counterpart of any PDF $\pi$ and by the definition of a PDF,

$$\Pi^{\sharp\kappa}\left(\mathcal{N}\right)=\int_{\mathcal{N}}\pi^{\sharp\kappa}\left(\theta\right)d\nu\left(\theta\right)$$

for all $\mathcal{N}\in\mathfrak{F}$, where $\Pi^{\sharp\kappa}$ is the $\kappa$-sharpened counterpart of $\Pi$. Thus, from Theorem 1 and (14)–(15),

$$\begin{aligned}\forall\mathcal{N}\in\mathfrak{F}'\int_{\mathcal{N}}\pi^{\sharp\kappa}\left(\theta\right)d\nu\left(\theta\right)&=\frac{P_{\Pi,\mathfrak{F}'}\left(\mathcal{N}\right)e^{-\kappa\,S\left(f_{\mathcal{N}}\right)}}{\sum_{\mathcal{N}'\in\mathfrak{F}'}P_{\Pi,\mathfrak{F}'}\left(\mathcal{N}'\right)e^{-\kappa\,S\left(f_{\mathcal{N}'}\right)}}\\&=\frac{\Pi\left(\mathcal{N}\right)e^{-\kappa\,S\left(f_{\mathcal{N}}\right)}}{\sum_{\mathcal{N}'\in\mathfrak{F}'}\Pi\left(\mathcal{N}'\right)e^{-\kappa\,S\left(f_{\mathcal{N}'}\right)}}\\&=\frac{\int_{\mathcal{N}}\pi\left(\theta\right)d\nu\left(\theta\right)e^{-\kappa\,S\left(f_{\mathcal{N}}\right)}}{\sum_{\mathcal{N}'\in\mathfrak{F}'}\left(\int_{\mathcal{N}'}\pi\left(\theta\right)d\nu\left(\theta\right)\right)e^{-\kappa\,S\left(f_{\mathcal{N}'}\right)}}\end{aligned}$$

for every finite partition $\mathfrak{F}'\subset\mathfrak{F}$ of $\Theta$, where $x\mapsto f_{\mathcal{N}}\left(x\right)\propto\int_{\mathcal{N}}\pi\left(\theta\right)f_{\theta}\left(x\right)d\nu\left(\theta\right)$ for all $\mathcal{N}\in\mathfrak{F}'$. In short, for every $\mathfrak{F}'$,

$$\int_{\mathcal{N}}\pi^{\sharp\kappa}\left(\theta\right)d\nu\left(\theta\right)\propto\int_{\mathcal{N}}\pi\left(\theta\right)d\nu\left(\theta\right)e^{-\kappa\,S\left(f_{\mathcal{N}}\right)}$$

for all $\mathcal{N}\in\mathfrak{F}'$. According to the mean value theorem, the continuity of both $\pi$ and $\theta\mapsto S\left(f_{\theta}\right)$ requires that there is a $\theta\in\mathcal{N}$ such that $\pi^{\sharp\kappa}\left(\theta\right)\propto\pi\left(\theta\right)e^{-\kappa\,S\left(f_{\mathcal{N}}\right)}$. That can only be the case for every arbitrarily small $\mathcal{N}$ if $\pi^{\sharp\kappa}\left(\theta\right)\propto\pi\left(\theta\right)e^{-\kappa\,S\left(f_{\theta}\right)}$ holds up to a $\Pi$-null set. Since $\pi^{\sharp\kappa}$, as a PDF, integrates to 1, it follows that (16) holds up to a $\Pi$-null set. □

Equation (16) reveals that sharpened prior distributions are not alternatives to elicited priors, objective Bayes priors such as reference priors, or other priors in the literature, for it requires as input $\pi$, the unsharpened prior density. On the contrary, any prior density might serve as $\pi$, as will be illustrated for commonly used default priors in Examples 6 and 7.

## 7 Adjusting priors for the simplicity of data PDFs

Since infinite sample spaces are valuable only as approximations of finite sample spaces (e.g., Evans, 2015, Appendix A), previous sections used finite sample spaces to determine how to adjust prior distributions of parameters to reflect the simplicity of each data distribution indexed by a parameter value. The results are now extended to infinite sample spaces, as anticipated in Bickel (2016), following Cover and Thomas (2006, §8.3).

The technical tools to accomplish that are relative entropy and the convergence of sample spaces of increasing cardinality. The *relative entropy* function $D$ has values equal to $D\left(\mu\,||\,\nu\right)=\int d\mu\log\left(d\mu/d\nu\right)$, the entropy of a probability measure $\mu$ relative to a measure $\nu$ that dominates $\mu$ (e.g., Maas, 2017). Complementary statistical applications of relative entropy include the prevention of overfitting models (Fúquene et al., 2016; Gelman et al., 2017), the idealization of Cromwell's rule for revising priors (Bickel, 2018), and the automatic construction of unsharpened priors (Section 8, Example 7).

"The convergence of sample spaces" refers to the weak convergence of the data distributions and other measures defined on them. Lemmas 2–3 both involve a sequence of finite sample spaces that approach $\mathcal{X}^{(1)}$, a countably infinite sample space. The (1) in the superscript indicates that 1 is the lowest possible distance between members of $\mathcal{X}^{(1)}$. Lemma 3 additionally uses a sequence of countably infinite sample spaces that approach $\mathcal{X}^{(0)}$, an uncountably infinite sample space. The (0) in the superscript indicates that the members of $\mathcal{X}^{(0)}$ may be arbitrarily close to each other.

**Lemma 2.** *Let $\mathcal{X}^{(1)}$ denote a set of integers and $\mu^{(1)}$ a probability measure on the power set (set of subsets) of $\mathcal{X}^{(1)}$ such that $\mu^{(1)}$ is dominated by $\lambda^{(1)}$, the counting measure on $\mathcal{X}^{(1)}$. Let $\mathcal{X}^{(1),1}, \mathcal{X}^{(1),2}, \dots$ denote a sequence of finite sets such that the sequence $\lambda^{(1),1}, \lambda^{(1),2}, \dots.$ of their counting measures converges weakly to $\lambda^{(1)}$, abbreviated by $\lambda^{(1),m} \xrightarrow{weak} \lambda^{(1)}$ as $m \to \infty$. Let $\mu^{(1),m}(\bullet) = \mu^{(1)}\left(\bullet | \mathcal{X}^{(1),m}\right)$. Then*

$$\lim_{m\to\infty} D\left(\mu^{(1),m} \,||\, \lambda^{(1),m}\right) = D\left(\mu^{(1)} \,||\, \lambda^{(1)}\right).$$

*Proof.* Since $D$ is a continuous function and since both $\lambda^{(1),m} \xrightarrow{\text{weak}} \lambda^{(1)}$ and $\mu^{(1),m} \xrightarrow{\text{weak}} \mu^{(1)}$ as $m \to \infty$, the claim follows. □

The "(1)" in the superscripts of Lemma 2 roughly corresponds to $\Delta = 1$ in the "$\Delta$" superscripts of Lemma 3.

**Lemma 3.** *Let $\mathcal{X}^{(0)}$ denote a measurable subset of the real line and $\mu$ a probability measure on the $\sigma$-field generated by the Borel subsets of $\mathcal{X}^{(0)}$ such that $\mu$ is dominated by $\lambda$, the Lebesgue measure on $\mathcal{X}^{(0)}$, and such that the probability density function $d\mu/d\lambda$ is continuous on $\mathcal{X}^{(0)}$. Consider $\mathcal{X}^{\Delta} = \mathcal{X}^{(0)} \cap \{\dots, -2\Delta, -\Delta, 0, \Delta, 2\Delta, \dots\}$, the probability measure $\mu^{\Delta}(\bullet) := \mu\left(\bullet | \mathcal{X}^{\Delta}\right)$ on the power set (set of all subsets) of $\mathcal{X}^{\Delta}$, and the normalized counting measure $\lambda^{\Delta}$ on $\mathcal{X}^{\Delta}$ for each $\Delta > 0$, where the normalization multiplies each counting measure by a constant such that $\lambda^{\Delta} \xrightarrow{weak} \lambda$ as $\Delta \to 0$. For every $\Delta > 0$, let $\mathcal{X}^{\Delta,1}, \mathcal{X}^{\Delta,2}, \dots$ denote a sequence of finite sets such that the sequence $\lambda^{\Delta,1}, \lambda^{\Delta,2}, \dots.$ of their normalized counting measures converges weakly to $\lambda^{\Delta}$, abbreviated by $\lambda^{\Delta,m} \xrightarrow{weak} \lambda^{\Delta}$ as $m \to \infty$. For every $\Delta > 0$ and $m = 1, 2, \dots,$ let $\mu^{\Delta,m}(\bullet) = \mu^{\Delta}\left(\bullet | \mathcal{X}^{\Delta,m}\right)$. Then*

$$\lim_{\Delta\to 0}\lim_{m\to\infty} D\left(\mu^{\Delta,m} \,||\, \lambda^{\Delta,m}\right) = D\left(\mu \,||\, \lambda\right).$$

*Proof.* Consider any $\Delta > 0$. Since $D$ is a continuous function and since both $\lambda^{\Delta,m} \xrightarrow{\text{weak}} \lambda^{\Delta}$ and $\mu^{\Delta,m} \xrightarrow{\text{weak}} \mu^{\Delta}$ as $m \to \infty$, it follows that

$$\lim_{m\to\infty} D\left(\mu^{\Delta,m} \,||\, \lambda^{\Delta}\right) = D\left(\mu^{\Delta} \,||\, \lambda^{\Delta}\right).$$

Likewise, since $D$ is continuous and since $\mu^{\Delta} \xrightarrow{\text{weak}} \mu$ and $\lambda^{\Delta} \xrightarrow{\text{weak}} \lambda$ as $\Delta \to 0$,

$$\lim_{\Delta \to 0} D\left(\mu^{\Delta} \,||\, \lambda^{\Delta}\right) = D\left(\mu \,||\, \lambda\right). \qquad \square$$

With either $\mathcal{X}^{(1)}$ or $\mathcal{X}^{(0)}$ as the limiting sample space, sharpened prior distributions may now be defined.

**Definition 8.** Consider a $\kappa > 0$ and probability measures $\Pi$ and $\Pi^{\sharp\kappa}$ on a measure space $(\Theta, \mathfrak{F})$. If the probability measure $\Pi^{\sharp\kappa}_{(1),m}$ on $(\Theta, \mathfrak{F})$ is the $\kappa$-sharpened counterpart of $\Pi$ with $\mathcal{X}^{(1),m}$ as the sample space for all $m = 1, 2, \ldots$, if $\Pi^{\sharp\kappa}_{(1),m} \xrightarrow{\text{weak}} \Pi^{\sharp\kappa}$ as $m \to \infty$, and if the convergence conditions of Lemma 2 hold for all finite-sample PMFs corresponding to each $\theta \in \Theta$, then $\Pi^{\sharp\kappa}$ is the $\kappa$-*sharpened counterpart* of $\Pi$ with $\mathcal{X}^{(1)}$ as the sample space and, for all $\theta \in \Theta$, with $\mu^{(1)}_{\theta}$ as the data probability measure on the power set of $\mathcal{X}^{(1)}$ and $f^{(1)}_{\theta} = d\mu^{(1)}_{\theta}/d\lambda^{(1)}$ as the data PDF. Similarly, if the probability measure $\Pi^{\sharp\kappa}_{\Delta,m}$ on $(\Theta, \mathfrak{F})$ is the $\kappa$-sharpened counterpart of $\Pi$ with $\mathcal{X}^{\Delta,m}$ as the sample space for all $m = 1, 2, \ldots$ and $\Delta > 0$, if $\Pi^{\sharp\kappa}_{\Delta,m} \xrightarrow{\text{weak}} \Pi^{\sharp\kappa}$ as $m \to \infty$ followed by $\Delta \to 0$, and if the convergence conditions of Lemma 3 hold for all finite-sample PMFs corresponding to each $\theta \in \Theta$, then $\Pi^{\sharp\kappa}$ is the $\kappa$-*sharpened counterpart* of $\Pi$ with $\mathcal{X}^{(0)}$ as the sample space and, for all $\theta \in \Theta$, with $\mu_{\theta}$ as the data probability measure on the measurable subsets of $\mathcal{X}^{(0)}$ and $f^{(0)}_{\theta} = d\mu_{\theta}/d\lambda$ as the data PDF.

Since (16) holds for each sharpened prior distribution based on a finite sample space, its equivalent holds for each limiting sharpened prior distribution based on an infinite sample space, where "equivalent" means writing $f^{(1)}_{\theta}$ or $f^{(0)}_{\theta}$ in place of $f_{\theta}$ when the infinite sample space is $\mathcal{X}^{(1)}$ or $\mathcal{X}^{(0)}$, respectively.

**Theorem 5.** *Consider a $\kappa > 0$ and probability measures $\Pi$ and $\Pi^{\sharp\kappa}$ on a measure space $(\Theta, \mathfrak{F})$. Let $\Pi^{\sharp\kappa}$ denote the $\kappa$-sharpened counterpart of $\Pi$ with $\mathcal{X}^{(1)}$ or $\mathcal{X}^{(0)}$ as the sample space, and $\pi^{\sharp\kappa} = d\Pi^{\sharp\kappa}/d\nu$ for a measure $\nu$ that dominates $\Pi$. Almost surely with respect to $\Pi$, if $\mathcal{X}^{(1)}$ is the sample space, then*

$$\pi^{\sharp\kappa}(\theta) = \frac{\pi(\theta)\, e^{-\kappa\, S\left(f^{(1)}_{\theta}\right)}}{\int \pi(\theta')\, e^{-\kappa\, S\left(f^{(1)}_{\theta'}\right)} d\nu(\theta')}, \tag{17}$$

*where $S$ is the entropy function defined by (2), but if $\mathcal{X}^{(0)}$ is the sample space, then*

$$\pi^{\sharp\kappa}(\theta) = \frac{\pi(\theta)\, e^{-\kappa\, S\left(f^{(0)}_{\theta}\right)}}{\int \pi(\theta')\, e^{-\kappa\, S\left(f^{(0)}_{\theta'}\right)} d\nu(\theta')}, \tag{18}$$

*where $S$ is the differential entropy function defined by*

$$S(f_{\theta}) = -\int f_{\theta}(x) \ln f_{\theta}(x)\, dx. \tag{19}$$

*Proof.* First, suppose $\Pi^{\sharp\kappa}$ is the $\kappa$-sharpened counterpart of $\Pi$ with $\mathcal{X}^{(1)}$ as the sample space. By Lemma 2, for all $\theta \in \Theta$,

$$\begin{aligned}
\lim_{m\to\infty} D\left(\mu_\theta^{(1),m} \,||\, \lambda^{(1),m}\right) &= D\left(\mu_\theta^{(1)} \,||\, \lambda^{(1)}\right) \\
\lim_{m\to\infty} \int d\mu_\theta^{(1),m}(x) \ln \frac{d\mu_\theta^{(1),m}}{d\lambda^{(1),m}}(x) &= \int d\mu_\theta^{(1)}(x) \ln \frac{d\mu_\theta^{(1)}}{d\lambda^{(1)}}(x) \\
\lim_{m\to\infty} \sum_{x\in\mathcal{X}^{(0)}} f_\theta^{(1),m}(x) \ln f_\theta^{(1),m}(x) &= \sum_{x\in\mathcal{X}^{(0)}} f_\theta^{(1)}(x) \ln f_\theta^{(1)}(x) \\
-\lim_{m\to\infty} S\left(f_\theta^{(1),m}\right) &= -S\left(f_\theta^{(1)}\right),
\end{aligned} \tag{20}$$

where $\mu_\theta^{(1),m}$ is a probability measure on the power set of $\mathcal{X}^{(1),m}$ and $f_\theta^{(1),m} = d\mu_\theta^{(1),m}/d\lambda^{(1),m}$ for all $m = 1, 2, \dots$. By Definition 8, $\Pi_{(1),m}^{\sharp\kappa} \xrightarrow{\text{weak}} \Pi^{\sharp\kappa}$ as $m \to \infty$, from which and from (16) it follows that

$$\pi_{(1),m}^{\sharp\kappa}(\theta) := \frac{d\Pi_{(1),m}^{\sharp\kappa}}{d\nu}(\theta) = \frac{\pi(\theta)\, e^{-\kappa\, S\left(f_\theta^{(1),m}\right)}}{\int \pi(\theta')\, e^{-\kappa\, S\left(f_{\theta'}^{(1),m}\right)} d\nu(\theta')} \to \pi^{\sharp\kappa}(\theta)$$

as $m \to \infty$ for all $\theta \in \Theta$. Therefore, (20) implies (17) $\Pi$-almost surely.

Next, suppose $\Pi^{\sharp\kappa}$ is the $\kappa$-sharpened counterpart of $\Pi$ with $\mathcal{X}^{(0)}$ as the sample space. Using reasoning analogous to the proof for the $\mathcal{X}^{(1)}$ case, by Lemma 3,

$$\lim_{\Delta\to 0} \lim_{m\to\infty} S\left(f_\theta^{\Delta,m}\right) = S\left(f_\theta^{(0)}\right); \tag{21}$$

$$\pi_{\Delta,m}^{\sharp\kappa}(\theta) := \frac{d\Pi_{\Delta,m}^{\sharp\kappa}}{d\nu}(\theta) = \frac{\pi(\theta)\, e^{-\kappa\, S\left(f_\theta^{\Delta,m}\right)}}{\int \pi(\theta')\, e^{-\kappa\, S\left(f_{\theta'}^{\Delta,m}\right)} d\nu(\theta')} \to \pi^{\sharp\kappa}(\theta) \tag{22}$$

as $m \to \infty$ followed by $\Delta \to 0$ for all $\theta \in \Theta$. Equations (21)–(22) entail (18) $\Pi$-almost surely. □

The next example presents a new Bayesian method of calibrating $p$ values.

**Example 3.** To address the problem of interpreting $p$ values, Held and Ott (2016) presented various lower bounds on the Bayes factor in favor of the null hypothesis. For example, suppose that under the alternative hypothesis $z(X) \sim \mathrm{N}\left(0, \sigma^2\right)$, where $\sigma$ is the alternative hypothesis's standard deviation of $z(X)$, the $p_{\text{one}}(X)$-quantile of the standard normal distribution, given $p_{\text{one}}(X)$, a single-sided p value or $p_{\text{one}}(X) = 1 - p(X)/2$, from $p(X)$, a two-sided $p$ value. Then one of the lower bounds is based on the observed Bayes factor

$$B(x;\sigma) = \sigma e^{-\frac{\left(1-\sigma^{-2}\right)z^2(x)}{2}}$$

if $|z(x)| \geq 1$. With $P(0)$ as the prior probability of the null hypothesis and $P(1) = 1 - P(0)$ as the prior probability of the alternative hypothesis, the null hypothesis's posterior probability is

$$P(0|z(X) = z(x)) = \frac{P(0)\, B(x;\sigma)}{P(0)\, B(x;\sigma) + P(1)} = \frac{P(0)}{P(0) + (1 - P(0))/B(x;\sigma)}$$

by Bayes's theorem. $P\left(0\right)$ may be interpreted as the probability that the null hypothesis component of a mixture model rather than the alternative hypothesis component generated the observation that $z\left(X\right)=z\left(x\right)$.

If the complexity of the distribution of $z\left(X\right)$ conditional on the alternative hypothesis is the differential entropy and was not considered when the value of $\sigma$ was elicited, then $P$ is an unsharpened prior PMF on $\left\{0,1\right\}$. Since $S\left(\mathrm{N}\left(0,\sigma^{2}\right)\right)$, the differential entropy of a normal distribution of standard deviation $\sigma$, is $\ln\sigma$ plus a constant (Michalowicz et al., 2013, p. 127), sharpening $P$ according to (18) leads to

$$\begin{aligned}
P^{\star\sharp\kappa}\left(0\right) &= \frac{P\left(0\right)e^{-\kappa\,S\left(\mathrm{N}\left(0,1\right)\right)}}{P\left(0\right)e^{-\kappa\,S\left(\mathrm{N}\left(0,1\right)\right)}+P\left(1\right)e^{-\kappa\,S\left(\mathrm{N}\left(0,\sigma^{2}\right)\right)}}\\
&= \frac{P\left(0\right)/1^{\kappa}}{P\left(0\right)/1^{\kappa}+P\left(1\right)/\sigma^{\kappa}} = \frac{P\left(0\right)}{P\left(0\right)+P\left(1\right)/\sigma^{\kappa}}\\
&= \left(1+\left(\sigma^{\kappa}\frac{P\left(0\right)}{P\left(1\right)}\right)^{-1}\right)^{-1} \qquad (23)\\
P^{\star\sharp\kappa}\left(0|z\left(X\right)=z\left(x\right)\right) &= \frac{P^{\star\sharp\kappa}\left(0\right)}{P^{\star\sharp\kappa}\left(0\right)+\left(1-P^{\star\sharp\kappa}\left(0\right)\right)/B\left(x;\sigma\right)}\\
&= \left(1+\left(B\left(x;\sigma\right)\frac{P^{\star\sharp\kappa}\left(0\right)}{1-P^{\star\sharp\kappa}\left(0\right)}\right)^{-1}\right)^{-1}\\
&= \left(1+\left(\left(\sigma^{\kappa}B\left(x;\sigma\right)\right)\frac{P\left(0\right)}{P\left(1\right)}\right)^{-1}\right)^{-1} \qquad (24)
\end{aligned}$$

as the sharpened prior and posterior probabilities of the null hypothesis. The sharpened prior is plotted in Figure 1.

Equation (24) suggests viewing $\sigma^{\kappa}B\left(x;\sigma\right)$ as the $\kappa$-*sharpened Bayes factor*, applicable regardless of the value of $P\left(0\right)$. Under the ideal value of $\kappa$ derived in Section 4, that simplicity adjustment, when coupled with an argument of Benjamin et al. (2017), leads to 0.001 or 0.01 rather than 0.005 or 0.05 as the default $p$-value threshold of statistical significance (Bickel, 2019c).

Alternatively, there is a true Bayes factor that is either the unsharpened Bayes factor $B\left(x;\sigma\right)$ or the $\kappa$-sharpened Bayes factor $\sigma^{\kappa}B\left(x;\sigma\right)$ for a $\kappa$ known only to lie within some interval. Let $\underline{\kappa}\geq0$ and $\overline{\kappa}\geq\underline{\kappa}$ denote the lowest and highest possible values of the true $\kappa$; hence, $\underline{\kappa}=0$ if the true Bayes factor is possibly unsharpened and $\overline{\kappa}=0$ if the true Bayes factor is necessarily unsharpened. Since $\underline{\kappa}\leq\kappa\leq\overline{\kappa}$, the true Bayes factor is in $\left[\sigma^{\underline{\kappa}}B\left(x;\sigma\right),\sigma^{\overline{\kappa}}B\left(x;\sigma\right)\right]$. Suppose a decision maker (DM) must decide whether or not to reject the null hypothesis under some loss for rejecting a true null hypothesis and some potentially different loss for failing to reject a false null hypothesis. The action the DM takes minimizes the expected loss with respect to the posterior distribution corresponding to the Bayes factor that a scientist reports. When deciding which Bayes factor in $\left[\sigma^{\underline{\kappa}}B\left(x;\sigma\right),\sigma^{\overline{\kappa}}B\left(x;\sigma\right)\right]$ to report, the scientist incurs

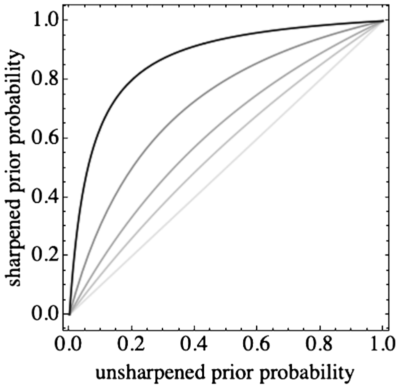

Figure 1: The sharpened prior probability $P^{\star\sharp\kappa}(0)$ that the null hypothesis is true, as a function of the unsharpened prior probability $P(0)$, according to (23) with $\sigma = 2$. The degrees of sharpness are $\kappa = 4$, $\kappa = 2$, $\kappa = 1$, $\kappa = 1/2$, and $\kappa = 0$ from the darkest curve to the lightest curve.

regret according to a caution parameter $C \in [0, 1]$ whenever the DM takes an action different than the action that minimizes the DM's loss function with respect to the posterior distribution corresponding to the unknown true Bayes factor. Accordingly, the scientist reports the *minimax Bayes factor* $\widehat{B}(x)$, the Bayes factor that minimizes the scientist's expected regret while an opponent chooses the true Bayes factor to maximize that expected regret. If the scientist does not know the DM's loss function but can assume certain invariance properties, then the minimax Bayes factor is a $C$-weighted geometric mean of the two most extreme Bayes factors (Bickel, 2019b, Proposition 1), in this case

$$\begin{aligned}\widehat{B}(x) &= \left(\sigma^{\underline{\kappa}} B(x;\sigma)\right)^{1-C} \times \left(\sigma^{\overline{\kappa}} B(x;\sigma)\right)^{C} \\ &= \sigma^{(1-C)\underline{\kappa}+C\,\overline{\kappa}} B(x;\sigma) = \sigma^{\widehat{\kappa}} B(x;\sigma),\end{aligned}$$

which is the $\widehat{\kappa}$-sharpened Bayes factor, where $\widehat{\kappa} = (1 - C)\,\underline{\kappa} + C\,\overline{\kappa}$, the weighted arithmetic mean called the *minimax sharpness*. That has some interesting consequences:

1. At least some sharpening is optimal unless the true Bayes factor is known or the regret is at the least cautious extreme. More precisely, $\widehat{\kappa} > 0$ unless $\overline{\kappa} = 0$ or both $C = 0$ and $\underline{\kappa} = 0$.

2. At the most cautious extreme $(C = 1)$, the minimax sharpness is the highest possible: $\widehat{\kappa} = \overline{\kappa}$.

3. In the case of intermediate caution $(C = 1/2)$, the minimax sharpness is the unweighted arithmetic mean of the lowest possible sharpness and the highest possible sharpness: $\widehat{\kappa} = (\underline{\kappa} + \overline{\kappa})/2$.

4. For any degree of caution, if the unsharpened Bayes factor could be true ($\underline{\kappa} = 0$), then the minimax sharpness is simply $\widehat{\kappa} = C\,\overline{\kappa}$.

5. Putting the conditions of consequences 3 and 4 together with taking Section 4's ideal sharpness as the maximum ($\overline{\kappa} = 1$) yields $\widehat{\kappa} = 1/2$ as a low-sharpness default. ▲

# 8 Adjusting priors for the simplicity of stochastic processes

In a typical Bayesian analysis, the data constitute a sample of $n$ observations that are conditionally independent given each value of $\theta$, the parameter. The data can be viewed in terms of making $n$ observations of an IID stochastic process labeled by an unknown value of $\theta$. More generally, the data consist of a time series of $n$ observations of a stationary stochastic process labeled by an unknown value of $\theta$. In either case, Bayesian coherence does not allow either the unsharpened prior distribution of $\theta$ or the sharpened prior distribution of $\theta$ to depend on $n$. Under that restriction, this section applies sharpened prior distributions to stochastic processes in order to facilitate Bayesian data analysis.

## 8.1 General stochastic processes

A *discrete-time Bayesian model* is a pair $\left(\left(X_{\bullet,1}, X_{\bullet,2}, \dots\right), \pi\right)$ such that $\pi$ is a PDF on $\Theta$ and $\left(X_{\theta,1}, X_{\theta,2}, \dots\right)$ is a stationary discrete-time stochastic process on $\mathcal{X}^{(*)} \in \left\{\mathcal{X}^{(1)}, \mathcal{X}^{(0)}\right\}$ for each $\theta \in \Theta$. To distinguish it from the Bayesian model $(f_\bullet, \pi)$ used in Sections 6–7, the latter is called a *basic Bayesian model*. As defined in information theory, the *entropy rate* of a stochastic process $(X_1, X_2, \dots)$ is

$$\begin{aligned} s\left(\left(X_1, X_2, \dots\right)\right) &= \lim_{t\to\infty} S\left(X_t | X_1, \dots, X_{t-1}\right) \\ &= \lim_{t\to\infty} E_{X_1,\dots,X_{t-1}}\, S\left(g_t\left(\bullet | X_1, \dots, X_{t-1}\right)\right), \end{aligned}$$

where $g_t\left(\bullet | X_1, \dots, X_{t-1}\right)$ is the conditional PDF of $X_t$ given $(X_1, \dots, X_{t-1})$, $S\left(g_t\left(\bullet | X_1, \dots, X_{t-1}\right)\right)$ is its entropy if $\mathcal{X}^{(*)} = \mathcal{X}^{(1)}$ (or its differential entropy (19) if $\mathcal{X}^{(*)} = \mathcal{X}^{(0)}$) as a function of the random $(X_1, \dots, X_{t-1})$, and $E_{X_1,\dots,X_{t-1}}$ gives the expectation value over $X_1, \dots, X_{t-1}$.

**Definition 9.** Let $\left(\left(X_{\bullet,1}, X_{\bullet,2}, \dots\right), \pi\right)$ denote a discrete-time Bayesian model and $(f_\bullet, \pi)$ a basic Bayesian model such that $s\left(\left(X_{\theta,1}, X_{\theta,2}, \dots\right)\right) = S(f_\theta^{(*)})$ for every $\theta \in \Theta$, where $f_\theta^{(*)} = f_\theta^{(1)}$ if $\mathcal{X}^{(*)} = \mathcal{X}^{(1)}$ and $f_\theta^{(*)} = f_\theta^{(0)}$ if $\mathcal{X}^{(*)} = \mathcal{X}^{(0)}$. Let, with respect to $(f_\bullet, \pi)$, $\pi_{f_\bullet}^{\sharp\kappa}$ denote the sharpened counterpart of $\pi$ for some $\kappa > 0$. The same PDF $\pi_{f_\bullet}^{\sharp\kappa}$ is also the $\kappa$-*sharpened counterpart* of $\pi$ with respect to $\left(X_{\theta,1}, X_{\theta,2}, \dots\right)$.

That definition ensures that the expression for sharpened priors over stochastic processes has the same form as those over other observables.

**Corollary 2.** *Consider a discrete-time Bayesian model* $\left(\left(X_{\bullet,1},X_{\bullet,2},\dots\right),\pi\right)$*. Let* $\pi^{\sharp\kappa}_{\left(X_{\bullet,1},X_{\bullet,2},\dots\right)}$ *denote the sharpened counterpart of* $\pi$ *for some* $\kappa>0$*. Then*

$$\pi^{\sharp\kappa}_{\left(X_{\bullet,1},X_{\bullet,2},\dots\right)}\left(\theta\right)=\frac{\pi\left(\theta\right)e^{-\kappa\, s\left(\left(X_{\theta,1},X_{\theta,2},\dots\right)\right)}}{\int\pi\left(\theta'\right)e^{-\kappa\, s\left(\left(X_{\theta',1},X_{\theta',2},\dots\right)\right)}d\nu\left(\theta'\right)}\tag{25}$$

*almost surely with respect to* $\Pi$*, the prior probability measuring defining* $\pi$ *by* $\pi=d\Pi/d\nu$*.*

*Proof.* Let $\left(f_{\bullet},\pi\right)$ denote any basic Bayesian model such that $s\left(\left(X_{\theta,1},X_{\theta,2},\dots\right)\right)=S(f_{\theta}^{(*)})$ for every $\theta\in\Theta$. By Definition 9, $\pi^{\sharp\kappa}_{\left(X_{\bullet,1},X_{\bullet,2},\dots\right)}=\pi^{\sharp\kappa}_{f_{\bullet}}$, allowing the substitution of $s\left(\left(X_{\theta,1},X_{\theta,2},\dots\right)\right)$ for every occurrence of $S(f_{\theta}^{(1)})$ or $S(f_{\theta}^{(0)})$ in (17) or (18) $\Pi$-almost surely according to Theorem 5. □

## 8.2 IID processes

Suppose $X_i\sim g_{\theta,1}$, IID conditional on $\theta$, for all $i=1,2,\dots$, where $g_{\theta,1}$ is a PDF on $\mathcal{X}^{(*)}$ for each $\theta\in\Theta$. Then $s\left(\left(X_{\theta,1},X_{\theta,2},\dots\right)\right)=S\left(g_{\theta,1}\right)=S(f_{\theta}^{(*)})$, and (25) simplifies to

$$\pi^{\sharp\kappa}\left(\theta\right)=\frac{\pi\left(\theta\right)e^{-\kappa\, S(f_{\theta}^{(*)})}}{\int\pi\left(\theta'\right)e^{-\kappa\, S(f_{\theta'}^{(*)})}d\nu\left(\theta'\right)},\tag{26}$$

where $S(f_{\theta}^{(*)})$ is either $S(f_{\theta}^{(1)})$, the entropy of a discrete-valued observable, or $S(f_{\theta}^{(0)})$, the differential entropy of a continuous-valued observable.

### Finite-$\Theta$ examples

In both of the following examples from Section 1, the sample space $\mathcal{X}^{(*)}=\mathcal{X}^{(1)}$ is finite, as is the parameter space $\Theta$. In (26), both $f_{\theta}^{(1)}$ and $\pi$ are PMFs, and $\nu$ is the counting measure.

**Example 4.** Example 1, continued. Since $f_{\theta}^{(1)}\left(x_i\right)=1/\theta$ for $i=1,\dots,\theta$ but $f_{\theta}^{(1)}\left(x_i\right)=0$ for $i=\theta+1,\dots,20$, the entropy is $S(f_{\theta}^{(1)})=\ln\theta$. Then the preliminary prior $P\left(\theta\right)=1/5$ for $\theta=4,6,8,12,20$ gives

$$P^{\sharp\kappa}_{\left(X_{\bullet,1},X_{\bullet,2},\dots\right)}\left(\theta\right)=\frac{\left(1/5\right)\left(e^{\ln\theta}\right)^{-\kappa}}{\sum_{\theta'=4,6,8,12,20}\left(1/5\right)\left(e^{\ln\theta'}\right)^{-\kappa}}=\frac{1/\theta^{\kappa}}{\sum_{\theta'=4,6,8,12,20}1/\theta'^{\kappa}}\tag{27}$$

according to (26). Thus, the ideal sharpened prior probability ($\kappa=1$) of a die is inversely proportional to how many sides it has. ▲

**Example 5.** Example 2, continued. With the uniform distribution on the configuration space $\Theta$ as the preliminary prior, reasoning analogous to that of Example 4 yields the analog of (27),

$$P^{\sharp\kappa}_{\left(X_{\bullet,1},X_{\bullet,2},\dots\right)}\left(\theta\right)=\frac{\left(1/\left|\Theta\right|\right)\left(e^{\ln\left|\theta\right|}\right)^{-\kappa}}{\sum_{\theta'\in\Theta}\left(1/\left|\Theta\right|\right)\left(e^{\ln\left|\theta'\right|}\right)^{-\kappa}}=\frac{1/\left|\theta\right|^{\kappa}}{\sum_{\theta'\in\Theta}1/\left|\theta'\right|^{\kappa}},$$

with $\theta$ as the configuration of colors. Substituting $\kappa = 1$ shows that the ideal sharpened prior probability of a configuration is inversely proportional to how many colors it has. ▲

The prior distributions resulting in the $\kappa = 1$ case of both examples are reciprocal distributions, which are important in studies of Benford's law (Hill, 1995; Pietronero et al., 2001; Kossovsky, 2014, p. 238). While those finite-domain examples motivate the theory, the remaining examples illustrate the sharpening of priors for statistical data analysis with infinite domains.

### Infinite-Θ examples

**Example 6.** Consider $n$ independent trials, each with an unknown probability $\theta$ of success. The entropy per trial is

$$S\left(f_{\theta}^{(1)}\right) = -\theta \log \theta - (1-\theta) \log (1-\theta) .$$

Jeffreys's prior density (Jeffreys, 1948) for a family of binomial distributions is proportional to $\theta^{-1/2} (1-\theta)^{-1/2}$ (Robert et al., 2012, p. 73). By (26), the corresponding sharpened density is

$$\pi^{\sharp\kappa}(\theta) \propto \theta^{-1/2} (1-\theta)^{-1/2} e^{\theta\kappa \log \theta + (1-\theta)\kappa \log(1-\theta)} = \theta^{\theta\kappa - 1/2} (1-\theta)^{(1-\theta)\kappa - 1/2} .$$

Instead of Jeffreys's prior density, Bickel (2016) considered the uniform density as the unsharpened prior. ▲

**Example 7.** Maximizing the entropy of an asymptotic posterior, relative to a prior, leads to the Berger-Bernardo (e.g., Berger and Bernardo, 1989) reference priors (Kass and Wasserman, 1996); see Berger et al. (2009). In the case of a normal data distribution with unknown mean $\mu$ and unknown standard deviation $\sigma$, the reference prior density $\pi(\mu, \sigma)$ is proportional to $1/\sigma$ (Ghosh et al., 2006, §5.1.0). As noted in Example 3, the differential entropy $S(f_{\mu,\sigma}^{(0)})$ of a normal distribution $f_{\mu,\sigma}^{(0)} = \mathrm{N}\left(\mu, \sigma^2\right)$ is $\ln \sigma$ up to a constant term. Then (26) prescribes the sharpened counterpart of the reference prior density:

$$\pi^{\sharp\kappa}(\mu, \sigma) \propto \frac{e^{-\kappa \ln \sigma}}{\sigma} = \frac{1}{\sigma^{1+\kappa}}, \tag{28}$$

which, in the case of Section 4's $\kappa = 1$, is the left-invariant measure (Bickel, 2016). Since the reference density in this case is a probability matching prior, (28) may also be used to adjust $p$ values and confidence intervals for simplicity (Bickel, 2019a). ▲

## 9 Discussion: Which priors should be adjusted for simplicity?

The idea of Examples 1–2 and Section 4 that priors are adjusted according to the coincidences involved in constructing the system studied has implications for whether and how to adjust priors for the simplicity of data distributions. First, simplicity may be warranted not only for default priors such as those in Examples 4–7 but also for other priors that do not account for the coincidences involved in the construction of the system studied (e.g., Example 3). Another implication is that prior distributions that represent known physical variability do not require adjustments for simplicity (cf. Bickel, 2019c), for their probabilities are limiting relative frequencies that do not depend on the construction of systems.

A third implication is that each $f_\theta$ used to adjust a prior for simplicity must reflect the variability intrinsic to the system studied as opposed to technical variability or measurement error. Otherwise, the sharpened prior, like some default priors, would depend on the details of the experiment or observational study, in violation of the likelihood principle. That charge of violating the likelihood principle is often made against default priors that depend on the sampling model (e.g., Ghosh et al., 2006, §5.2; Kadane, 2011, §12.8).

## References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2017). "Redefine statistical significance." *Nature Human Behaviour*, 1. 1314

Berger, J., Bernardo, J., and Sun, D. (2009). "The formal definition of reference priors." *Annals of Statistics*, 37(2): 905–938. MR2502655. doi: https://doi.org/10.1214/07-AOS587. 1318

Berger, J. O. and Bernardo, J. M. (1989). "Estimating a Product of Means: Bayesian Analysis with Reference Priors." *Journal of the American Statistical Association*, 84(405): 200–207. MR0999679. 1318

Bickel, D. R. (2016). "Computable priors sharpened into Occam's razors", working paper, HAL-01423673. URL https://hal.archives-ouvertes.fr/hal-01423673. 1307, 1310, 1318

Bickel, D. R. (2018). "Bayesian revision of a prior given prior-data conflict, expert opinion, or a similar insight: A large-deviation approach." *Statistics*, 52: 552–570. MR3806564. doi: https://doi.org/10.1080/02331888.2018.1427752. 1311

Bickel, D. R. (2019a). "Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors." *Communications in Statistics – Theory and Methods*. doi: https://doi.org/10.1080/03610926.2019.1580739. 1318

Bickel, D. R. (2019b). "Reporting Bayes factors or probabilities to decision makers of unknown loss functions." *Communications in Statistics – Theory and Methods*, 48: 2163–2174. 1315

Bickel, D. R. (2019c). "Sharpen statistical significance: Evidence thresholds and Bayes factors sharpened into Occam's razor." *Stat*, 8(1): e215. 1314, 1319

Blanchard, T. (2018). "Bayesianism and Explanatory Unification: A Compatibilist Account." *Philosophy of Science*. doi: https://doi.org/10.1086/699157. 1301

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. New York: John Wiley & Sons. MR2239987. 1310

Evans, M. (2015). *Measuring Statistical Evidence Using Relative Belief*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. New York: CRC Press. MR3616661. 1310

Fúquene, J., Steel, M., and Rossell, D. (2016). "On choosing mixture components via non-local priors." *arXiv preprint* arXiv:1604.00314. 1311

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy*, 19(10). 1311

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer. MR2247439. 1299, 1318, 1319

Held, L. and Ott, M. (2016). "How the Maximal Evidence of P-Values Against Point Null Hypotheses Depends on Sample Size." *American Statistician*, 70(4): 335–341. MR3574785. doi: https://doi.org/10.1080/00031305.2016.1209128. 1313

Hill, T. P. (1995). "A Statistical Derivation of the Significant-Digit Law." *Statistical Science*, 10(4): 354–363. MR1421567. 1318

Jefferys, W. H. and Berger, J. O. (1992). "Ockham's Razor and Bayesian Analysis." *American Scientist*, 80(1): 64–72. 1301

Jeffreys, H. (1948). *Theory of Probability*. London: Oxford University Press. MR0000924. 1301, 1318

Kadane, J. (2011). *Principles of Uncertainty*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. 1319

Kass, R. E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, 91: 1343–1370. 1299, 1318

Kossovsky, A. (2014). *Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications*. Series in Computer Vision. World Scientific Publishing Company. 1318

Maas, J. (2017). *Entropic Ricci curvature for discrete spaces*, 159–174. Springer International Publishing. MR3727579. 1310

Michalowicz, J. V., Nichols, J. M., and Bucholtz, F. (2013). *Handbook of Differential Entropy*. New York: CRC Press. MR3157477. 1314

Myrvold, W. (2017). "On the evidential import of unification." *Philosophy of Science*, 84(1): 92–114. MR3558492. doi: https://doi.org/10.1086/688937. 1301

Pietronero, L., Tosatti, E., Tosatti, V., and Vespignani, A. (2001). "Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf." *Physica A: Statistical Mechanics and its Applications*, 293(1): 297–304. 1318

Poston, T. (2014). *Reason and Explanation: A Defense of Explanatory Coherentism*. Palgrave Innovations in Philosophy. London: Palgrave Macmillan. 1301

Press, S. (2009). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Ltd. MR1941390. 1299

Rényi, A. (1965). "On the Foundations of Information Theory." *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 33: 1–14. MR0886242. doi: https://doi.org/10.1080/03081078408934872. 1301

Robert, C., Christensen, R., Johnson, W., Branscum, A., and Hanson, T. (2012). "Bayesian Ideas and Data Analysis." MR2682928. 1318

Rosenkrantz, R. D. (1976). *Simplicity*, 167–203. Dordrecht: Springer Netherlands. 1301

Teixeira, A., Matos, A., and Antunes, L. (2012). "Conditional Rényi entropies." *IEEE Transactions on Information Theory*, 58(7): 4273–4277. MR2943089. doi: https://doi.org/10.1109/TIT.2012.2192713. 1308

White, R. (2005). "Why favour simplicity?" *Analysis*, 65(287): 205–210. 1300

## Acknowledgments

I gratefully acknowledge the detailed comments of two anonymous referees, an associate editor, and Michele Guindani. Their efforts led to substantial improvements in clarity. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009).