**New Foundations for Imprecise Bayesianism**

by

Jason Paul Konek

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2013

Doctoral Committee:

Professor James M. Joyce, Chair
Professor Allan F. Gibbard
Assistant Professor Ezra R. Keshet
Assistant Professor Sarah E. Moss

# Acknowledgments

I am indebted and grateful to far too many people to list here. But I must mention a few. Thanks first to my dissertation chair, Jim Joyce. He offered wisdom, guidance and patience in abundance, much more than I deserve. Thanks to Sarah Moss, Allan Gibbard and Eric Swanson. Each of you shaped me in profound ways. Thanks to Gordon Belot for his tireless efforts as placement director. Thanks to Maria Lasonen-Aarnio, Brian Weatherson, David Manley, Rich Thomason, Ezra Keshet, Brandon Fitelson, Mike Titelbaum, Richard Pettigrew, Samir Okasha, James Ladyman, Jan-Willem Romeijn, Hannes Leitgeb and Jon Williamson. A special thanks to Martha Gibson and Dennis Stampe for their early mentorship and continued friendship. Thanks to all the recent Michigan philosophy graduate students, especially Steve Campbell, Dave Wiens, Billy Dunaway, Nate Charlow, Sven Nyholm, Shen-yi Liao, Sara Protasi, Dustin Tucker, Lina Jansson, Dan Singer and Dmitri Gallow. Thanks to my mother, father and grandparents for their love, support, and inspiration. And most of all, thanks to my dearest Monica. I've sapped your strength for years, and will continue doing so for years to come.

# Contents

**Acknowledgments** **ii**

**List of Figures** **iv**

**List of Tables** **viii**

**Abstract** **ix**

**Chapter 1: An Anti-Luck Solution to the Problem of the Priors** **1**

**Chapter 2: Precise Priors without Total Comparative Probability** **44**

**Chapter 3: Cliffordian Conservatism & Imprecise Prior Probabilities** **96**

**Appendix A** **141**

**References** **145**

# List of Figures


**Figure 1.1** Left: $u$ and $u_D$. Right: $b$ and $b_D$. **15**

**Figure 1.2** Cramer-von Mises distance. **18**

**Figure 1.3** $u_D$ and $b_D$. **19**

**Figure 1.4** The objective expected posterior accuracy of $u$ from the perspective of chance hypotheses $H$ of the form $B = x$. **20**

**Figure 1.5** The marginal chance distribution for $\mathfrak{C}(u_D, H)$, which stays fairly constant across hypotheses $H$ about the coin's bias. **22**

**Figure 1.6** The objective expected posterior accuracy of $b$ from the perspective of various chance hypotheses $H$ of the form $B = x$. **23**

**Figure 1.7** The beta distribution with $\alpha = \beta = 1.45$. **26**

**Figure 1.8** Approximation of $f$ given $n = 8$. **27**

**Figure 1.9** Expected posterior accuracy of $s$ (from the perspective of various chance hypotheses $H$ of the form $B = x$). **28**

**Figure 1.10** Expected posterior accuracy of both $u$ and $s$ (rescaled to emphasize difference in variation across hypotheses $H$). **28**

**Figure 1.11** Non-uniform prior over hypotheses $B = x$ defined by $f$. **38**

**Figure 1.12** MaxSen prior $s$ for $B$. **39**

**Figure 1.13** MaxSen prior $s^*$ for $\theta$. **39**

**Figure 1.14** Non-MaxSen prior $s^*$ for $B$. **39**

**Figure 1.15** Bottom curve: objective expected posterior inaccuracy of the truth-value estimate $s_D(X) = \int_0^1 x \cdot s_D(x)dx$ (measuring inaccuracy by the Brier score). Top: expected inaccuracy of $s^*_D(X) = \int_0^1 x^2 \cdot s^*_D(x)dx$. **40**

**Figure 2.1** $f_{Q_X}$ and $f_{Q_Y}$, as well as their means, $p(X)$ and $p(Y)$. **59**

**Figure 2.2** $u$ and $u_D$. **61**

**Figure 2.3** $u$, $b$ and $b^*$. **63**

**Figure 2.4** A sequence of mean-preserving spreads. **64**

**Figure 2.5** Prior distribution over chance hypotheses $ch(W_1 = x) = y$. **66**

**Figure 2.6** Prior distribution $p$ over utility hypotheses $u = x$ conditional on $B$. **66**

**Figure 2.7** Prior and posterior distributions over utility hypotheses $u = x$ conditional on $B$. **67**

**Figure 2.8** Adjacency matrix representing $\preceq$. **75**

**Figure 2.9** A range of distributions which concentrate probability more and more heavily on smaller and smaller subsets of theoretical hypotheses. **81**

**Figure 2.10** Cramer-von Mises distance. **83**

**Figure 2.11** $Exp_u(eu(\mathscr{I}_x))$. **86**

**Figure 3.1** Inaccuracy of $p(H_i) = x$ when $H_i$ is true, relative to Brier and log scores, respectively. **108**

**Figure 3.2** Inaccuracy of $p(H_i) = x$ when $H_i$ is true, according to increasingly convex exponential scores. **113**

**Figure 3.3** Inaccuracy of $p(H_i) = x$ when $H_i$ is true, according to increasingly concave exponential scores. **113**

**Figure 3.4** The Cramer-von Mises distance between two beta distributions, given by a function of the area between their respective cumulative distribution functions. **117**

**Figure 3.5** MaxEnt prior $u$ over hypotheses $B = x$. **119**

**Figure 3.6** Beta distributions with concentration $s = 10$. **120**

**Figure 3.7** Top curve: $\mathscr{M}_{10}$'s objective expected (Jamesian) disutility, relative to chance hypotheses $B = x$. Bottom: $u$'s objective expected disutility, relative to $B = x$. **121**

**Figure 3.8** Top curve: $u$'s objective expected (Cliffordian) disutility, relative to chance hypotheses $B{=}x$. Bottom: $\mathscr{M}_{10}$'s objective expected disutility, relative to $B{=}x$. **122**

**Figure 3.9** Top curve: $u$'s objective expected disutility, relative to chance hypotheses $B{=}x$, measuring disutility by $\mathscr{D}_{0.9}$. Bottom: $\mathscr{M}_{10}$'s objective expected disutility, relative to $B{=}x$. **123**

**Figure 3.10** Left: $\mathscr{M}_{10}$'s objective expected disutility across hypotheses $B{=}x$, compared to the beta distribution $p$ with $\alpha = 0.8$ and $\beta = 2.4$ (entropy: -0.425). Right: $\mathscr{M}_{10}$'s objective expected disutility compared to $q$ with $\alpha = \beta = 0.5$ (entropy: -0.242). **124**

**Figure 3.11** Uniform prior $u$ (bottom) and more concentrated beta prior $b$ (top). **127**

**Figure 3.12** $u_D$ and $b_D$. **128**

**Figure 3.13** The marginal chance distribution $p$ for $\mathscr{D}(u_D, H)$, relative to the true hypothesis $H$ about the coin's bias, $B = 5/7$. **128**

**Figure 3.14** The objective expected posterior epistemic disvalue of $u$, relative to chance hypotheses $B = x$. **130**

**Figure 3.15** The objective expected posterior epistemic disvalue of $b$, relative to chance hypotheses $B = x$. **131**

**Figure 3.16** Top curve: $\mathscr{M}_3$'s objective expected disutility, relative to chance hypotheses $B{=}x$, measuring disutility by $\mathscr{D}_{0.708}$. Bottom: $b$'s objective expected disutility, relative to $B{=}x$. **133**

**Figure 3.17** Beta distributions $p$ with entropy $H(p) \geq 0.24$ and $Exp_p(B) = 1/2$. **135**

**Figure 3.18** Top curve: $p$'s objective expected disutility, relative to chance hypotheses $B{=}x$, measuring disutility by $\mathscr{D}_{0.708}$. Bottom: $\mathscr{E}_{0.24}$'s objective expected disutility, relative to $B{=}x$. **135**

**Figure A.1** Epistemic disutility of $S_{Abstain}$ (bottom) and $S_{Heads \geq Tails}$ (top), respectively, as measured by the linear score $\mathscr{D}_{0.925}$. **143**

# List of Tables


| | | |
|---|---|---|
| **Table 2.1** | Conditional probabilities across priors of decreasing variance. | **63** |
| **Table 2.2** | Joyce's measure of weight across priors of decreasing variance. | **65** |
| **Table 2.3** | Ellsberg problem payoff table. | **68** |
| **Table 2.4** | Epistemic payoff accepting/rejecting/abstaining. | **74** |
| **Table 2.5** | Epistemic payoff of judging $X \preceq Y$, or abstaining. | **76** |
| **Table 2.6** | Epistemic payoff of judging $X \preceq Y$, or abstaining. | **76** |
| **Table 2.7** | Epistemic payoff of judging *Heads* $\preceq$ *Tails*, *Heads* $\succ$ *Tails*, or abstaining. | **77** |
| **Table 2.8** | Posterior probabilities that GC/IMLD/PK beats time $T$. | **82** |
| **Table 2.9** | The effect of weight on $d(f_{X\|D}, f_{Y\|D})/\|u_D(X) - u_D(Y)\|$. | **84** |
| **Table 2.10** | Epistemic payoff of judging *Heads* $\preceq$ *Tails*, *Heads* $\succ$ *Tails*, or abstaining. | **86** |
| **Table 2.11** | Expected epistemic utility of $\mathscr{I}_x$ from the perspective of $u$. | **86** |
| **Table 2.12** | Old 'conservative' payoff matrix. | **87** |
| **Table 2.13** | New 'liberal' payoff matrix. | **88** |
| **Table A.1** | ..... | **141** |
| **Table A.2** | ..... | **144** |

## Abstract

My dissertation examines two kinds of statistical tools for taking prior information into account, and investigates what reasons we have for using one or the other in different sorts of inference and decision problems.

Chapter 1 describes a new objective Bayesian method for constructing 'precise priors'. Precise prior probability distributions are statistical tools for taking account of your 'prior evidence' in an inference or decision problem. 'Prior evidence' is the wooly hodgepodge of information that you come to the table with. 'Experimental evidence' is the new data that you gather to facilitate inference and decision-making. I leverage this method to provide the seeds of a solution to *the problem of the priors*, the problem of providing a compelling epistemic rationale for using some 'objective' method or other for constructing priors. You ought to use the proposed method, at least in certain contexts, I argue, because it minimizes your need for *epistemic luck* in securing accurate 'posterior' (post-experiment) beliefs.

Chapter 2 addresses a pressing concern about precise priors. Precise priors, some Bayesians say, fail to adequately summarize certain kinds of evidence. As a class, precise priors capture improper responses to unspecific and equivocal evidence. This motivates the introduction of imprecise priors. We need imprecise priors, or *sets* of distributions to summarize such evidence. I argue that, despite appearances to the contrary, precise priors are, in fact, flexible enough to capture proper responses to unspecific and equivocal evidence. The proper motivation for introducing imprecise

priors, then, is not that they are required to summarize such evidence. We ought to search for new epistemic reasons to introduce imprecise priors.

Chapter 3 explores two new kinds of reasons for employing imprecise priors. We ought to adopt imprecise priors in certain contexts because they put us in an unequivocally better position to secure epistemically valuable posterior beliefs than precise priors do. We ought to adopt imprecise priors in various other contexts because they minimize our need for epistemic luck in securing such posteriors. This points the way toward a new, potentially promising epistemic foundation for imprecise Bayesianism.

Thesis Supervisor: James M. Joyce

Title: Cooper Harold Langford Collegiate Professor of Philosophy

# Chapter 1

## An Anti-Luck Solution to the Problem of the Priors

"In realistic problems of decision or inference," Edwin Jaynes notes, "we often have prior information which is highly relevant to the question being asked; to fail to take it into account is to commit the most obvious inconsistency of reasoning and may lead to absurd or dangerously misleading results" (Jaynes 1968, 1). When a microbiologist, for example, designs and performs an experiment to adjudicate between competing theoretical hypotheses, *e.g.*, whether over expression of a certain gene causes chromosomal instability in breast tumors, it would be both epistemically irresponsible and practically disastrous for her to ignore the great deal of prior information at her disposal. This includes information about the levels of different genes expressed in past patients, as well as their various clinical symptoms, recurrence rates, etc., information about the broader causal mechanisms that give rise to breast cancer, and so on. Unfortunately, finding a well-motivated, practically useful method for taking prior information into account is difficult. Prior information such as the microbiologist's is incredibly multifarious and complex.

Bayesians argue that the best method for incorporating prior evidence $E$ in decision and inference problems is to specify a 'prior' probability distribution $p$ over the competing hypotheses $H_1, ..., H_n$ which somehow "summarizes [the] great deal

of heterogeneous information" contained in $E$ (Suppes 1966, 203). We can think of these probabilities as estimates of the truth-values of $H_1, ..., H_n$ which (i) satisfy constraints imposed by $E$ while intuitively (ii) going no further than those constraints require. **Subjective** Bayesians say that experienced physicists, biologists, medical researchers, engineers, etc. — agents who are typically quite adept at synthesizing multifarious and complex data — ought to look to their own opinions to furnish these priors. They ought to specify some prior probability distribution which captures their best estimates of the truth of $H_1, ..., H_n$, and in turn reflects their prior evidence $E$ (as well their personal inductive quirks and hunches).

Frequentist statisticians object: if the best method for taking account of prior information requires expert researchers to look to their own opinions to determine 'subjective priors', then we ought to simply ignore this information. Better to make do with statistical tests that "could be described as independent of these [prior] probabilities" than to rob scientific practice of its objectivity (Pearson 1962, 55). Any method for incorporating prior information in inference and decision problems, if it is to have any relevance to science, must be 'objective' in at least the following sense: it prescribes adopting the same prior probability distribution in any two problems where the prior evidence imposes the same constraints (*cf.* Jaynes 1968, 3). The subjectivist method violates this demand. Expert opinions about the plausibility of competing theoretical hypotheses may differ — sometimes significantly — even if they agree, broadly, on the constraints forced on us by the prior evidence.

Contemporary **objective** Bayesians, in contrast, generally endorse the *maximum entropy method* (MaxEnt), which satisfies the demand for objectivity:

- Summarize your prior evidence by constraints $C_1, ..., C_n$, which you model by a set of probability distributions $\mathscr{C}$.

- Adopt the prior $p$ that maximizes entropy $H(p) = -\sum_i p(H_i) \cdot log(p(H_i))$ on $\mathscr{C}$.

Though a researcher's own opinions may be important for determining the *constraints* imposed by her prior evidence (*cf.* Jaynes 1976, 181-194), they should *not* be used to determine the prior distribution in its entirety, on the objectivist view. Instead, all researchers who arrive at the same evidential constraints should proceed in the same manner. They ought to adopt the prior that maximizes entropy on the set of probabilities that satisfy those constraints.

While MaxEnt may provide an 'objective' method for incorporating prior information, in the sense that it prescribes adopting the same prior in any two problems where we have the same evidential constraints, frequentists and subjectivists doubt that there is any compelling epistemic rationale undergirding it. More generally, they think, there is no compelling rationale for us to use *any* 'objective' method for constructing prior probability distributions. *This is the problem of the priors.* In addition, John Venn (1866), J.M. Keynes (1921) and R.A. Fisher (1922) all provide examples that seem to show that MaxEnt yields inconsistent results in a range of cases, depending on how you describe them.[1] This paper outlines and defends a new kind of objective Bayesian solution to the problem of the priors.

In §1.1, I describe Jaynes' rationale for employing MaxEnt, and give some reason to find this rationale wanting. In §1.2, I outline a novel, anti-luck rationale for adopting an alternative prior, the maximally sensitive (MaxSen) prior. In §1.3-1.6, I fill in this outline. In §1.3, I investigate the *theoretical role of priors*, to elucidate the form that a proper response to the problem of the priors ought to take. I suggest that the central role of priors is to help us secure accurate posterior beliefs, and to minimize our need for *epistemic luck* in securing those beliefs. In §1.4, I distinguish two importantly different types of epistemic luck. In §1.5, I illustrate how one prior might depend

more on luck for success than another. In §1.6, I explore the extent to which the MaxEnt prior ameliorates the need for such luck. In §1.7, I describe the MaxSen prior. I argue that this prior does more to ameliorate the need for luck than the MaxEnt prior. In fact, it *minimizes* the need for luck in securing accurate posteriors, and so is *best* suited to play the primary theoretical role of priors. In §1.8, I draw together the preceding threads, to resolve the problem of the priors. Finally, in §1.9, I address two pressing concerns, including a concern about MaxSen's consistency.

## 1.1 The Problem of the Priors

In situations of complete ignorance regarding hypotheses $H_1, ..., H_n$, when our evidence provides *no* constraints on probabilities over $H_1, ..., H_n$, the maximum entropy distribution is just the *uniform* distribution.[2] So MaxEnt agrees with Laplace's Principle of Insufficient Reason (*PIR*):

> *PIR.* In situations of complete ignorance regarding hypotheses $H_1, ..., H_n$, when there is no reason to think that any hypothesis is more or less probable than any other, the uniquely correct prior to adopt is the uniform distribution $u$, so that $u(H_i) = u(H_j)$ for all $i$ and $j$.

Proponents of MaxEnt and PIR *disagree* however about *why* you ought to adopt the uniform prior. Laplace reasons as follows: "when we see no reason that makes one [hypothesis] more probable than the other... this uncertainty makes us regard them as equally probable" (Laplace 1774, 378). But frequentists and subjectivists see this as no better than an admission that there is no good rationale for adopting any particular prior in situations of ignorance, coupled with an arbitrary selection of the uniform distribution. Here is Fisher: the choice of the uniform prior is "evidently extremely arbitrary... evolving a vitally important piece of knowledge, that of the

exact form of the distribution... out of an assumption of complete ignorance" (Fisher 1922, 324-5). In situations of ignorance, we *lack* reason to think any one hypothesis more probable than any other. It would indeed be arbitrary to simply suppose that this forces us to pretend that we have one set of reasons — reasons that speak equally strongly in favor of each hypothesis — rather than any other set of reasons.

Jaynes offers a different rationale. "The maximum-entropy distribution may be asserted," he says, "for the *positive reason* that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the *negative one* that there was no reason to think otherwise" (Jaynes 1957, 623; emphasis mine). The entropy of a distribution $p$, $H(p) = -\sum_i p(H_i) \cdot log(p(H_i))$, is uniquely reasonable, Jaynes argues, as a measure of the *informativeness* of that distribution. This "supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason" (Jaynes 1957, 623). Gone is the old Laplacian rationale, *viz.*, that we are forced to see ourselves as having reasons that speak equally strongly in favor of all hypotheses whenever we lack reasons altogether. In is the new, information-theoretic rationale: in situations of ignorance, our prior evidence is minimally informative. The uniform distribution encodes the minimum amount of information about theoretical hypotheses $H_1, ..., H_n$, since it maximizes entropy (and informativeness decreases as entropy increases). Hence, the uniform distribution best reflects our prior evidence about $H_1, ..., H_n$, at least in terms of informational content.

There is good reason, however, to doubt Jaynes' rationale, and in turn, to doubt the adequacy of the maximum entropy method. The primary theoretical role of priors is *not* to best reflect your prior evidence, as I will argue shortly. Rather, it is to help you secure accurate posterior beliefs by updating on your evidence, and to minimize your need for epistemic luck in securing those beliefs. A proper justification for the

maximum entropy method, if there were one, would illuminate why the MaxEnt distribution is best suited to play the theoretical role of priors. In fact, however, an alternative distribution, the MaxSen prior, is best suited to play this role.

## 1.2 Main Argument

The remainder of this chapter is devoted to outlining and defending a new kind of objective Bayesianism: the maximum sensitivity method, or MaxSen. Schematically, the argument for MaxSen goes as follows.

1. You ought to adopt whichever prior is best suited to play the primary theoretical role of priors, if there is one.

2. The primary role of priors is to help you secure accurate beliefs by updating on your evidence, and to minimize your need for *epistemic luck* in securing those beliefs.

3. Various priors put you in a position to secure accurate posteriors by updating on your evidence.

4. Only the MaxSen prior, however, minimizes your need for epistemic luck in securing accurate posteriors.

C. You ought to adopt the MaxSen prior to incorporate prior information in inference and decision problems.

I qualify this conclusion a bit in §1.5 and §1.7-1.8. I also do not *fully* defend any of premises 1-4. To defend premise 4, for example, I construct the MaxSen prior in toy cases involving simple theoretical hypotheses (about the bias of a coin) and binomial data (data that comes in the form of a sequence of 'successes' and 'failures').

I then show that the MaxSen prior minimizes the need for epistemic luck in these toy cases. This simple approach, however, is sufficient for the modest end of this paper: to *gesture* toward a promising, anti-luck rationale undergirding the MaxSen method, and in turn, to draw attention to a promising resolution to the problem of the priors.

## 1.3 The Theoretical Role of Priors

When we ask,"Is there a good rationale for adopting any particular prior?" we are asking for a certain *kind* of reason in response. If adopting priors suddenly made us better cooks, or lovers, or conversationalists, that would be one reason — a pragmatic reason — to adopt them. But our question demands reasons that speak to *the primary theoretical role of priors*. A *proper* answer to our question takes the form: we ought to adopt this prior or that because it is best suited to play the relevant theoretical role (whatever that may be).

We must, then, be clear about what this theoretical role *is*. Some objective Bayesians, such as Jaynes, assume that the primary role of priors is *representational*. Jaynes prescribes adopting the maximum entropy prior for the "positive reason that it is... maximally noncommittal with regard to missing information" (Jaynes 1957, 623); the maximum entropy prior best *reflects* or *represents* the informational content of our prior evidence.

> *Informational Account.* The primary theoretical role of prior probabilities is to accurately reflect the informational content of the agent's prior evidence.

Certain subjective Bayesians agree that the primary role of priors is representational, but insist that Jaynes and others ought not restrict their attention to evidence. Prior probabilities ought to represent an agent's all-things-considered prior judgments

about the plausibility of hypotheses, which might depend not only on her prior evidence, but also on her assessment of their intrinsic plausibility, her personal inductive quirks, etc.

> *Subjectivist Account.* The primary theoretical role of priors is to accurately represent the agent's prior opinions about the plausibility of hypotheses.

Still other Bayesians, such as Jon Williamson, claim that the primary role of priors is *practical*. Williamson prescribes adopting the maximum entropy prior because it yields the most 'cautious' decision-making policy consistent with the prior evidence (Williamson 2007, 12-7).

> *Practical Account.* The primary theoretical role of priors is to yield the most sensible decision-making policy under conditions of ignorance.

To illustrate Williamson's proposal, suppose that you would like to visit a friend in the city, but you have no evidence about whether the train that you need is running or not. You also have an important Skype meeting in an hour. Your roommate is willing to give you a ride to the station. As long as the train is running, you will make the meeting and see your friend. But if the train is not running, you will have to take an expensive cab home, and may well miss your meeting. If you adopt the MaxEnt (uniform) prior, Williamson observes, your credence that the train is running is $1/2$, and hence (given that the costs of taking an expensive cab and missing your important meeting outweigh the benefits of seeing your friend) the expected utility of staying home is higher than the expected utility of going with your roommate to the station. This 'cautious' decision-making policy, Williamson claims, is clearly the sensible one, given how scant your evidence is. (You have none!) Hence, the MaxEnt prior, in virtue of delivering this sensible policy, is well-suited to play the relevant theoretical role.

Each of these accounts is inadequate. The practical account is difficult to make sense of in a non-question-begging manner. The reason: which decision-making policy counts as most cautious depends on which epistemic perspective you evaluate it from. Suppose, for example, that you are nearly certain that the train is running, despite having no evidence about the matter. Then the decision-making policy that MaxEnt recommends appears downright foolish from your perspective, not cautious. It will not do, by the way, to insist that in light of your evidence (you have none) you ought to evaluate MaxEnt's decision-making policy from a more 'even-handed' epistemic perspective, *e.g.*, one in which you treat the competing hypotheses — that the train is running, and that it is not — as equally probably. This is just to evaluate MaxEnt's decision-making policy *from its own perspective*. And the decision-making policy yielded by *any* prior appears most cautious from that prior's own perspective.

The informational and subjectivist accounts, on the other hand, are inadequate because they pay insufficient attention to the *theoretical role of evidence itself*. Evidence helps us secure *accurate* posterior credences, or truth-value estimates. Credences are more accurate the closer they are to the actual truth-values. And accuracy is a 'basic epistemic good'. Whatever else is true of them, credences are more valuable, from the epistemic perspective, the more accurate a picture of the world they paint (*cf.* Joyce 2009, 267-71). But evidence does more than just this. It also helps us secure accurate posteriors *in a way that minimizes our need for epistemic luck*. For example, gathering ballistic evidence, DNA evidence, etc. minimizes the detective's need for epistemic luck in arriving at a true belief about who killed Jones.

This fully characterizes the theoretical role of evidence. Evidence is important to our epistemic lives, at bottom, *exactly* because it helps us secure accurate posteriors in a luck-minimizing fashion. Plausibly, then, prior probabilities — statistical tools for taking prior evidence into account — are important exactly to the extent that they

enable evidence to play *its* role, *i.e.*, to assist us in securing accurate posteriors in a luck-minimizing fashion. This suggests the following position about the theoretical role of priors:

> *Instrumental Account.* The primary theoretical role of priors is to put us in a position to secure accurate, minimally luck-dependent posterior credences by updating on new data.

When the various roles listed above conflict, it is clear that this final role takes precedence. When, for example, the prior that best represents a researcher's opinions about the plausibility of hypotheses happens to put her in a rather poor position to secure accurate, minimally luck-dependent posterior credences, it would be absurd to advise her to adopt that prior (the same goes for the prior that most accurately reflects the informational content of her prior evidence).[3] Suppose, for example, that a scientist has scant prior evidence about the causal mechanism under investigation (a particular virus' infection mechanism, perhaps). She *does*, however, find one particular hypothesis extremely intrinsically plausible. But she does not find it plausible for any good reason. Her hunch reflects no particular *skill* at assessing intrinsic plausibility. She simply 'feels it in her bones'. Then advising her to adopt a prior that reflects this hunch, by concentrating probability on her favorite hypothesis, would be absurd. It would result in her discounting new data that she really ought to be more sensitive to (in much the way that a conspiracy theorist discounts data that tells against her favorite hypothesis, *e.g.*, that an alien spacecraft crashed near Roswell, New Mexico in 1947).

This illustrates what should be clear: whichever prior best enables evidence to play *its* theoretical role is *ipso facto* best suited to play the theoretical role of priors. It is worth noting that the instrumental account does, in fact, enjoy a certain measure of

support in the literature. Here, for example, is Patrick Suppes: "It is of fundamental importance to any deep appreciation of the Bayesian viewpoint to realize that the particular form of the prior distribution expressing beliefs held before the experiment is conducted is not a crucial matter... The well-designed experiment is one that will swamp divergent prior distributions with the clarity and sharpness of its results" (Suppes 1966, 204). The reason that it is not a crucial matter exactly which form the prior distribution takes is that, in a 'well-designed' experiment, the data we receive is fairly *weighty*. And when the data we receive is weighty, the 'washing out' theorems show that a range of priors converge on the true theoretical hypothesis (with high objective probability).[4] As a result, those priors are all likely to yield fairly accurate — and minimally luck-dependent — posterior distributions. Hence, they all play the primary theoretical role of priors close to equally well. And they do so even though some priors do a rather poor job representing, for example, the agent's prior opinions about the plausibility of hypotheses. This latter fact is — or at least ought to be — "not a crucial matter" from the Bayesian viewpoint.

One final aside: unfortunately, not all experiments yield data weighty enough to "swamp divergent prior distributions with [its] clarity and sharpness" in the way Suppes envisions (Suppes 1966, 204). Limits on time, personnel, funding, etc. keep scientific researchers from gathering as much data as they would like. And in those pitiable, but all-too-common circumstances, the washing out theorems do not have much purchase. Many priors will depend *significantly* on luck for success (accuracy). If there is one prior that is minimally luck-dependent, then, at least in these circumstances, there will be good reason to use *it* to take your prior information into account.

## 1.4 Two Kinds of Epistemic Luck

*Now* when we ask, "Is there a good rationale for adopting any particular prior?" we have something of an answer. Our answer: there *is* a good rationale if there is some prior distribution that minimizes the need for epistemic luck in securing accurate posterior beliefs. Such a distribution would be uniquely suited to play the primary theoretical role of priors. I will argue that there is such a distribution: the maximally sensitive, or MaxSen prior. Before describing the MaxSen prior, though, we ought to get clearer on the target phenomenon: epistemic luck.

There are various kinds of epistemic luck. If the ground under an archer could easily have shifted, but did not, and she fires off a skillful shot which hits the bullseye, then her success is subject to what virtue epistemologists such as Pritchard (2009) call *environmental luck*. This is the sort of luck that enables agents to exercise skill. Without it, our archer would not have gotten her shot off, and so would not have been successful (hit the bullseye). Even so, note: certain important contrastive facts about her success are explained primarily by her *skill*, an *internal* factor, *e.g.*, the fact that she hit the bullseye dead on, rather than two (or three, or four) inches below the bullseye (or above the bullseye, or to the left of the bullseye, etc.).

In contrast, another sort of luck — *intervening* luck — severs this explanatory link. If an expert archer's shot is knocked off-track and then back on-track by subsequent gusts of wind, then she is subject to intervening luck. Her shot is, to a high degree, successful, but not *because* it was skillful (her shot is not *apt*, in Sosa's terminology; *cf.* Sosa 2007, 79). Her particular degree of success (the fact that it hit the bullseye, rather than two, or three, or four inches below, etc.) is not explained primarily by internal factors (the agent's skill). Rather, it is explained by external factors (fortuitous gusts of wind). We will take this to be the defining characteristic of intervening luck: it

is in play when external factors are primarily responsible for explaining an agent's particular *degree* of success (why she achieved exactly *this* degree of success, rather than some other degree).

Prior distributions are also subject to intervening epistemic luck, in the following sense: when you update a prior on evidence, it yields a posterior which is more or less accurate (more or less successful). This particular degree of accuracy (why the posterior is accurate to exactly *this* degree, rather than some other degree), in turn, is *explained* more or less by two different kinds of factors. On the one hand, internal factors — facts about the prior's intrinsic properties, such as how *resilient* it is (*cf.* §1.5) — might bear the bulk of the explanatory burden. On the other hand, external factors — facts about the prior's extrinsic properties, such as the proximity of a coin's true bias to the prior's expected bias — might end up shouldering a bigger part of this burdern.

Of course, no prior minimizes dependence on luck *tout court*. There are various kinds of both environmental and intervening luck that adopting a prior — any prior — will simply not mitigate. No prior mitigates the environmental luck in play when a researcher's heart keeps functioning normally, rather than failing (as it easily could have, perhaps). No prior helps eliminate the luck involved in stumbling upon a friend returning from a movie, and learning that the ending was a disaster ("...and then she opened her eyes, and it was all a dream!"). (No prior mitigates this sort of luck in receiving new evidence.) And no prior (fully) ameliorates the luck involved in avoiding wildly misleading evidence, of the sort that a gambler faces if she observes a coin with bias $B = 0.9$ (biased strongly toward heads) come up tails 19 of 20 tosses.[5] In searching, then, for a distribution best suited to play the primary theoretical role of priors, we ought to attempt to identify a prior that yields posterior credences which depend minimally on a *special kind* of intervening epistemic luck (the sort of

luck susceptible to mitigation by savvy prior construction), not epistemic luck *tout court*. A plausible candidate: luck in having the true chances fall close to one's prior estimates of the chances. When we talk of epistemic luck from here on out, we will have this special kind of luck in mind.

## 1.5 Dependence on Luck: An Example

To illustrate how one distribution might depend more on luck for success than another — in particular, luck in having the true chances fall close to its prior estimates — compare priors of varying *resilience*. A prior distribution $p$ is resilient with respect to a datum $D$ to the extent that the posterior distribution $p_D$ ($p$ conditioned on $D$) is close to $p$. Compare, for example, the maximum entropy (uniform) distribution $u$ over hypotheses about the bias of a coin, $B = x$, on the one hand, and a more concentrated distribution $b$ on the other hand (*e.g.*, a beta distribution with $\alpha = 10$ and $\beta = 4$).[6] (Beta distributions $b$ are parameterized by two quantities, $\alpha$ and $\beta$. These 'shape parameters' determine which hypotheses $B = x$ the distribution $b$ focuses its probability mass on. The larger (smaller) $\alpha$ is compared to $\beta$, the more $b$ focuses probability mass on $B = x$ with $x \approx 1$ ($x \approx 0$). For more information, see endnote 7.[7]) Suppose that you flip the coin 15 times. It comes up heads 12 times and tails 3 times. When you condition the maximum entropy distribution on this data sequence ($H^{12}T^3$), it moves quite a bit: the distance from $u$ to $u_D$ is 0.107 (at least when you measure distance using one plausible distance function, Cramer-von Mises, detailed in §1.6).[8] The more concentrated distribution, in contrast, moves much less: the distance from $b$ to $b_D$ is 0.007. (Both priors and posteriors are pictured right, next page.) Even if you had observed the data sequence that makes the more concentrated distribution move *most* ($H^0T^{15}$), it would not have moved *much* more

than the maximum entropy distribution: 0.35 as compared to 0.3.

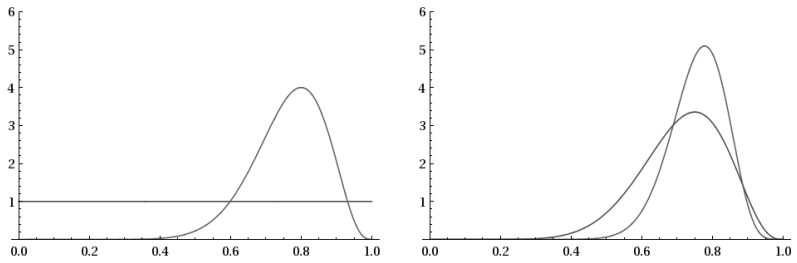


Figure 1.1: Left: $u$ and $u_D$. Right: $b$ and $b_D$.

Priors that are more resilient than others with respect to a wide range of data tend to depend more on luck for success (*i.e.*, luck in have the true chances fall close to its prior estimates). Consider Betsy and Jim, for example, two climate scientists working on papers about the future impact of climate change on coastal areas for the new IPCC report (intergovernmental panel on climate change). They are both competent, let's suppose. But neither are Sherlock Holmesian 'super sleuths'. Neither are especially skilled at assessing the 'intrinsic plausibility' of the various candidate climate models. They do, however, have access to large experimental data sets (the same data sets). This includes data about current land air temperature, sea surface temperature, sea level, ozone, etc. It also includes data about these quantities over the last 100 years. In addition, they have a great deal of *prior* information about how these quantities interact, about the broader causal mechanisms that give rise to climate change, and so on (the same prior information, suppose). In order to incorporate their prior information, both Betsy and Jim adopt priors over theoretical hypotheses (climate models). Jim adopts a prior that concentrates probability almost exclusively on one particular theoretical hypothesis (a concentrated beta distribution). In contrast, Betsy adopts a prior that is much less resilient than Jim's with respect to a wide range of data (the MaxEnt prior). The effect is that Betsy's prior is much more malleable, much more prone to change in the face of new data.

Both Betsy and Jim consult the climate data for the last 100 years. As Betsy pores over it, she updates her prior, which tends to move quite a bit and forces her to revise her credences for observing different sorts climate-related effects in coastal areas (conditional on the current ozone levels being one way or another, on greenhouse emission rates staying constant, etc.). In contrast, as Jim pores over the data for the last 100 years, his prior tends to not move much at all. So he revises his credences minimally.

Finally, Betsy and Jim both update their priors on the data regarding current climatic conditions and deliver their reports to the IPCC. Their total data is non-misleading, let's suppose, and both Betsy and Jim are successful. They both end up making *accurate* predictions about what sorts of effects to expect in coastal areas over the next 10 years (*e.g.*, erosion, ecosystem loss, coral bleaching). But Jim's success depends more on luck than Betsy's. In particular, it depends more on luck in having the true theoretical hypothesis (climate model) fall close to his prior estimate. Had the true climate model been rather dissimilar from Jim's preferred model, and had Jim received similarly non-misleading evidence, his posterior distribution would have been much further from the truth than it currently is. In turn, his predictions about what sorts of effects to expect in coastal areas over the next 10 years would have been much less accurate. Not so for Betsy. Her posterior distribution would have converged on the true climate model to nearly the same extent that it actually does (see §1.6 for more detail).

One might wonder, "Why restrict our attention to non-Holmesian researchers? Why not imagine that Jim is especially *skilled* at assessing the intrinsic plausibility of theoretical hypotheses? Suppose he is. Suppose the bias in his prior reflects this skill. And suppose that intrinsic plausibility is a reliable guide to the truth. Then we could say that Jim's success depends rather minimally on luck as well." True enough, but be-

side the point. Remember, our aim is to identify a general, impersonal, well-motivated method for constructing priors over theoretical hypotheses, which researchers can use to incorporate prior evidence in inference and decision problems. The sort of Holmesian skill imagined by our objector, however, presumably delivers information much too complex to be summed up (even approximately) by constraints on expectations, and does so in a manner much too complex to be captured by any tractable algorithm (perhaps similar to skill at diagnosing obscure medical conditions). But then it is hopeless to write such skill into the actual protocol for constructing priors. And, unfortunately, many researchers lack this skill. So no sufficiently general method for constructing priors simply advises individual researchers to exercise this skill, while staying silent on what this amounts to. The upshot: Sherlock Holmes and his ilk are well-advised to exercise their skill to arrive at a prior, rather than employing an 'objective' method like MaxSen. But it is, nonetheless, worthwhile to identify a general, impersonal method for incorporating prior information in inference and decision problems, which non-Holmesian researchers can use to arrive at accurate, minimally luck-dependent posteriors.

## 1.6 Ameliorating Dependence on Luck

Priors that are more resilient than others with respect to a wide range of data tend to depend more on luck for success. To make this claim a bit more precise, and to substantiate it, consider one attractive measure of datum-relative resilience. Recall, a distribution $p$ is *resilient* with respect to a datum $D$ to the extent that $p_D(\cdot) = p(\cdot|D)$ is close to $p$. Deza and Deza (2009) survey a wide range of distance functions on the space of probability distributions, each of which gives us a different way of saying exactly how close $p_D$ is to $p$. I will focus on one in particular, *Cramer-von Mises*

*distance*:

$$\mathfrak{C}(p, q) = \int_0^1 |P(x) - Q(x)|^2 dx$$

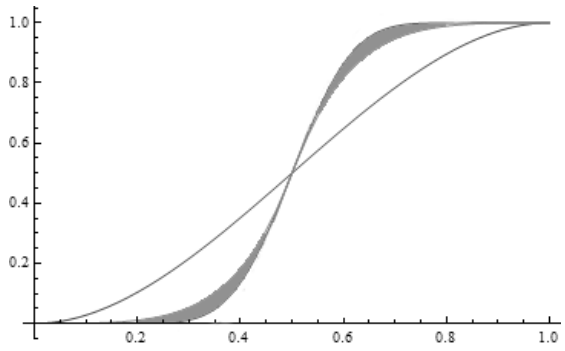


Figure 1.2: Cramer-von Mises distance.

$\mathfrak{C}$ specifies the distance between distributions $p$ and $q$ as a function of the area between their corresponding cumulative distribution functions, $P$ and $Q$ (counting regions of smaller divergence for less and regions of greater divergence for more; pictured left).[9] (It is the squared $L_2$ metric between $P$ and $Q$.) It is attractive because (i) it is an analogue of squared Euclidean distance on the space of probability densities, and (ii) it yields the correct verdict about comparative closeness in those cases where obviously correct answers are to be had.[10] In addition, note that the Brier score, $I(p, w) = (1/N) \cdot \sum_{i=1}^{N} (p(X_i) - w(X_i))^2$ — a paradigmatically reasonable scoring rule (see Joyce 1998, 2009 and Leitgeb and Pettigrew 2010) — measures the inaccuracy of discrete distributions by squared Euclidean distance. Because $\mathfrak{C}$ provides a natural extension of squared Euclidean distance to the space of continuous distributions, I will sometimes speak of $\mathfrak{C}(p, H)$ (the Cramer-von Mises distance between $p$ and the indicator distribution $i_H$ which places all of its probability on $H$) as the *accuracy* of $p$ with respect to $H$.

On the proposed view, a distribution $p$ is resilient with respect to a datum $D$ to the extent that the following is close to zero: $\mathfrak{C}(p, p_D) = \int_0^1 |P(x) - P_D(x)|^2 dx$. To see why resilient priors tend to depend more on luck for success (posterior accuracy), consider an illustrative case. Compare, once more, the maximum entropy (uniform) distribution $u$ over hypotheses about the bias of a coin, and a more concentrated beta

distribution $b$ (with $\alpha = 10$ and $\beta = 4$). A bookie hands you and your friend a coin, and offers you a bet. Neither of you have any prior evidence about the coin's bias. The bookie allows you to flip the coin for awhile prior to deciding whether or not to take the bet. You adopt the maximum entropy prior $u$. Your friend adopts the more biased beta prior $b$ (she feels it in her bones that the coin's bias is roughly $b$'s mean, *viz.*, 5/7). Note: $b$ is more resilient than $u$ with respect to a wide range of data.

You flip the coin 14 times. It comes up heads 10 times and tails 4 times. When you both condition on this data $D$, you arrive at the posteriors $u_D$ and $b_D$, respectively (right). Suppose that $D$ is perfectly non-misleading evidence; the true hypothesis $H$ about the bias of the coin is $B = 5/7$ (exactly the frequency of heads in your data sequence). Then your friend is more successful (accurate). Her distribution converges more on $H$ than yours: $\mathfrak{C}(u_D, H) = 0.028 > 0.020 = \mathfrak{C}(b_D, H)$. But her success also depends more on luck in having the coin's true bias fall close to her prior estimate than yours. Had the coin's true bias fallen further from her prior estimate, and had she received similarly non-misleading evidence, then her posterior distribution would have ended up much further from the truth than it currently is. Not so for you. Your posterior distribution would have converged on the true hypothesis about the coin's bias to nearly the same extent.

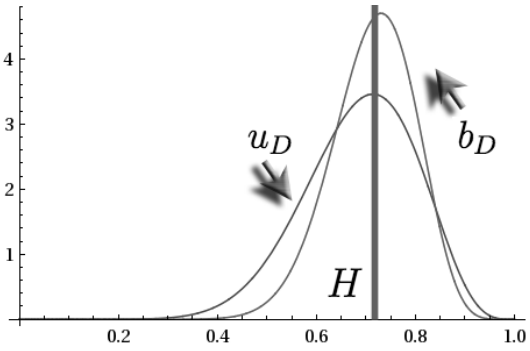


Figure 1.3: $u_D$ and $b_D$.

Even more to the heart of the matter, your distribution $u$'s expected posterior accuracy:

$$\sum_{k=0}^{14} \binom{14}{k} \cdot x^k \cdot (1-x)^{14-k} \cdot \mathfrak{C}(u_D, H)$$

stays fairly constant across hypotheses $H$ of the form $B = x$ (*i.e.*, hypotheses about the coin's bias; left). To see that this is the crux of why her success depends more on luck than yours, consider an example.

> *The Rain Machine.* You stumble upon a machine with the potential to affect the rainfall in London. Let $R$ be the amount of rainfall in London tomorrow. The machine (somehow) graphs the marginal chance distribution $p$ for $R$. It also has two knobs, set all the way to the left, in the 'off' position. As you spin the top knob, $p$ changes fairly significantly. As you spin the bottom knob, it remains largely unaltered. Before you leave the machine, you spin both knobs all the way to the right.

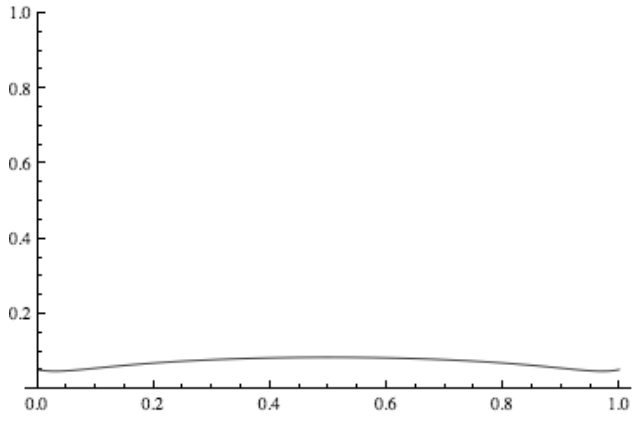


Figure 1.4: The objective expected posterior accuracy of $u$ from the perspective of chance hypotheses $H$ of the form $B = x$.

The next day London gets 3mm of rain. The fact that it gets *this* amount of rain (3mm), rather than something less (2mm, 1mm, etc.), is explained, in part, by the position of the top knob. The position of the bottom knob, in contrast, is more or less irrelevant. Why? Well, the explanation of the fact that London gets 3mm of rain, rather than 2mm, 1mm, etc. is probabilistic. The most proximate explanatory factor is that the marginal chance distribution $p$ for $R$ has a particular character (a particular, mean, variance, etc.). To explain why London gets 3mm of rain, rather than 2mm, 1mm, we must cite not only probability mass that $p$ assigns to $R = 3$, but also the

mass that $p$ assigns to $R = 2$, $R = 1$, and so on; the entire distribution is relevant (all of its moments: mean, variance, etc.). In addition, $p$ serves as an explanatory screen. Any other factor relevant for explaining why London gets 3mm of rain, rather than 2mm, 1mm, etc. is only relevant in virtue of explaining why $p$ takes the exact form that it does. Now, the position of the top knob, clearly, is relevant for explaining why $p$ has the character it does. Had you only turned it half way to the right, rather than all the way to the right, $p$ would have had a much different character. In contrast, the position of the bottom knob is next to irrelevant for explaining why $p$ has the character it does. Regardless of how you turn the bottom knob, $p$ remains almost entirely unaltered. Plausibly, then, the position of the bottom knob is (more or less) irrelevant for explaining why London gets 3mm of rain, rather than 2mm, 1mm, etc.[11]

Similarly, the explanation of the fact that $u_D$ is inaccurate to a particular degree ($\mathfrak{C}(u_D, H) = 0.028$), rather than some other degree (0.027, 0.026, etc.) is probabilistic. The most proximate explanatory factor is that, immediately prior to your experiment (flipping the coin), the true marginal chance distribution $p$ for $\mathfrak{C}(u_D, H)$ had a particular character (pictured right). And just as above, to explain why $u_D$ is inaccurate to the exact degree that it is, rather than something slightly higher or lower, we must cite not only probability mass that $p$ assigns to the hypothesis $\mathfrak{C}(u_D, H) = 0.028$, but also the mass that $p$ assigns to $\mathfrak{C}(u_D, H) = 0.027$, $\mathfrak{C}(u_D, H) = 0.026$, etc.; the entire distribution is relevant. In addition, $p$ serves as an explanatory screen. Any other factor relevant for explaining why $u_D$ is inaccurate to exactly the degree that it is (0.028), rather than some other degree (0.027, 0.026, etc.), is only relevant in virtue of explaining why $p$ takes the exact form that it does.[12]

Now note: $p$ is more or less invariant across hypotheses $H$ about the coin's bias. Whether the true bias is 5/7, 11/64 or 82/97, the marginal chance distribution $p$ for $\mathfrak{C}(u_D, H)$ will look more or less the same.[13] This is reflected in the fact that $p$'s

mean — $u$'s expected posterior accuracy — stays fairly constant across hypotheses $H$ (see figure 1.4, p. 20). The upshot: the external factor in question — how close the coin's true bias happened to fall to $u$'s prior estimate — is not terribly relevant to explaining why $p$ takes the exact form that it does. Hence, it is also not terribly relevant to explaining why $u_D$ is inaccurate to exactly degree 0.028, rather than 0.027, 0.026, etc. The moral: $u$ depends fairly minimally on luck in having the true chances fall close to its prior estimates for success (posterior accuracy).

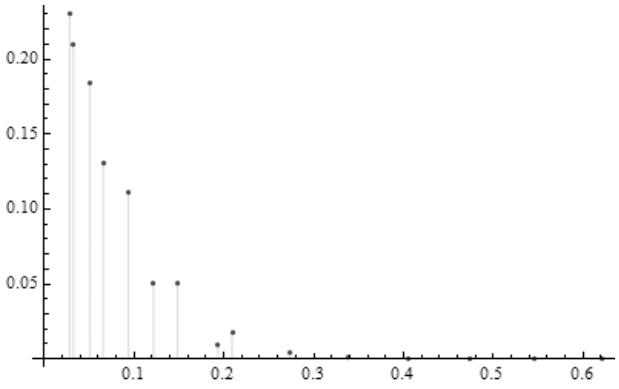


Figure 1.5: The marginal chance distribution for $\mathfrak{C}(u_D, H)$, which stays fairly constant across hypotheses $H$ about the coin's bias.

To hammer this point home, consider a less fanciful analogy than the rain machine.

> *The Expert Archer.* A highly skilled archer faces a number of different targets $T$ arranged at varying distances. Given her expertise, the marginal chance distribution $p$ for $D$ (distance of her arrow from the center of the target) looks more or less the same, regardless of which target she takes aim at. Whether she aims at some target $T$ rather close by, or some $T'$ rather far away (within reasonable bounds, of course), $p$ assigns roughly the same (high) probability mass to the hypothesis $D = 0$ (hitting the target dead center), roughly the same (low) probability mass to the hypothesis $D = 15$ (hitting 15cm off target), and so on.

Because $p$ remains largely unaltered across targets $T$, the initial proximity of our archer to $T$ is plausibly (more or less) irrelevant for explaining why $p$ takes the

exact form that it does. And because facts about the form that $p$ takes serve as an explanatory screen vis-á-vis $D$ — any other factor relevant for explaining why $D = 0$ (she hits the target dead center), rather than $D = 1$, $D = 2$, etc., is only relevant in virtue of explaining why $p$ takes the exact form that it does — that initial proximity is (more or less) irrelevant for explaining why our archer is successful to the exact degree that she is. This mirrors the coin flipping case. Because $p$ remains largely unaltered across chance hypotheses $H$, the (initial) proximity of your prior $u$ to $H$ is plausibly (more or less) irrelevant for explaining why $p$ takes the exact form that it does. And because facts about the form that $p$ takes serve as an explanatory screen vis-á-vis posterior accuracy — any other factor relevant for explaining why $\mathfrak{C}(u_D, H) = 0.028$, rather than $\mathfrak{C}(u_D, H) = 0.027$, $\mathfrak{C}(u_D, H) = 0.026$, etc., is only relevant in virtue of explaining why $p$ takes the exact form that it does — that initial proximity is next to irrelevant for explaining why your posterior $u_D$ is successful (accurate) to the exact degree that it is (0.028).

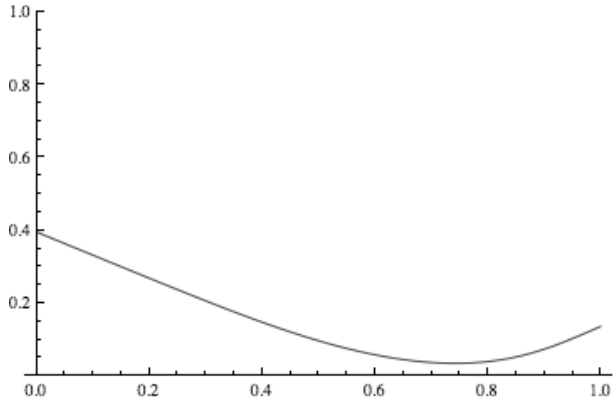


Figure 1.6: The objective expected posterior accuracy of $b$ from the perspective of various chance hypotheses $H$ of the form $B = x$.

Your friend, however — the one who adopts the more biased beta prior $b$ — is in a different boat. The marginal chance distribution $q$ for $\mathfrak{C}(b_D, H)$ varies rather significantly across chance hypotheses $H$. This is reflected in the fact that $q$'s mean — the expected posterior accuracy of *her* distribution $b$ — varies significantly across $H$ (left, previous page). The upshot: the (initial) proximity of her prior $b$ to $H$ *is* relevant for explaining why $q$ takes the exact form that it does. In turn, it *is* relevant

for explaining why her posterior $b_D$ is successful (accurate) to the exact degree that it is (0.020).

The situation here is not unlike that of an unskilled archer. Such an archer might face targets $T$ arranged at varying distances. Suppose she aims at a close one and hits the bullseye dead center. Unlike in the expert archer case, the marginal chance distribution $q$ for $D$ (distance of *her* arrow from the center of the target) varies significantly across $T$. If she aims at some target $T$ rather close by, the mean of $q$ (*i.e.*, the expected value of $D$) might be close to 0. There is a high chance of hitting the bullseye dead center, a lower chance of hitting 1cm off target, an even lower chance of hitting 5 cm off target, etc. But if, instead, she aims at some $T'$ far away, the mean of $q$ might be much higher. There is a much higher chance of missing the bullseye by quite a bit. The upshot: the unskilled archer's initial proximity to her target *is* relevant for explaining why $q$ takes the exact form that it does. In turn, it *is* relevant for explaining why she is successful to the exact degree that she is.

This all serves to highlight an important virtue of the MaxEnt prior. It renders external factors less explanatorily relevant than certain other priors (more concentrated beta priors), and thereby does more to ameliorate dependence on intervening epistemic luck. The MaxEnt prior is better suited, then, to play the primary theoretical role of priors. But an alternative prior, *viz.*, the MaxSen prior, is even better suited to play this role.

## 1.7 The MaxSen Method

When a scientist designs and performs an experiment aimed at adjudicating between competing theoretical hypotheses, $H_1, ..., H_n$, she ought to, according to the MaxSen method, take her prior information into account as follows:

- Summarize her prior evidence by constraints $C_1, ..., C_m$, which we model by a set of probability distributions $\mathscr{C}$.

- Adopt the prior $s$ in $\mathscr{C}$ that is 'maximally sensitive' to evidence in the following sense: the experimental data, rather than the initial proximity of the true theoretical hypothesis $H$ to $s$'s prior estimate, explains, to the greatest extent possible, posterior accuracy.

  - Formally: minimize $f(p) = max_i\ Exp_{H_i}(d(p_D, H_i))$ - $min_j\ Exp_{H_j}(d(p_D, H_j))$ on $\mathscr{C}$.[14] (Read $d(p_D, H_i)$ as the distance between $p_D$ and the indicator distribution $i_{H_i}$ which places all of its probability on $H_i$.)

    * We continue to use $\mathfrak{C}$ to measure the distance between probability distributions, though it is open to the proponent of MaxSen, of course, to use an alternative distance function.

To illustrate the MaxSen method, imagine once more that you have a coin of unknown bias. You plan to perform $n$ independent coin flips, in order to adjudicate between competing chance hypotheses. You have no relevant prior information, suppose, save for the following (which we assume only to limit computational complexity): your prior ought to take the form of a beta distribution. So $\mathscr{C}$ is the set of all beta distributions.

At this point, MaxEnt prescribes adopting the uniform prior. The distribution that maximizes entropy on $\mathscr{C}$ is the beta prior with $\alpha$=$\beta$=1, which is just the uniform prior. MaxSen, in contrast, prescribes adopting a non-unform beta prior. Which prior it prescribes depends on the details of the experimental set-up. I address this issue in §1.9.1. For now, just note that this is to be expected, if what we have said about

the theoretical role of priors is correct. If what we have said is correct, then priors are merely instrumentally valuable tools for securing accurate, minimally luck-dependent posterior credences by updating on new data. It is no surprise that which tools are best suited for this end depends on what new experimental data you stand to receive. And it is no surprise that what data you stand to receive depends on the details of the experimental set-up. This includes details about what *kind* of evidence the experiment is designed to yield: evidence about the outcomes of coin flips, or about levels of gene expression, or about sea surface temperature, etc. It also includes details about, for example, the number of times $n$ that the experiment is to be repeated. In our coin flip example, if you are going to flip the coin 8 times ($n = 8$), then the MaxSen prior is the beta distribution with $\alpha = \beta = 1.45$ (right). If instead you are going to flip the coin twenty times ($n = 20$), then the MaxSen prior is the beta distribution with $\alpha = \beta \approx 2$.

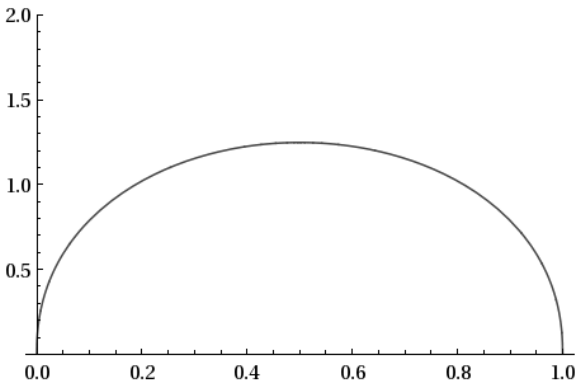

Figure 1.7: The beta distribution with $\alpha = \beta = 1.45$.

To construct the MaxSen prior, one might use any number of optimization algorithms, *e.g.*, a Markov Chain Monte Carlo algorithm. I use simple regression analysis here, since my purposes are merely illustrative. Consider, for example, the case of $n = 8$ (you flip the coin eight times). In this case, if you choose a reasonably fine partition of $\mathscr{C}$, and evaluate $f(p) = max_i\ Exp_{H_i}(d(p_D, H_i))$ - $min_j\ Exp_{H_j}(d(p_D, H_j))$ at the upper and lower bounds of the elements of this partition, regression analysis yields the polynomial approximation $f^*$ of $f$ (pictured left, next page).[15] As is clear from inspection of this graph, $f$ takes a minimum, roughly, at the beta distribution

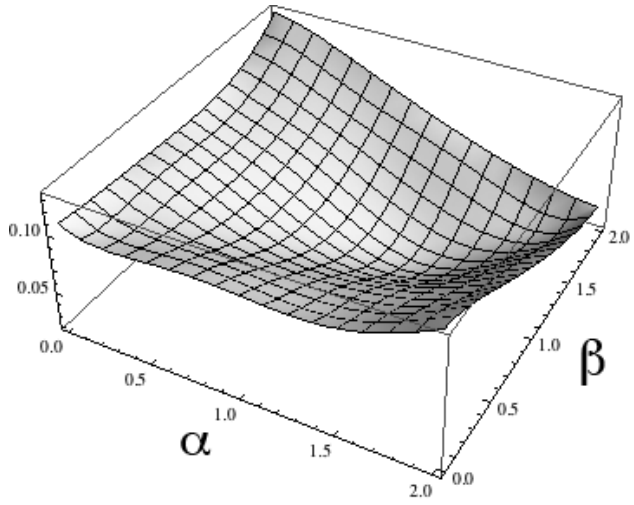


Figure 1.8: Approximation of $f$ given $n$=8.

with $\alpha = \beta = 1.4$. Standard Lagrangian methods reveal the unique minimum to be, more precisely, at the beta distribution $s$ with $\alpha = \beta = 1.45$. This is the MaxSen prior. Similar techniques can be used to approximate the MaxSen distribution for any $n$. In addition, Orbanz and Teh (2010) describe how to use standard inference techniques in a way that can be leveraged to construct the MaxSen distribution in more difficult inference and decision problems (we return to this issue in §1.8.2).

The MaxSen prior $s$, rather than the MaxEnt (uniform) prior $u$, minimizes the need for epistemic luck in securing accurate posteriors. Recall, $u$'s expected posterior accuracy varies fairly minimally across chance hypotheses $H$. There is a fairly high chance that the experiment will yield data that causes it to converge significantly on the true chance hypothesis $H$, regardless of which $H$ is true. As a result, the MaxEnt prior $u$ performs *better* than many other priors (low variance beta priors) vis-á-vis ameliorating dependence on epistemic luck. This notwithstanding, the MaxSen prior $s$'s expected posterior accuracy varies significantly less than $u$'s with changes in $H$ (pictured next page). In fact, the difference between the maximum and minimum expected accuracy is only 0.012 ($f(s) = 0.012$). The result: facts about how close the true chances happened to fall to $s$'s (prior) estimates — an external factor — play virtually no role in explaining the (posterior) accuracy of $s_D$. Not only does the MaxSen prior perform *better* than the MaxEnt prior vis-á-vis ameliorating dependence

on luck, but it performs (more or less) as well as any prior could perform in this regard.

Once more, the situation is not unlike that of an expert archer. Whether she aims at some target rather close by, or another far away (within reasonable bounds), she has roughly the same (high) chance of hitting the target dead center, the same (lower) chance of hitting 1cm off target, and so on. The result: facts about how close she happened to be to her target play virtually no role in explaining her success. Her skill ameliorates her dependence on luck — in particular, luck in having her target fall close to some 'preferred' distance — to the maximum extent possible.

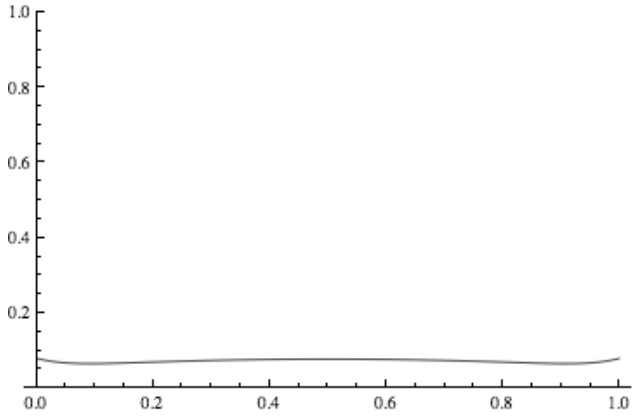


Figure 1.9: Expected posterior accuracy of $s$ (from the perspective of various chance hypotheses $H$ of the form $B = x$).

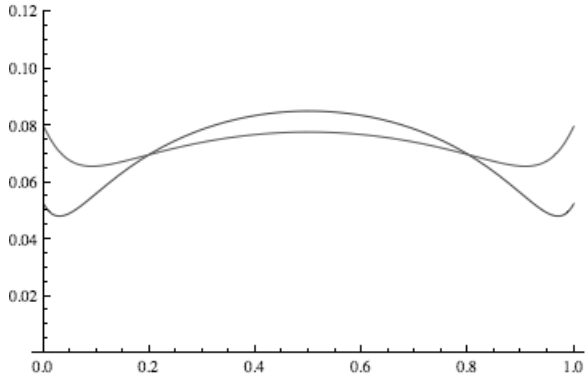


Figure 1.10: Expected posterior accuracy of both $u$ and $s$ (rescaled to emphasize difference in variation across hypotheses $H$).

The moral is this: the MaxSen prior renders external facts — facts about how close the true chances happened to fall to $s$'s initial estimates — virtually explanatorily irrelevant. In turn, it minimizes the need for epistemic luck in securing accurate posterior beliefs. Hence, the MaxSen prior is uniquely suited to play the primary theoretical role of priors (at least in simple decision and inference problems like those considered here).

## 1.8 Concluding Remarks

### 1.8.1 Recap

I have argued that there is a promising, anti-luck rationale for employing the MaxSen method to incorporate prior information in inference and decision problems. If correct, this resolves the problem of the priors. To recap, the argument goes as follows:

1. You ought to adopt whichever prior is best suited to play the primary theoretical role of priors, if there is one.

2. The primary role of priors is to help you secure accurate beliefs by updating on your evidence, and to minimize your need for epistemic luck in securing those beliefs.

3. Various priors put you in a position to secure accurate posteriors by updating on your evidence.

4. Only the MaxSen prior, however, minimizes your need for epistemic luck in securing accurate posteriors.

C. You ought to adopt the MaxSen prior to incorporate prior information in inference and decision problems.

The majority of this chapter is devoted to defending premise 4. It is worthwhile to restate that defense here, in a more succinct form.

1′. No prior mitigates the need for epistemic luck *tout court*. Rather, the prior that minimizes your need for luck, if there is one, does so by mitigating a special kind of intervening epistemic luck: luck in having the true chances fall close to one's initial estimates of the chances.

2′. Intervening luck is the sort of luck in play when external factors are primarily responsible for explaining success. Mitigating intervening luck is a matter of rendering such factors explanatorily irrelevant.

3′. The MaxSen prior $s$ renders facts about how close the true chances happen to fall to $s$'s initial estimates less explanatorily relevant than any other prior.

4′. So, the MaxSen prior does more to mitigate the relevant kind of luck than any other prior. (*From 2 and 3*)

C′. This is all that a prior can do to ameliorate dependence on epistemic luck *tout court*; so the MaxSen prior minimizes the need for such luck in securing accurate posteriors. (*From 1 and 4*)

### 1.8.2 Outstanding Issues

This chapter motivates and details a new kind of objective Bayesian method, the MaxSen method, for constructing priors. But it does not provide a full defense of this method. Consider premise 3′. In arguing for this premise, we restricted our attention to inference and decision problems involving simple theoretical hypotheses about the bias of a coin and binomial data. But, a full defense must consider a much broader class of theoretical hypotheses and data. Microbiologists, for example, are not concerned with the biases of coins or binomial data. They design and perform experiments aimed at adjudicating between more complex theoretical hypotheses, *e.g.*, causal models that describe how over expression of a certain gene produces chromosomal instability in breast tumors. And the data that such experiments yield — qualitative data about the reorganization of certain cellular structures, quantitative data about levels of DNA replication, etc. — is certainly not binomial (does not come in the form of a sequence of 'successes' and 'failures').[16] In inference and

decision problems of this sort, Bayesian priors are nonparametric.[17] A fuller defense of the MaxSen method would illustrate how standard inference techniques (MCMC, sequential Monte Carlo, variational inference, expectation propagation) can be used to determine a nonparametric MaxSen prior in such problems, and to compute its posterior (see Orbanz and Teh 2010 and Neal 2000). It would also identify the conditions under which these techniques are applicable.

This paper also does not address some pressing concerns. For example, Venn (1866), Keynes (1921) and Fisher (1922) all provide examples that seem to show that MaxEnt yields inconsistent results in a range of cases, depending on how you describe them. I briefly address Fisher's concern below. But it is incumbent on the proponent of MaxSen to show definitively that these problems do not extend to her proposal.

Examining the boundaries of the class of contexts in which MaxSen is applicable, and responding fully to description or parameterization dependency concerns are tasks that require separate investigation. Our aim here was simply to highlight the kind of epistemic rationale undergirding MaxSen, and in turn, to highlight a promising route for resolving the problem of the priors. I conclude by raising a few additional questions to be addressed in future research.

- We specified the MaxSen prior using one particular distance function on the space of probability densities, *viz.*, Cramer-von Mises distance. Are our results robust under a range of metrics, *e.g.*, the Lévy metric? the $L_p$ metrics?

- The MaxSen prior outperforms alternative precise priors, including the MaxEnt prior, vis-á-vis ameliorating dependence on luck. But we have not compared the MaxSen prior to *imprecise* priors, or sets of probability functions. Is there good epistemic reason to prefer the MaxSen prior to alternative imprecise priors, at least in certain contexts of inquiry? (I address this issue in chapter 3.)

- In §1.3, we remarked that, in certain circumstances, limits on time, personnel, funding, etc. keep scientific researchers from gathering weighty enough data for the washing out theorems to have much purchase. And in such circumstances, many priors will depend significantly on luck for success (accuracy). When is an experiment sufficiently "well-designed" for some fairly large class of priors to depend fairly minimally on luck for success?

## 1.9 Objections

### 1.9.1 Likelihood Principle

The MaxSen method seems to run afoul of the Likelihood Principle:

> *Likelihood Principle* (LP). For any two experiments aimed at adjudicating between theoretical hypotheses $H_1, ..., H_n$, and any two data sequences $D$ and $D'$ produced by those experiments, if $D$ and $D'$ determine the same likelihood function (up to an arbitrary positive constant), *i.e.*, there is some $k > 0$ such that $p(D|H_i) = k \cdot p(D'|H_i)$ for all $H_i$, then the 'evidential meaning' or 'evidential import' of $D$ and $D'$ for $H_1, ..., H_n$ is the same. (*cf.* Edwards, Lindman and Savage 1963, 237)

Many Bayesian statisticians, such as Savage, de Finetti and Berger (as well as 'frequentist' statisticians such as Fisher) take the LP to be central to rational inductive inference. Birnbaum (1962) summarizes the standard Bayesian rationale for the LP as follows. First, on the Bayesian view, according to Birnbaum, the aim of rational inductive inference is to use "experimental results along with other available [prior] information" to determine a posterior that provides "an appropriate final synthesis of available information" (Birnbaum 1962, 299). Posteriors 'synthesize' the total

available data $E$ by specifying truth-value estimates for the theoretical hypotheses under investigation, $H_1, ..., H_n$, which capture the 'evidential meaning' or 'evidential import' of $E$ for $H_1, ..., H_n$.

Second, Bayes' theorem tells us that posteriors, $p(\cdot|D)$, are fully determined by two components: a prior, $p(\cdot)$, and a likelihood function for the experimental data $D$, $p(D|\cdot)$, which specifies how probable the various theoretical hypotheses $H_1, ..., H_n$ render $D$.

$$\begin{aligned} \textit{Bayes' Theorem.} \quad p(H_i|D) &= [p(D|H_i) \cdot p(H_i)]/p(D) \\ &= [p(D|H_i) \cdot p(H_i)]/\textstyle\sum_j p(D|H_j) \cdot p(H_j) \end{aligned}$$

Finally, because the *prior* distribution captures the 'evidential meaning' of the *prior* data (no more, no less), the likelihood function must capture the 'evidential meaning' of the experimental data, on the Bayesian view (no more, no less). "In this sense," Birnbaum says, "we may say that [Bayes' theorem] implies [the likelihood principle]" (Birnbuam 1962, 299). "The contribution of experimental results to the determination of posterior probabilities is always characterized just by the likelihood function and is otherwise independent of the structure of an experiment" (*ibid.*).

MaxSen seems to violate the LP by making 'extraneous' features of the experimental set-up — in particular, its 'stopping rule' — relevant to the 'evidential meaning' or 'evidential import' of experimental data. Stopping rules are rules that specify when to stop gathering new data. They are extraneous, according to the LP, because they have no influence on likelihoods.

The argument that MaxSen violates the LP, by making stopping rules relevant to evidential force, goes as follows. First, as any proponent of the method would happily admit, stopping rules *are* relevant to which prior you ought to adopt, according to MaxSen. Suppose, for example, that you and your friend are going to flip a coin, in

order to adjudicate between competing hypotheses about its bias. You adopt different fixed stopping rules — you plan to flip the coin 8 times; your friend plans to flip it 20 times. Then MaxSen recommends that you adopt the beta prior $s$ with $\alpha = \beta = 1.45$. It recommends that your friend adopt the beta prior $s'$ with $\alpha = \beta = 2$.

From here, it seems, we are just a few small steps from showing that MaxSen violates the LP.

1. Posteriors reflect the 'evidential meaning' of the total available data (prior and experimental) for the theoretical hypotheses under investigation.

2. MaxSen renders posteriors sensitive to stopping rules.

3. So, according to MaxSen, the 'evidential meaning' of the total available data is sensitive to stopping rules. (*From 1 and 2*)

4. Stopping rules are obviously irrelevant to the meaning of the *prior* data. If they are relevant to the meaning of the total data at all, it must be because they impact the meaning of the *experimental* data.

5. Hence, the meaning of the experimental data is sensitive to stopping rules, according to MaxSen. (*From 3 and 4*)

6. The LP says: stopping rules are irrelevant to the meaning of the experimental data.

C. MaxSen violates the LP. (*From 5 and 6*)

In fact, though, MaxSen is perfectly consistent with the LP. The problem with this argument: premise 1 is false. The primary role of priors is *not* to reflect the 'evidential meaning' of the prior data, or anything of the sort, but rather, to help us secure

accurate posteriors in a luck-minimizing fashion (or so I argued in §1.3). Similarly, the primary role of posteriors is *not* to reflect the 'evidential meaning' of the total data (prior and experimental together), but rather, to encode accurate, minimally luck-dependent truth-value estimates for both theoretical hypotheses (*e.g.*, models of viral infection mechanisms, climate models, etc.) and non-theoretical propositions (regarding ecosystem loss, etc.).

Rational inductive inference, on this view, is simply *not* aimed at using "experimental results along with other available [prior] information" to determine a posterior that provides "an appropriate final synthesis of available information" (Birnbaum 1962, 299). Of course, summarizing 'what the data says' — its 'evidential meaning' or 'evidential import' — is important for various purposes, *e.g.*, reporting experimental results in science journals. But rational inductive inference aims at something different: getting at the truth (securing accurate truth-value estimates) in a minimally luck-dependent fashion.

The proponent of MaxSen might elaborate as follows: the 'evidential meaning' or 'evidential import' of a body of evidence is almost always best summarized by a *set* probabilities. Such 'meanings' are rarely specific enough to single out a unique distribution. Prior evidence, for example, typically imposes constraints on prior probabilities, constraints satisfied by a range of distributions. And, given that the LP is true, capturing the correct 'meanings' for experimental data items is a matter of encoding the correct likelihoods; many priors encode the correct likelihoods. Normally, then, there will be a *set* of distributions that, when updated on the experimental data, reflect the evidential meaning of the total evidence as well as any other distribution.

Still, if MaxSen is correct, one of these priors is uniquely well-suited to play the primary role of priors, *viz.*, to help us secure accurate posteriors in a luck-minimizing fashion. The crucial point is this: its distinguishing properties — the properties that

set it apart from the other priors that adequately summarize 'what the data says' — are important *merely* because they make it (the MaxSen prior) well-suited to play the relevant theoretical role. *These distinguishing properties do not reflect anything about the 'evidential meaning' of the data, prior or experimental.* Such 'meaning' is characterized by the constraints that the prior evidence imposes (which the MaxSen prior satisfies), and the likelihood functions for potential experimental data items (which the MaxSen prior correctly encodes).

To recap: MaxSen is consistent with the LP, despite its sensitivity to stopping rules. Stopping rules *do* determine certain features of the MaxSen prior. But these features do not reflect 'what the data says'. They are merely instrumentally valuable 'design features' which make the MaxSen prior well-suited to play its particular theoretical role.

A final note: appreciating the proper role of priors — to help us secure accurate, minimally luck-dependent posteriors — not only squares MaxSen with the LP, but also makes clear why one's choice of a prior *should* be sensitive to stopping rules. Consider a practical analogy. Monica has an investment advisor. The advisor's goal is to deliver the largest return that she can on Monica's investments at some time point, *e.g.*, 10 years from now. She has two tools to achieve this goal: (i) the investment capital that Monica provides each month, and (ii) an investment strategy.

Now, *some* features of Monica's circumstances are irrelevant to which investment strategy her advisor ought to adopt: whether she hopes to retire in Montana or Monterrey, for example. This has no effect on which investment strategy will yield the highest return. But other features of her circumstances clearly *do* matter. It clearly matters how much investment capital Monica has available. If she can invest $1,000 per month, and her friend can invest $5,000 per month, then it would be foolish for their respective advisors to adopt the same investment strategy. Her advisor might be

best served opting for a conservative investment strategy (government bonds, etc.), while her friend's advisor is better served by focusing more on higher risk/higher reward options.

Similarly, a researcher might have a couple tools at her disposal for achieving her epistemic goal (securing accurate, minimally luck-dependent posteriors): (i) the data that her experiment yields, and (ii) an inductive strategy (encoded by her prior). Some features of her experimental set-up are clearly irrelevant to which inductive strategy (prior) she ought to adopt: whether her pipettes were made by company A rather than company B, for example. This has no effect on which inductive strategy is likely to yield the highest 'epistemic return' (the most accurate, least luck-dependent posteriors). But other features of her circumstances clearly *do* matter. It clearly matters how much data that experiment will yield (which depends on the stopping rule she employs). Just as one financial advisor might be better served opting for a more 'aggressive' investment strategy than another, if her client has more investment capital to work with, so too might one researcher be better served opting for a more 'aggressive' inductive strategy than another, if she has more (weightier) experimental data to work with. She can afford to adopt a prior that concentrates probability more on less 'extreme' theoretical hypotheses (hypotheses that assign less extreme objective probabilities to experimental data items), without increasing her dependence on luck. The reason: the more extreme hypotheses will do more to 'make themselves heard'; given the weightiness of the data, they will either be very strongly confirmed or very strongly disconfirmed.

It is no surprise, then, that which prior you ought to adopt depends on how much data your experiment is designed to yield (and other important features of the experimental set-up, *e.g.*, whether the *kind* of data that it is designed to yield is particularly probative vis-á-vis the relevant theoretical hypotheses). And it is no

strike against MaxSen that it respects this fact.

### 1.9.2 Parameterization Dependence

As frequentists and subjectivists often note, MaxEnt seems to yield inconsistent results in a range of cases, depending on how you describe them. The following example, adapted from Fisher 1922 (pp. 324-5), illustrates the point. One final time, consider a coin of unknown bias. You plan to flip the coin $n$ times, in order to adjudicate between the competing chance hypotheses. You have no relevant prior information, save for the following: your prior ought to take the form of a beta distribution.

Given these evidential constraints, MaxEnt prescribes adopting the uniform prior over hypotheses $B = x$. But, Fisher points out, you "might never have happened to direct [your] attention to the particular quantity" $B$ (Fisher 1922, 325). Instead, you might have maximized entropy with respect to $\theta = \sqrt{B}$. "The quantity, $\theta$," Fisher says, "measures the degree of probability, just as well as [$B$], and is even, for some purposes, the more suitable variable" (*ibid.*, 325). If, however, you maximize entropy with respect to $\theta$, you will adopt the uniform prior over hypotheses of the form $\theta = x$, which is equivalent to adopting a *non*-uniform prior over hypotheses $B = x$ defined by the probability density $f(x) = 1/(2\sqrt{x})$ (right).

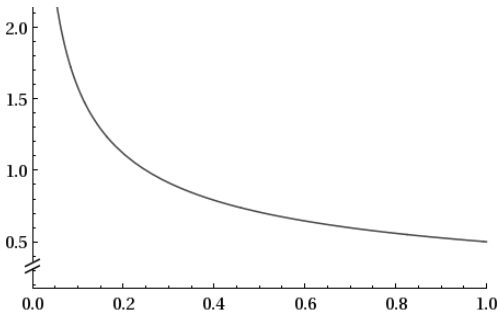


Figure 1.11: Non-uniform prior over hypotheses $B = x$ defined by $f$.

The upshot: depending on which parameter you focus on, $B$ or $\theta$, MaxEnt will prescribe a different prior. In turn, you will make different (inconsistent) judgments in the two cases. For example, if you maximize entropy with respect to $B$, and then observe two heads in a row, your new best estimate of the coin's bias is 0.75. In

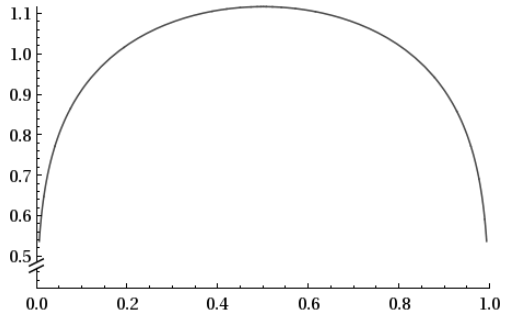


Figure 1.12: MaxSen prior $s$ for $B$.

MaxSen prior over hypotheses $B = x$ is the beta distribution $s$ with $\alpha = \beta \approx 1.2$ (left, above). If, in contrast, you use the MaxSen method to determine a beta prior over hypotheses of the form $\theta = x$, you will arrive at the distribution $s^*$ with $\alpha \approx 0.9$ and $\beta \approx 1.5$ (right). And $s^*$ is *not* equivalent to $s$. Adopting the prior $s^*$ over hypotheses $\theta = x$ is equivalent to adopting the distribution over hypotheses $B = x$ defined by the probability density $g(x) = 0.65581\sqrt{1-\sqrt{x}}/x^{0.55}$ (left, below).

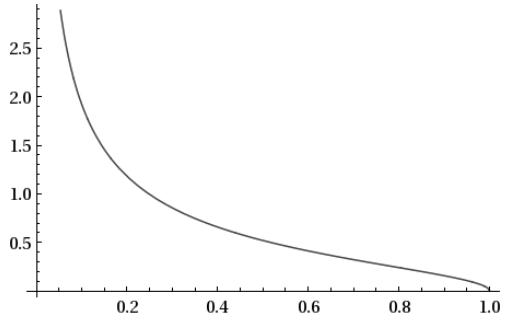


Figure 1.14: Non-MaxSen prior $s^*$ for $B$.

contrast, if you maximize entropy with respect to $\theta$ and observe two heads, your best estimate is 0.71.

One might suspect that the MaxSen method is subject to a similar sort of parameterization dependence. And it is. But this is a *feature*, not a *bug*. Suppose that you plan to flip the coin of unknown bias 5 times. The

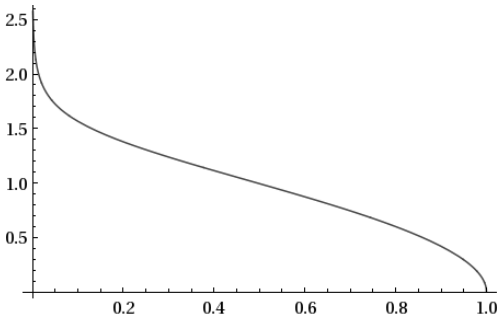


Figure 1.13: MaxSen prior $s^*$ for $\theta$.

This is a feature, I claim, not a bug, because (i) there *are* grounds for focusing certain parameters, rather than others; the MaxSen method does *not* leave you in the precarious position of yielding different prescriptions relative to different parameters, with no good reason to choose between them; (ii) *in virtue* of its parameterization dependence, the MaxSen

method is *flexible* enough to yield appropriate prescriptions for a variety epistemological tasks.

Which parameter you ought to focus on, I claim, depends on what your epistemic aims are in your context of inquiry. Suppose, for example, that your aim is two-fold: (i) to arrive at an accurate, minimally-luck dependent posterior $p_D$ over theoretical hypotheses (about the bias of the coin, about the square root of the bias of the coin, etc.), and (ii) to arrive at an accurate, minimally-luck dependent estimate, $p_D(X)$, of the truth-value of $X$, the proposition that the coin will come up heads on the next toss. Perhaps you care about the former *because* it is a good means to the latter. When $p_D$ is sufficiently accurate (it concentrates probability significantly enough on the true theoretical hypothesis), it will yield a truth-value estimate for $X$ that is (objectively) likely to be accurate too.

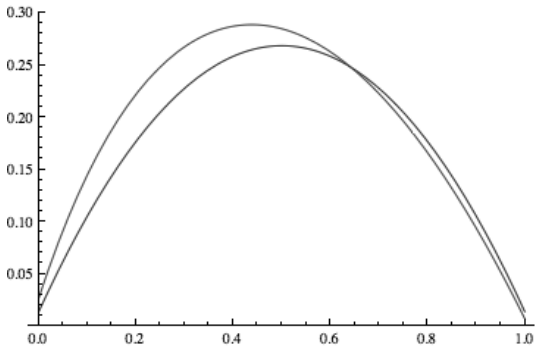


Figure 1.15: Bottom curve: objective expected posterior inaccuracy of the truth-value estimate $s_D(X) = \int_0^1 x \cdot s_D(x)dx$ (measuring inaccuracy by the Brier score). Top: expected inaccuracy of $s^*_D(X) = \int_0^1 x^2 \cdot s^*_D(x)dx$.

If this your aim — to get at the truth of $X$ in a minimally luck-dependent fashion — then you ought to focus on the parameter $B$, rather than $\theta$, I claim. The reason: the objective expected accuracy of $s_D(X) = \int_0^1 x \cdot s_D(x)dx$ (the posterior probability for $X$ determined by the MaxSen prior $s$ over hypotheses $B = x$) varies less across hypotheses $B = x$ (and $\theta = x$) than the expected accuracy of $s^*_D(X) = \int_0^1 x^2 \cdot s^*_D(x)dx$ (the posterior probability determined by the MaxSen prior $s^*$ over hypotheses $\theta = x$). So $s$ depends less on luck in yielding a successful (accurate) posterior probability for

$X$ than $s^*$ does.

The reason, informally, is that the MaxSen prior $s^*$ over hypotheses $\theta = x$ is particularly adept at converging on the true value of $\theta$. But small changes in $\theta$ correspond to large changes in $B$. So it is not quite as adept at converging on the true value of $B$ (the chance of $X$). Moreover, $s^*$'s posterior probability (truth-value estimate) for $X$ is just its estimate of $B$. And the accuracies of these two estimates hang together. The less (objectively) likely it is to yield an accurate estimate of $B$, the less (objectively) likely it is to yield an accurate posterior probability for $X$. The upshot: it is not particularly adept at converging on the actual truth-value of $X$. It requires more *epistemic luck* for success (accuracy) than the MaxSen prior $s$ over chance hypotheses $B = x$.

So there *are* grounds for focusing on $B$ rather than $\theta$. If your aim is to get at the truth of $X$ in minimally luck-dependent fashion, then you would be better served by focusing on $B$ than on $\theta$. But, in *other* contexts of inquiry, it might be epistemically important to arrive at accurate, minimally luck-dependent estimates of *other* quantities. For this end, it might be better to adopt a prior that is better at converging on the true value of $\theta$. It is precisely because the MaxSen method offers non-equivalent prescriptions for $B$ and $\theta$ that it is able to furnish priors that are well-suited for these different tasks.

The moral is this: researchers are typically concerned not only with securing accurate, minimally luck-dependent truth-value estimates for theoretical hypotheses, but also with securing such estimates for *various other quantities*. We might not only care about how over expression of some gene influences instability in breast tumors. We might also care about whether a patient will go into remission if she receives certain therapies or treatments. Which other quantities we are concerned with — *e.g.*, the truth-values of pertinent non-theoretical propositions (about remission, and so on)

— can and ought to inform how we go about investigating theoretical hypotheses, by informing how we parameterize them.

# Notes

[1]Venn (1866), Keynes (1921) and Fisher (1922) aim their objections at Laplace's Principle of Insufficient Reason (*PIR*). But these problems for *PIR* extend directly to MaxEnt.

[2]More carefully, when our evidence provides no constraints on probabilities over theoretical hypotheses, *and a countably additive uniform distribution over that space exists*, then the MaxEnt distribution is just the uniform distribution. In certain contexts, however, the MaxEnt prior exists while a countably additive uniform prior does not. For example, Furrer et al. (2011) claim to specify techniques that can be used to derive an 'infinite-dimensional generalization of the entropic uncertainty relation' (Furrer et al. 2011, 12). But, it is well known that there is no Lebesgue measure on infinite-dimensional spaces, and hence, no analogue of the standard uniform distribution.

[3]See ch. 3, §1.

[4]See Savage 1972, pp. 46-50. See also Barron, Schervish and Wasserman (1999), or Hawthorne (1993) for discussion of conditions that guarantee convergence.

[5]Rather, no prior fully amerliorates the luck involved in avoiding misleading evidence without rendering facts about evidence irrelevant to explaining posterior accuracy *altogether*. If a gambler adopts a close-to-perfectly dogmatic prior — one which is nearly perfectly resilient with respect to nearly all data — then of course the accuracy or inaccuracy of her posterior estimate of the coin's bias will depend minimally on the misleading nature of her new evidence. But it does so because it depends minimally on the character of *any* new data. The moral: no prior singles out and fully ameliorates *exactly* the sort of epistemic luck involved in avoiding misleading evidence.

[6]By the uniform prior $u$ over hypotheses $B = x$, I mean the prior $u$ defined by the uniform density function $f(x) = 1$.

[7]A beta prior $b$ is a probability distribution defined by a density function of the form $f(x) = ((1-x)^{-1+\beta}x^{-1+\alpha})/Beta[\alpha,\beta]$. Beta distributions are characterizable in terms of $\alpha$ and $\beta$, and hence, fairly computationally tractable. They also form a very flexible class of distributions. For these two reasons, we restrict our attention to beta distributions in many of our examples.

[8]For simplicity, I only consider sequences of coin tosses that are exchangeable, from the perspective of the true chance distribution.

[9]The cumulative distribution function $P$ corresponding to a distribution $p$ over chance hypotheses (defined by density $f$) is defined by $P(ch(X) \leq x) = \int_0^x f(y)\,dy$, and specifies the probability that the chance of $X$ is less than or equal to $x$.

[10]For example, for any beta densities $f$, $g$ and $h$, if they all have the same mean but increasing variance, then $f$ is closer to $g$ than to $h$. Similarly, if they all have the same variance but larger and larger means, then $f$ is closer to $g$ than to $h$.

[11]The usual caveats are needed: there is no demon intervening to make the bottom knob causally inefficacious, except when it's turned all the way to the right, or anything of the sort. If that were the case, of course, then the position of the bottom knob might well be relevant to explaining why London gets 3mm of rain, rather than 2mm, 1mm, etc., despite not being counterfactually relevant.

[12]Save, of course, for the fact that, at the end of the day, your experiment produced exactly the outcome that it did.

[13]Of course, when $B \approx 0$ or $B \approx 1$, this distribution will concentrate probability almost exclusively on one value for $\mathfrak{C}(u_D, H)$.

[14]It would be better to minimize $g(p) = max_{i,j} d(ch_{H_i}, ch_{H_j})$, where $ch_{H_i}$ is the objective marginal distribution for $d(p_D, H_i)$ determined by $H_i$. But the extra layer of complexity that this would add would, I suspect, obscure the more important, underlying philosophical point. It would draw attention away from the anti-luck rationale undergirding the MaxSen method.

[15]I evaluated $f(p) = max_i\ Exp_{H_i}(d(p_D, H_i))$ - $min_i\ Exp_{H_i}(d(p_D, H_i))$ at all beta distributions $p$ with $\alpha, \beta \in \{0.001, 0.25, 0.5, 0.75, ..., 2\}$. For the raw data, the details of the polynomial ($n = 5$) least-squares fit of the data, or the Mathematica code needed to run the simulations, please email jpkonek@gmail.com.

[16]Of course, like any version objective Bayesianism, the MaxSen method will not be applicable in *all* realistic problems of inference or decision.

[17]Nonparametric Bayesian models specify a joint distribution over an infinite number of parameters, *e.g.*, each of the uncountably many values of a probability density function.

# Chapter 2

## Precise Priors Without Total Comparative Probability

In the casino, your evidence might be perfectly specific. You might, for example, know exactly which cards have been dealt, exactly which cards your opponent needs to beat you on the river, and so have evidence that justifies a perfectly precise credence, say 3/52, that she will win and take the rest of your money. But evidence is often unspecific and equivocal. Consider your current evidence that it will snow next new year's eve, or your evidence about the price of copper twenty years from now, or about interest rates on home equity loans forty years from now. "About these matters," Keynes says, "there is no scientific basis on which to form any calculable probability whatever" (Keynes 1937, 213-4).

Imprecise Bayesians say that when your evidence is unspecific and equivocal, your opinions should be unspecific and equivocal too. Precise priors (single probability functions), however, do not allow for such opinions. If you adopt a precise prior, in an attempt to incorporate unspecific/equivocal evidence in an inference or decision problem, you will be stuck with perfectly specific opinions. Taking account of unspecific and equivocal prior evidence requires imprecise priors (sets of probability functions). Call this *the preclusion problem* for precise Bayesianism.

*Preclusion Problem.* Precise priors preclude unspecific and equivocal opin-

ions, and so invariably capture improper responses to unspecific and equivocal evidence.

The preclusion problem captures the central epistemic motivation for imprecise Bayesianism. My aim in this chapter is to demonstrate how flexible precise priors are. Despite first appearances, precise priors *do* indeed allow for unspecific and equivocal opinions.

In §2.1, I sketch the Bayesian approach to inductive inference. In §2.2, I detail the preclusion problem. In §2.3, I identify the background assumption that generates the preclusion problem, which I call *locality*. In §2.4-2.6, I present a number of reasons for doubting locality. In §2.7, I outline a broadly Bayesian, locality-free approach to inductive inference, and show that it avoids the preclusion problem. Finally, in §2.8, I explore independent reason for thinking that this approach, or something much like it is correct. If it is, this undercuts the central epistemic motivation for introducing imprecise priors. This, in turn, provides impetus to search for new epistemic foundations for imprecise Bayesianism.

## 2.1 The Bayesian Approach to Inductive Inference

All Bayesians agree on certain facts about inductive inference. They agree, for example, that when a researcher designs and performs an experiment aimed at adjudicating between competing theoretical hypotheses, $H_1, ..., H_n$, she ought to (i) take her prior evidence $E$ for $H_1, ..., H_n$ into account by adopting a 'prior', which somehow summarizes the information in $E$, (ii) update that 'prior' on her experimental data, to obtain a 'posterior', and (iii) read her new opinions about $H_1, ..., H_n$ (as well as the propositions $X$ that $H_1, ..., H_n$ render more or less likely) off of this 'posterior'. They also agree that (iv) probabilities are useful for constructing priors; constructing a

prior involves specifying a probability distribution $p$ (a precise prior), or a set of distributions (an imprecise prior) over $H_1, ..., H_n$. Finally, they agree, for the most part, on (v) which comparative and qualitative judgments any 'posterior' (precise or imprecise) commits its bearer to making; they agree on how to 'read off' new opinions from a posterior.[18]

Imagine, for example, a virologist who designs and performs an experiment to adjudicate between competing hypotheses $H_1, ..., H_n$ about a protein interaction in a virus. ($H_1, ..., H_n$ might be causal models that represent how this interaction works.) She comes to the table with a great deal of *prior* information, of course, *e.g.*, information about how these sorts of interactions work in similar viruses. Then her experiment yields *new* data. On the Bayesian view, to take her prior evidence $E$ into account in her inference problem, she ought to adopt a prior over $H_1, ..., H_n$. Different Bayesians, however, prescribe adopting different priors.

*Subjective* Bayesians say that agents ought to look to their own opinions to furnish priors. If an agent's opinions are rich enough to pin down a single truth-value estimate for each of the $H_1, ..., H_n$, then she ought to adopt a *precise* prior, a single probability distribution $p$ over $H_1, ..., H_n$ that summarizes $E$ (*viz.*, the prior $p$ that encodes 'her' truth-value estimates).[19] If, in contrast, her prior opinions fail to pin down a single probability distribution $p$ over $H_1, ..., H_n$, then she ought to adopt an *imprecise* prior, or a set of a probabilities. In particular, she ought to adopt the set of probabilities that are (rationally) compatible with her comparative and qualitative judgments (rationally permissible to adopt given those judgments).

*Objective* Bayesians, in contrast, endorse methods for constructing priors that do not depend, in the same way, on the agent's prior opinions (and inductive quirks, hunches, etc.). Edwin Jaynes (1957, 1968, 1973), for example, endorses the *maximum entropy method* (MaxEnt):

- Summarize your prior evidence by constraints $C_1, ..., C_n$, which you model by a set of probability distributions $\mathscr{C}$.

- Adopt the prior $p$ that maximizes entropy $H(p) = -\sum_i p(H_i) \cdot log(p(H_i))$ on $\mathscr{C}$.

On this view, any researcher who arrives at the same evidential constraints should proceed in the same manner. She should adopt a *precise* prior, *viz.*, the probability distribution $p$ over $H_1, ..., H_n$ that maximizes entropy on the set of probabilities that satisfy those constraints. See Kass and Wasserman (1996) for an overview of precise, objective Bayesian approaches to inductive inference. Alternatively, objective Bayesians might prescribe adopting a particular *imprecise* prior. Jeffrey (1983) and Dalkey (1985), for example, propose measures of entropy for imprecise models (sets of probabilities).[20] Imprecise objective Bayesians might prescribe adopting the maximum entropy imprecise prior consistent with your evidence (*viz.*, $\mathscr{C}$ itself).

So different Bayesians prescribe adopting different priors. They agree, nonetheless, about how inductive inference proceeds once an agent has a prior in hand. They agree, for example, that our virologist ought to proceed by updating her prior on her new, experimental data $D$. If she adopts a precise prior, *i.e.*, a single probability distribution $p$ over hypotheses $H_1, ..., H_n$ (about the relevant protein interaction) that summarizes her prior evidence $E$ (about how these sorts of interactions work in similar viruses, etc.), then this involves conditioning $p$ on $D$, *i.e.*, adopting the posterior $p_D(\cdot) = p(\cdot|D)$. If she adopts an imprecise prior, a set of probabilities $S$, then this involves conditioning every $p$ in $S$ on $D$. She ought to then 'read off' her new opinions about $H_1, ..., H_n$ from her posterior. If her prior $p$ is precise, this involves making the comparative and qualitative judgments that her posterior $p_D$ rationally commits her to making, according to Bayesian orthodoxy:

- She is committed to judging that $X$ is more plausible than $Y$ if $p_D(X) > p_D(Y)$.

- She is committed to judging that $D$ provides positive incremental support for $X$ if $p_D(X) > p(X)$.
- She is committed to judging that action $A$ is preferable to $B$ if the expected utility of $A$ (from the perspective of $p_D$) is greater than the expected utility of $B$ (from the perspective of $p_D$).

If her prior $S$ is imprecise, reading new opinions off of her posterior $S_D$ involves making the comparative and qualitative judgments that $S_D$ is univocal about, *i.e.*, the judgments that *all* elements of $S_D$ commit her to making.

This last bit of orthodoxy — about which comparative and qualitative judgments a precise prior/posterior commits you to making — though extremely entrenched, is also extremely implausible. It forces agents to ignore a great deal of information about the *quality of their evidence* — in particular, the *weight* of their evidence — in inquiry and decision-making. This makes for bad inductive and practical policy in a wide range of contexts, I will argue.

The reason this matters: *this bad bit of orthodoxy generates the preclusion problem.* Fixing this bug, I will argue, reveals how flexible precise priors are. Fixing this bug shows that precise priors plausibly capture adequate responses to unspecific and equivocal evidence.

## 2.2 The Preclusion Problem

### 2.2.1 The Basic Issue

The question that divides precise and imprecise Bayesians is this: should you invariably use a single probability distribution to incorporate your prior information in inference and decision problems? Or are there circumstances in which imprecise priors (sets of distributions) are called for? Precise Bayesians say that you should

invariably adopt a precise prior. Imprecise Bayesians say that there are circumstances in which imprecise priors are called for.

It may seem obvious that there are circumstances in which imprecise priors are called for. After all, you might think, the primary theoretical role of priors is just to represent an agent's actual prior opinions about the plausibility of hypotheses. (This is a natural enough subjectivist thought.) And normal researchers' actual prior opinions often fail to pin down a single truth-value estimate for each of the theoretical hypotheses $H_1, ..., H_n$ under investigation. Here, for example, are Kyburg and Pittarelli:

> Suppose that the judgments "$A$ is at least as probable as $B$" and "$B$ or $C$ is at least as probable as $A$" are made for mutually exclusive and exhaustive events $A$, $B$, and $C$. Any of the infinitely many solutions to the system of linear inequalities
>
> $$p(A) + p(B) + p(C) = 1$$
> $$p(A) \geq p(B)$$
> $$p(B) + p(C) \geq p(A)$$
>
> for example
>
> $$p(A) = 0.2,\ p(B) = 0.1,\ p(C) = 0.7$$
>
> is compatible with these judgments. If nothing stronger than these comparisons is forthcoming, then there is no basis for choosing a single one of these functions as representative of the probability information. (Kyburg and Pittarelli 1996, 325)

An agent's comparative probability judgments $\preceq$ 'leave open' any distribution $p$ that weakly represents them, *i.e.*, is such that $H_i \preceq H_j$ only if $p(H_i) \leq p(H_j)$

(in the sense that $p$ is not impermissible to adopt simply in virtue of her making judgments $\preceq$). And, as Kyburg and Pittarelli stress, a normal researcher's actual opinions normally 'leave open' many distributions. When they do, the imprecise prior $S$, which contains all of these 'open' distributions $p$, best represents her actual prior opinions. If the primary theoretical role of priors is just to represent actual prior opinions, then she ought to adopt that imprecise prior $S$.

Alternatively, one might contend that the primary theoretical role of priors is to represent the opinions about the plausibility of hypotheses that are *best supported* by her prior evidence. If this is right, it is no longer *obvious* that certain circumstances call for imprecise priors. The mere fact that *actual* researchers have less than maximally specific prior opinions (prior opinions that fail to pin down precise truth-value estimates) no longer settles the dispute. The important question to ask now is: do certain bodies of prior evidence *support* less than maximally specific states of opinion? If so, then some evidential circumstances call for imprecise priors. If not, then not.

Imprecise Bayesians such as Levi (1980), Walley (1991) and Joyce (2005) argue that certain bodies of evidence do indeed support less than maximally specific states of opinion. In particular, when your evidence is unspecific and equivocal, your opinions should be unspecific and equivocal too. The upshot: on either account of the theoretical role of priors, certain circumstances call for imprecise priors. Here, for example, is Walley:

> If there is little evidence concerning [a hypothesis,] then beliefs about [that hypothesis] should be indeterminate, and probability models imprecise, to reflect the lack of information. (Walley 1991, 212-3)

And here is Joyce:

> ...the proper response to symmetrically ambiguous or incomplete evidence

> is not to assign probabilities symmetrically, but to refrain from assigning precise probabilities at all... Imprecise credences have a clear epistemological motivation: they are the proper response to unspecific evidence. (Joyce 2005, 171)

*Unspecific* evidence regarding a proposition $X$ is evidence that fails to discriminate $X$ from incompatible alternatives $X'$ (Joyce 2005, 167). For example, your evidence about interest rates $R$ on home equity loans forty years from now might be very specific with respect to the claim that $R$ will be higher than 15% (it might nearly rule it out), but be relatively unspecific with respect to the claim that $R$ will be exactly 3% (it might fail to discriminate that claim from competitors, *e.g.*, $R = 2.9$, $R = 3.1$, etc.). *Equivocal* evidence regarding $X$ is evidence that is open to different readings, and whose significance for $X$ varies on those different readings (*ibid.*). For example, your evidence about whether you will have health problems later in life might be equivocal if you have some alarming symptoms, but very little information about the underlying condition causing them (perhaps your symptoms are equally plausible on a range of hypotheses about their cause). On different suppositions about the underlying condition, the significance of your current symptoms (and family history, etc.) shifts; it tells a different story, so to speak, about whether you will have health problems later in life.

To see why imprecise Bayesians like Walley, Joyce and others hold that unspecific and equivocal prior evidence calls for imprecision in one's prior probabilities, consider a case adapted from Williamson (2010, 116-20). An oncologist prescribes hormonal treatment $T$ to a breast cancer patient. She performs a test to determine whether the patient's tumor is estrogen-receptor-positive (ER+). She also has auxiliary evidence about $R$, whether her patient's breast cancer will recur given $T$, which includes (i)

data about the patient's symptoms, (ii) data from clinical databases about age, tumor size, survival time, etc. of past patients, (iii) quantitative molecular data on tumor cells, etc.

Our oncologist has a wealth of prior information relevant to $R$, much more than we typically have about other matters: whether it will snow next new year's eve, what the price of copper will be twenty years from now, how high/low interest rates on home equity loans will climb/fall forty years from now. This relatively weighty evidence might be specific enough to impose the following constraints on any prior that summarizes it:

- the tumor is at least 9/10-likely to be estrogen-receptor-positive;
- the patient's cancer is 1/4-likely to recur given that her tumor is estrogen-receptor-positive ($T$ is a fairly effective treatment for estrogen-receptor-positive tumors);
- the patient's cancer is 3/4-likely to recur given that her tumor is estrogen-receptor-negative ($T$ is much less effective for estrogen-receptor-negative tumors).

If her evidence does impose these constraints, then it plausibly commits her to making certain comparative and qualitative judgments, *e.g.*, "It is more probable that the patient's tumor is estrogen-receptor-positive than it is that there will be an earthquake in London today." But, it does *not* commit her to making certain other judgments. She is not committed to judging, "$ER+$ is more probable than drawing a black ball at random from an urn containing 92 black balls and 8 red balls," just as you or I am not committed to judging, "Snow next new year's eve is more probable than drawing a black ball at random from an urn containing 43 black balls and 57 red balls." Even

weighty, quantitative evidence, such as our oncologist's, is too unspecific to be this demanding. Our oncologist's evidence only commits her to a *partial* comparative probability ordering, not a *total* ordering. An agent's comparative probability ordering $\preceq$ is total if it is a partial order (reflexive, antisymmetric, transitive) and also satisfies *totality*: $X \preceq Y$ or $Y \preceq X$, for all $X$ and $Y$. In this sense, her less than maximally specific evidence at least permits (and perhaps positively requires) less than maximally specific (merely partial) opinions, or comparative/qualitative judgments.

Suppose, however, that our oncologist adopts some precise prior, in order to incorporate her prior information in her inference and decision problem (to help her figure out what to think about the prospects of sustained remission, whether to prescribe additional treatments, etc.). Perhaps she adopts the MaxEnt prior, *i.e.*, the prior $p$ that maximizes entropy on the set $S$ of priors $q$ that satisfy the constraints imposed by her evidence:

- $9/10 \leq q(ER+) \leq 1$
- $q(R|ER+) = 1/4$
- $q(R|ER-) = 3/4$.

Then her prior probabilities for $ER+$ and $R$ are $p(ER+) = 0.9$ and $p(R) = 0.3$, respectively. But, according to Bayesian orthodoxy, this means that she is committed to making *exactly the sorts of comparative probability judgments that we claimed she need not make, given her evidence.*

- She must judge that it is definitely less probable that the patient's tumor is estrogen-receptor-positive than it is that she will select a black ball if she randomly draws from an urn containing 92 black balls and 8 red balls.

- She must judge that it is definitely more probable that her patient's cancer will recur than it is that she will see 3 heads if she flips a fair coin 7 times.

Something stronger is true, in fact. *All* precise priors $p$ are such that $p(X) \leq p(Y)$ or $p(X) \geq p(Y)$, for any $X$ and $Y$. And since an agent who adopts $p$ is committed to judging $X \preceq Y$ if $p(X) \leq p(Y)$, according to Bayesian orthodoxy, it follows straightaway that your comparative probability judgments $\preceq$ must form a *total* order if you adopt such a prior. That is, you are either committed to judging $X \preceq Y$ or $Y \preceq X$, for any $X$ and $Y$. (Similarly, since the expectation operator $Exp_p$ totally orders actions, your preferences must form a total order as well.) But having perfectly specific opinions — a total comparative probability ordering, total preferences, etc. — is the wrong way to respond to unspecific and equivocal evidence.

This is why imprecise Bayesians like Walley, Joyce and others hold that unspecific and equivocal prior evidence calls for imprecision in one's prior probabilities. If our oncologist adopts an imprecise prior, in order to incorporate her prior information in her inference and decision problem, she is *not* necessarily committed to comparative and qualitative judgments that all form total orders. Suppose, for example, that she adopts the set $S$ of distributions that satisfy the constraints imposed by her evidence. The marginally unspecific nature of her prior evidence for $R$ (the proposition that her patient's breast cancer will recur) is reflected in the spread $\{p(R)|p \in S\} = [0.25, 0.3]$, on the imprecise Bayesian view. The greater this spread, typically, the fewer comparative and qualitative judgments she will be committed to making with respect to $R$. For example, if all $p$ in $S$ agree that the probability of observing 3 heads on 7 independent flips of a fair coin is 0.273 ($p(3H) = 0.273$), and some $p$ in $S$ say that the probability of the patient's cancer recurring is 0.25 ($p(R) = 0.25$), while other $p'$ in $S$ say that the probability is 0.3 ($p(R) = 0.3$) — and of course there *are* such

$p$ and $p'$ in $S$ — then our oncologist is neither committed to judging $R \preceq 3H$ nor $3H \preceq R$ (according to Bayesian orthodoxy). She is only committed to making the comparative and qualitative judgments that $S$ is univocal about, *i.e.*, the judgments that *all* elements of $S$ commit her to making.

The moral: agents who adopt imprecise priors are typically committed to *merely partial* comparative probability judgments, comparative preferability judgments, judgments of incremental support, etc. They are permitted to abstain from judgment on various issues. In this way, imprecise priors allow for genuinely unspecific and equivocal opinions.

### 2.2.2 Adopting a Prior vs. Being Representable by a Prior

The remainder of this chapter is devoted to demonstrating just how flexible precise priors are. Agents who adopt precise priors are typically committed to *merely partial* comparative and qualitative judgments, I claim. They are permitted to abstain from judgment on various issues. Precise priors are flexible enough, then, to allow for genuinely unspecific and equivocal opinions.

We ought to address one concern now, though, at the outset. One might worry that it betrays a rather basic confusion to suggest that a precise prior could allow for a merely partial comparative probability ordering. The reason: to count as *adopting* a precise prior, one must *already have* a total comparative probability ordering. Adopting a precise prior is just *equivalent*, one might suggest, to having a comparative probability ordering that is rich enough to pin down a single truth-value estimate for each of the theoretical hypotheses under investigation $H_1, ..., H_n$. Scott (1964) shows us just what this 'richness' amounts to. A comparative probability ordering $\preceq$ is representable by a unique probability distribution $p$, in the sense that $X \preceq Y$ only if $p(X) \leq p(Y)$, if and only if $\preceq$ satisfies Scott's axiom:

> *Scott's Axiom.* If $\langle X_1, ..., X_n\rangle$ and $\langle Y_1, ..., Y_n\rangle$ contain the same number of truths as a matter of logic, so that $\sum_i w(X_i) = \sum_i w(Y_i)$ for any world $w$, then it is not true that $X_i \preceq Y_i$ for all $i$ while $X_j \prec Y_j$ for some $j$.

and, in addition, satisfies two other structural axioms: totality and non-atomicity (*cf.* Scott 1964, p. 246; Joyce 2010, p. 285).[21] The upshot: any comparative probability ordering that is 'rich enough' to pin down a single truth-value estimate for each of $H_1, ..., H_n$ *must already be total.* It is nonsense, then, to suggest that an agent *should* adopt a precise prior, in order to incorporate her prior information in her inference or decision problem, and yet is *not* rationally required to have total comparative beliefs.

This worry runs together two distinct notions: *having* or *being representable by* a probability distribution $p$ over hypotheses $H_1, ..., H_n$, on the one hand, and *adopting* $p$ as one's prior, on the other hand. Distinguishing these notions resolves our objector's worry. **Having** or **being representable by** a probability distribution $p$ over theoretical hypotheses is a matter of having opinions (making comparative and qualitative doxastic judgments) that rationally commit you to estimating truth-values via $p$.[22] **Adopting** a distribution $p$ over theoretical hypotheses as one's prior, in contrast, is a matter of making the comparative and qualitative judgments that $p$ rationally commits you to making, of allowing $p$ to guide your inferential practices and decision-making, in this sense. Importantly, the comparative/qualitative judgments that a distribution commits you to making ***differ significantly*** from the judgments that commit you to estimating truth-values via that distribution (or so I will argue). It is perfectly possible for the former to be merely partial while the latter are total.

Richard Jeffrey (1987, p. 589) illustrates the distinction when, discussing the Ellsberg paradox (*cf.* §5). He says, "I think you do well to find a definite probability function to express your uncertainty, if you can... in the Ellsberg problems (were I ever

to face them) I think I would try to express my uncertainty via a single probability assignment — the uniform one, I imagine. If so, I differ from [Mark] Kaplan, who would see my adoption of the uniform distribution as unjustifiable precision, whereas *I think I would adopt it as a precise characterization of my uncertainty*" (emphasis mine). Jeffrey's point, of course, is not that you would do well to make some set of total, non-atomic, Scott's-axiom-satisfying comparative probability judgments, which commit you to estimating truth-values via some single probability assignment. "We humans are not capable of adopting opinions gratuitously, even if we cared to do so" (Jeffrey 1983, 145). The point, rather, is just that you would do well to *use* some probability assignment to facilitate decision-making in the Ellsberg problem, by making the qualitative and comparative judgments (judgments of comparative preferability, in this case) that it commits you to making (in a limited domain).

Adopting a precise prior is a common practice too. Objective Bayesian statisticians will attest that using MaxEnt to facilitate inquiry in computational biology, computer vision, or natural language processing (just a few of the areas where MaxEnt has proved enormously useful) does not require making an incredibly rich set of comparative and qualitative judgments. It only requires *using* the MaxEnt prior to guide your inferential practices and decision-making, by making the judgments that it commits you to making (in a limited domain).

Once we distinguish the notions of *having* or *being representable by* a probability distribution $p$ over theoretical hypotheses, on the one hand, and *adopting* $p$ as one's prior, on the other hand, our objector's worry dissolves. It is not confused to suggest that a researcher could (and perhaps should) adopt a precise prior, in order to incorporate her prior information in her inference or decision problem, despite not being representable by a precise prior. Neither is it confused to suggest that this researcher might not be rationally required to make comparative probability judgments that

form a total order, as a result of adopting such a prior.

## 2.3 Rational Commitment

According to Bayesian orthodoxy, an agent who adopts a precise prior $p$ is committed to making the following comparative judgments:

- She is committed to judging that $X$ is more probable than $Y$ if $p(X) > p(Y)$.
- She is committed to judging that data $D$ provides positive incremental support for $H$ if $p_D(H) > p(H)$.
- She is committed to judging that action $A$ is preferable to $B$ if the expected utility of $A$ (from the perspective of $p$) is greater than the expected utility of $B$ (from the perspective of $p$).

This bit of orthodoxy is implausible, however. To show this, I will first identify some *pro tanto* reason to expect it to be false (in §2.3-2.5). I will then turn to the main argument argument against it (in §2.6-2.8).

We ought to expect the orthodoxy about rational commitment to be false because it directs agents to ignore a great deal of information about the quality of their evidence — in particular, the *weight* of their evidence — when making comparative and qualitative judgments, judgments which not only have epistemic value in their own right, plausibly, but also structure subsequent inquiry, and so have downstream epistemic consequences. Reasonable agents, however, take *all* information about the quality of their evidence into account for these purposes.

To see this, note that according to orthodoxy, whether $p$ carries a commitment to judging $X \succ Y$ (that $X$ is more plausible than $Y$) depends exclusively on $p(X)$ and $p(Y)$. These probabilities, however, are merely the first moments of certain marginal

distributions (they are expected values, *viz.*, the expected objective probabilities of $X$ and $Y$, respectively); $p(X) = \sum_i p(X|H_i) \cdot p(H_i) = \sum_x x \cdot p(Q_X = x)$, where $Q_X = x$ if and only if some theoretical hypothesis $H_i$ with $p(X|H_i) = x$ is true. (When considering uncountably many theoretical hypotheses, $p(X) = \int_0^1 x \cdot f_{Q_X}(x)\,dx$, where $f_{Q_X}$ is the density that defines the marginal distribution of $Q_X$.) And the higher moments of these distributions (variance, skewness, etc.) plausibly reflect important information about the quality of one's evidence (in particular, the weight of one's evidence) for $X$ and $Y$.

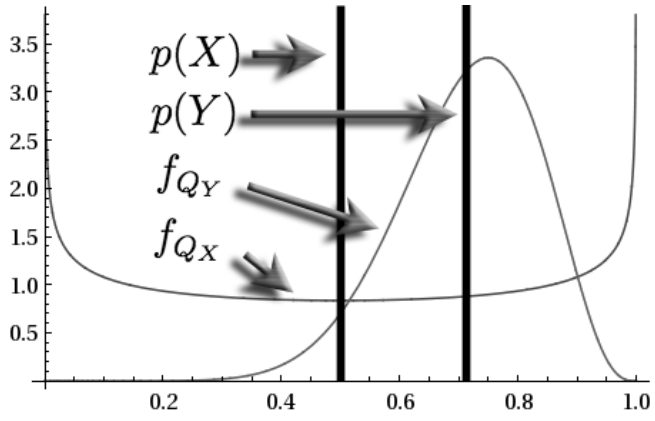


Figure 2.1: $f_{Q_X}$ and $f_{Q_Y}$, as well as their means, $p(X)$ and $p(Y)$.

Similarly, whether $p$ carries a commitment to judging that data $D$ provides incremental support for $X$ depends exclusively on $p(X)$ and $p(X|D)$, according to the orthodox account. These probabilities, however, are merely the first moments of certain marginal distributions; $p(X) = \sum_i p(X|H_i) \cdot p(H_i) = \sum_x x \cdot p(Q_X{=}x)$ and $p_D(X) = \sum_i p_D(X|H_i) \cdot p_D(H_i) = \sum_x x \cdot p_D(V_X = x)$, where $V_X = x$ if and only if some theoretical hypothesis $H_i$ with $p_D(X|H_i) = x$ is true.[23] Again, the higher moments of these distributions plausibly reflect important information about the quality (weight) of the evidence for $X$ (before and after learning $D$).

Finally, whether $p$ carries a commitment to judging that $A$ is preferable to $B$ depends exclusively on $Exp_p(A)$ and $Exp_p(B)$, according to orthodoxy. These expectations, once more, are merely the first moments of certain marginal distributions;

they are the (evidential, let's suppose) expected utilities of $A$ and $B$, respectively. $Exp_p(A) = \sum_w u(w) \cdot p_A(w) = \sum_x x \cdot p_A(u = x)$.[24] The higher moments of these distributions might reflect important information about the quality of one's evidence that $A$ and $B$ will produce good outcomes.

The general theme is this: for each type of comparative judgment $J_{X,Y}$ between $X$ and $Y$, the orthodox account of rational commitment says that there are certain marginal distributions, $f$ and $f'$, that encode the information relevant for determining whether $p$ commits its bearer to making the judgment $J_{X,Y}$. Moreover, that information is encoded in a specific 'spot', so to speak, in $f$ and $f'$. It is encoded *locally*, we might say, in the mean, or first moment of $f$ and $f'$, respectively. It is not encoded *globally*, across all of the moments of $f$ and $f'$ (mean, variance, skewness, etc.). Call this *the locality thesis*.

We ought to expect the locality thesis to be false. Information about the quality of one's evidence — in particular, the weight of one's evidence — for $X$ and $Y$, is distributed across all of the moments of $f$ and $f'$ (mean, variance, skewness, etc.), as we will see in §4. And such information is plausibly relevant for determining whether and which comparative/qualitative judgments you are rationally committed to making, in virtue of adopting $p$, as we will see in §2.5 and §2.6.

## 2.4 How Probabilities Reflect Weight

The weight of an agent's total evidence for a proposition is a matter of "how much relevant information the data contains, irrespective of which way it points" (Joyce 2005, 159). Keynes introduces the notion of weight as follows:

> As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, according as

> the new knowledge strengthens the unfavourable or the favourable evidence; but *something* seems to have increased in either case, — we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the *weight* of an argument. New evidence will sometimes decrease the probability of an argument, but it will always increase its 'weight'. (Keynes 1921, 77)

To illustrate, consider Popper's paradox of ideal evidence (Popper 1959, 425-7). A bookie hands you a coin and offers you a bet. You have no prior evidence about the coin's bias. Before you decide what to do, the bookie hits you over the head and knocks you out. When you come to, she reports that she flipped it 1000 times, and that it came up heads $500 \pm 20$ times. To take account of your prior information (*viz.*, none) in your decision problem, you decide to adopt the maximum entropy prior $u$ over hypotheses $B = x$ about the coin's bias. You then condition $u$ on your new data $D$.

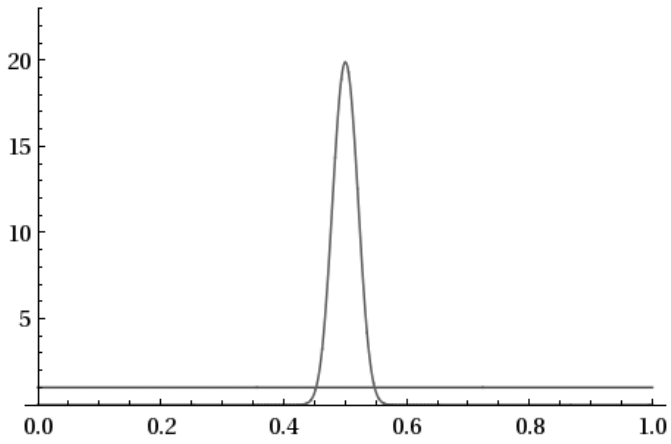


Figure 2.2: $u$ and $u_D$.

After you receive your new data, you have weightier evidence about whether the coin will come up heads on its next flip (and so weightier evidence about whether you will make or lose money if you take the bookie's bet). You have a "more substantial basis upon which to rest [your] conclusion," as Keynes says. But this is not reflected in your probability for $H$ (the proposition that the coin will come up heads on the

next flip). Both your prior $u(H)$ and posterior probability $u_D(H)$ equal 1/2 (right).[25]

Richard Jeffrey notes that while $u$ and $u_D$ assign the same probability to $H$, they nonetheless "assign different values to any proposition $A(n)$ that asserts, concerning $n \geq 2$ distinct tosses, that all of them yield heads" (Jeffrey 1965, 184). For example, $u(A(6)) = 1/7 \approx 0.143$ and $u_D(A(6)) \approx 0.016 \approx 1/2^6$. The reason: $u_D$ is much more *resilient* than $u$, much more steadfast in the face of new data. This is reflected in the fact that $u_D(H|X)$ is close to $u_D(H)$ for a wide range of potential new data items $X$. For example, suppose that you are going to flip the coin 6 times. Let $H_i$ be the proposition that it comes up heads on the $i^{th}$ toss. Then we have $u_D(H_1) = 1/2$, $u_D(H_2|H_1) \approx 1/2$, $u_D(H_3|H_1\&H_2) \approx 1/2$, etc. $U_D$ is resilient; conditioning on new data — $H_1$, $H_1\&H_2$, $H_1\&H_2\&H_3$, etc. — does not alter the probability of observing a heads on the next flip very much. That is why $u_D(A(6)) = u_D(H_1\&...\&H_6) = u_D(H_1) \cdot u_D(H_2|H_1) \cdot u_D(H_3|H_1\&H_2) \cdot ... \cdot u_D(H_6|H_1\&...\&H_5) \approx 1/2^6$.

Skyrms sums up Jeffrey's view as follows: "In a word, the ideal evidence" — extremely *weighty* evidence — "has changed not the *probability* of tails on toss $a$, but rather the *resiliency* of the probability of tails on toss $a$" (Skyrms 1977, 707). The characteristic effect of weighty evidence is to render one's posterior resilient with respect to new data. This is a matter of stabilizing its conditional probabilities (making $u_D(H|X)$ close to $u_D(H)$ for a wide range of $X$). The important point to note, for our purposes, is that while $u$'s unconditional probability for $H$ depends exclusively on its first moment (mean) — $u(H)=\int_0^1 x \cdot f(B=x)dx=\int_0^1 x\,dx=1/2$ — its conditional probabilities are a function of *all of its moments* (mean, variance, skewness, etc.). To see this, compare the probability that $u$ assigns to $H$ conditional on $D$ with the probability that lower/higher variance beta priors assign (pictured left, next page), for a range of data sequences $D$:

Table 2.1: Conditional probabilities across priors of decreasing variance.

| | $u$: $\alpha = \beta = 1$<br>*mean*: 0.5<br>*variance*: 0.83 | $b$: $\alpha = \beta = 0.5$<br>*mean*: 0.5<br>*variance*: 0.125 | $b^*$: $\alpha = \beta = 10$<br>*mean*: 0.5<br>*variance*: 0.01 |
|---|---|---|---|
| $D = H^7T^3$ | $u_D(H) = 0.667$ | $b_D(H) = 0.682$ | $b^*_D(H) = 0.567$ |
| $D = H^5T^{20}$ | $u_D(H) = 0.222$ | $b_D(H) = 0.212$ | $b^*_D(H) = 0.333$ |
| $D = H^1T^{49}$ | $u_D(H) = 0.038$ | $b_D(H) = 0.029$ | $b^*_D(H) = 0.157$ |

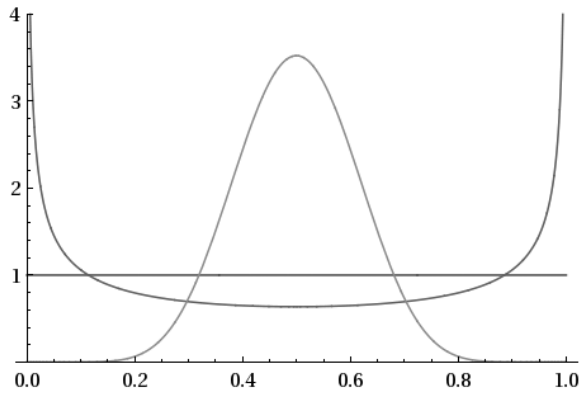


Figure 2.3: $u$, $b$ and $b^*$.

The probability of $H$ conditional on $D$ varies as you move from $u$ to $b$ to $b^*$, despite the fact that the first moment (mean) of each distribution is the same (=0.5). How resilient these priors are with respect to data sequences $D$, then, depends not just on their first moments (means), but on their higher moments as well (variance, etc.). This means that higher moments encode information about the weight of one's evidence, on Jeffrey's view.

Joyce (2005) suggests that the characteristic effect of weight is somewhat different. Weighty evidence for $H$ tends to cause a prior $p$'s probabilities to "concentrate more and more heavily on increasingly smaller subsets of chance hypotheses" (Joyce 2005, 167). On Joyce's picture, this effect is measured roughly by how small the following quantity is, across potential data sequences $D$:

$$w_p(H, D) = \int_0^1 |f(B = x) \cdot (x - p(H))^2 - f_D(B = x) \cdot (x - p_D(H))^2| dx$$

(where $f$ is the density that defines $p$). Importantly, $f(B = x) \cdot (x - p(H))^2$ tends to be small if $p$ concentrates probability on a small, connected subset of chance hypotheses

(hypotheses about the coin's bias, $B = x$). The reason: $f(B = x)$ is quite small when $(x - p(H))^2$ is large; $(x - p(H))^2$ is quite small when $f(B = x)$ is large. Similarly, $f_D(B = x) \cdot (x - p_D(H))^2$ tends to be small too. The upshot: $w_p(H, D)$ is close to zero, for a wide range of data $D$.

The important point to note, for our purposes, is that $w_p(H, D)$ is a function of $p$'s higher moments. To illustrate this, compare the values that this quantity takes relative to the uniform prior $u$, as opposed to lower/higher variance beta priors $b$ and $b^*$, across a range of data sequences $D$:

Table 2.2: Joyce's measure of weight across priors of decreasing variance.

| | $u$: $\alpha = \beta = 1$<br>*mean*: 0.5<br>*variance*: 0.83 | $b$: $\alpha = \beta = 0.5$<br>*mean*: 0.5<br>*variance*: 0.125 | $b^*$: $\alpha = \beta = 10$<br>*mean*: 0.5<br>*variance*: 0.01 |
|---|---|---|---|
| $D = H^7T^3$ | $w_u(H, D) = 0.078$ | $w_b(H, D) = 0.121$ | $w_{b^*}(H, D) = 0.007$ |
| $D = H^5T^{20}$ | $w_u(H, D) = 0.079$ | $w_b(H, D) = 0.121$ | $w_{b^*}(H, D) = 0.011$ |
| $D = H^1T^{49}$ | $w_u(H, D) = 0.083$ | $w_b(H, D) = 0.125$ | $w_{b^*}(H, D) = 0.012$ |

Joyce's quantity $w(H, D)$ varies as you move from $u$ to $b$ to $b^*$, despite the fact that the first moment (mean) of each distribution is the same (=1/2). The value of this quantity depends not just on the first moments (means) of the respective priors, but on their higher moments as well (variance, etc.). This means that higher moments encode information about the weight of one's evidence on Joyce's view as well.

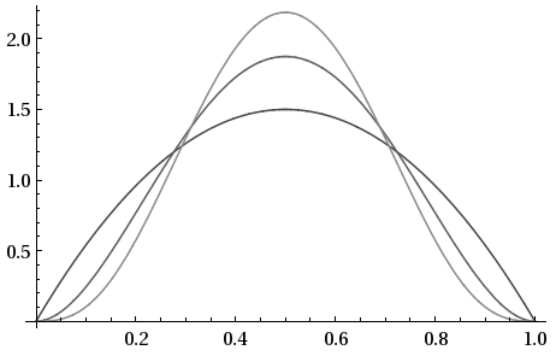


Figure 2.4: A sequence of mean-preserving spreads.

There are various other accounts of how prior (and posterior) probabilities reflect the weight of evidence. You might, for example, eschew quantitative measures of

weight altogether. Instead, you might borrow something like the notion of a mean-preserving spread from Machina and Rothschild (1990). Machina and Rothschild introduce the notion as follows (1990, p. 233):

> Intuitively, such a spread consists of moving probability mass from the centre of a probability distribution to its tails in a manner which preserves the expected value of the distribution.

You might then suggest that a prior $p$ reflects weightier evidence for $X$ than $p'$ if $p'$'s marginal for $X$ is a mean-preserving spread of $p$'s. If this is right, then again, higher moments (variance, skewness, etc.) encode information about the weight of one's evidence.

Whether or not any one of these proposals is fully adequate is beside the point. The point is just this: on *any* plausible account of how priors $p$ reflect the weight of one's evidence — Jeffrey's, Joyce's, or some alternative account — information about weight is not encoded 'locally', exclusively in $p$'s first moment (mean). It encoded 'globally', across all of $p$'s moments (mean, variance, skewness, etc.).

## 2.5 Decision-Making and the Weight of Evidence

Gärdenfors and Sahlin (1982, pp. 361-2) consider a case much like the following, in order to illustrate the importance of the weight of one's evidence for determining which practical (comparative preferability) judgments she is committed to making. Julie sits down to watch a tennis match between players 1 and 2. The players are to play a fixed number $N$ of games. At the outset, Julie has no information about players 1 and 2 (does not know their strengths and weaknesses, their track records against similar opponents, their present physical conditions, etc.). Her friend sits down and offers her a bet $B$. $B$ costs \$45. But it pays out \$100$x$, where $x$ is the

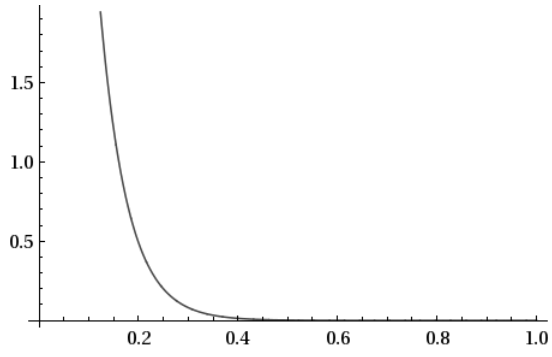


Figure 2.5: Prior distribution over chance hypotheses $ch(W_1 = x) = y$.

proportion of the $N$ games that player 1 wins. So, if player 1 wins every game, $B$ pays out \$100. If player 1 wins half of the games, $B$ pays out \$50. If player 1 wins $1/4$ of the games, $B$ pays out \$25, and so on.

Julie adopts some prior distribution $p$ over hypotheses $ch(W_1 = x) = y$ about the chance that player 1 will win a certain proportion $x$ of the games, to take account of her prior evidence (none) about player 1's prospects for winning, for the purposes of decision-making. She makes whatever judgment $p$ (together with her desires $u$) commits her to making and chooses accordingly. Suppose, for concreteness, that $p$ is the MaxEnt prior, that $u(\$k) = k$, and that $N = 18$ (they are going to play 3 6-game sets). So, for each $x$, her distribution over chance hypotheses $ch(W_1 = x) = y$ is given by the density $f_x(y) = 19e^{-19y}$ (left), her prior probability for $W_1 = x$ is $p(W_1 = x) = \int_0^1 y \cdot f_x(y)dy = 1/19$, and her expected utilities for accepting and declining, respectively, are $Exp_p(Accept\ B) = \sum_{i=0}^{18} p(W_1 = i/18) \cdot ((100i/18) - 45) = 5$ and $Exp_p(Reject\ B) = 0$. The upshot: accepting $B$ is, according to her best estimate, preferable to the status quo. So she accepts.

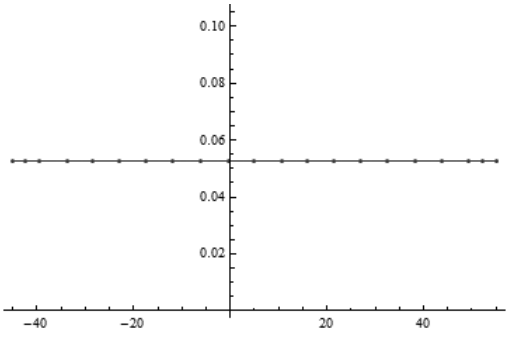


Figure 2.6: Prior distribution $p$ over utility hypotheses $u = x$ conditional on $B$.

Julie then acquires new data $D$. She watches players 1 and 2 for an entire day, and learns a great deal about their strengths and weaknesses, etc. She learns that

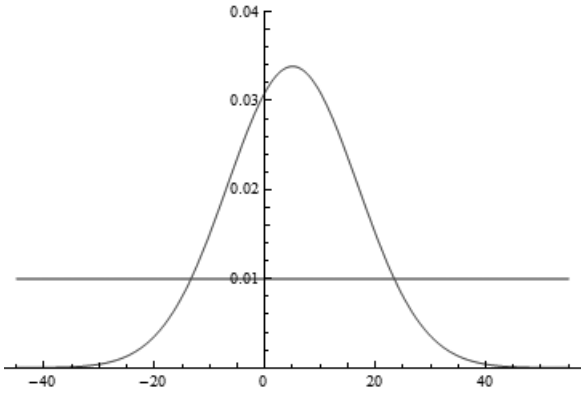


Figure 2.7: Prior and posterior distributions over utility hypotheses $u = x$ conditional on $B$.

while player 1 has an excellent service game, player 2 has an excellent return game. While player 2 is very effective at the net, player 1 has a very effective passing shot. They are very evenly matched. The next day, she sits down to watch another match. Her friend offers her bet $B$ again. In order to incorporate her new data in her decision problem, Julie conditions $p$ on $D$. The result: $p_D$ is nearly certain that players 1 and 2 have an even chance of winning each game; $p_D$ concentrates probability largely on $ch(W_1 = x) = 1$ if $x \approx 1/2$ and largely on $ch(W_1 = x) = 0$ if $x \not\approx 1/2$. This causes $p_D$ to concentrate probability around $W_1 = 1/2$, which in turn causes $p_D$ to concentrate probability around on $U$=5.

Importantly, *none of this is reflected in the expected utility of B* (just as in Popper's paradox of ideal evidence). Julie's prior expected utility for accepting bet $B$ is $Exp_p(Accept\ B) = \sum_{i=0}^{18}(1/19) \cdot ((100i/18) - 45) = 5$. Julie's posterior expected utility is:

$$Exp_{p_D}(Accept\ B) \approx \sum_{i=0}^{18} \begin{pmatrix} 18 \\ i \end{pmatrix} (1/2)^{18}((100i/18) - 45) = 5.$$

So Julie's best estimate of $B$'s utility remains unchanged (despite the fact that she has much weightier evidence undergirding that estimate). According to Bayesian orthodoxy, then, Julie ought to take bet $B$ on day 2, after acquiring a wealth of new data, if and only if she takes it on day 1, when she knows next to nothing about the

players. "It seems, however, perfectly rational," Gärdenfors and Sahlin say, "if... Julie decides to bet on [the second day's match], but not on [the first]" (Gärdenfors and Sahlin 1982, 362).

The lesson of cases like this, Gärdenfors and Sahlin posit, is that "the amount and quality of information which the decision maker has concerning the possible states and outcomes of the decision situation in many cases is an important factor when making the decision" (Gärdenfors and Sahlin 1982, 362). Information about the amount or weight of one's evidence that an action/bet will produce good outcomes is not, however, reflected in that action/bet's expected utility. Julie's expectations remain constant as the weight of her evidence varies. *If* this is right, then locality is wrong. Whether or not Julie is committed to judging *Accept B* preferable to *Reject B* depends on more than just $Exp_p(Accept\ B)$ and $Exp_p(Reject\ B)$. It depends on whatever 'global' properties of her prior encode information about the weight of her evidence.

The Gärdenfors and Sahlin case provides *pro tanto* reason to expect locality to be false (though certainly not conclusive reason; for an orthodox Bayesian response, see Broome 1991, ch. 5). The Ellsberg paradox seems to provide similar reason. In the Ellsberg paradox (1961, pp. 653-5), a friend offers you two pairs of bets, $A/A^*$ and $B/B^*$, on a random draw from an urn containing 90 balls. You know that 30 balls are yellow, and that the other 60 are either red or black. But you have no prior information about the proportion of red to black. The bets pay out as follows:

| | *Yellow* | *Red* | *Black* |
|---|---|---|---|
| $A$ | \$100 | \$0 | \$0 |
| $A^*$ | \$0 | \$100 | \$0 |
| $B$ | \$100 | \$0 | \$100 |
| $B^*$ | \$0 | \$100 | \$100 |

Table 2.3: Ellsberg problem payoff table.

Many *prima facie* reasonable agents prefer $A$ to $A^*$ and $B^*$ to $B$ (Ellsberg 1961, 669). This pattern of preferences, however, violates Savage's Sure-Thing Principle. The Sure-Thing Principle says that when you evaluate two options, you ought to ignore alternatives in which they produce the same outcome. So you ought to prefer $A$ to $A^*$ if and only if you prefer $B$ to $B^*$. According to Bayesian orthodoxy, however, any precise prior commits you to preferences that satisfy the Sure-Thing Principle.

The lesson, Ellsberg imagines a respondent saying, is not that there is anything wrong with adopting a precise prior, or that there is anything wrong with the pattern of preferences. Rather, the lesson is this:

> ...having exploited knowledge, guess, rumor, assumption, advice, to arrive at a final judgment [precise prior probabilities for the events on which the utility of one's alternative actions depends]... one can still stand back from this process and ask: "How much, in the end, is all this worth? How much do I really know about the problem? How firm a basis for choice, for appropriate decision and action, do I have?" (Ellsberg 1961, 659-60).

And when the answers to these questions are, "It's not worth much," or "I don't know very much," or "I don't have a very firm basis for choice, for appropriate decision and action," rationality might not demand very much from you (Ellsberg 1961, 660). You may not be rationally committed to aligning your preferences with your best estimates of utility. Instead, you may be permitted to "search for additional grounds for choice," such as an action's 'security level' (its minimum expected utility relative to the priors not ruled out by your evidence; Ellsberg 1961, 662).

The Ellsberg case provides additional *pro tanto* reason to expect locality to be false (though still not conclusive reason). It seems to highlight the fact that information about the weight of one's evidence is relevant for determining whether or not you are

rationally committed to judging one action preferable to another. But, information about the weight of one's evidence that an action/bet will produce good outcomes is not reflected in that action/bet's expected utility. Again, *if* this is right, then locality is wrong. Rational commitments (to make comparative preferability judgments, and perhaps also other qualitative/comparative judgments, *e.g.*, comparative probability judgments) supervene on whatever 'global' properties of priors happen to encode information about the weight of one's evidence.

## 2.6 Rejecting Locality

### 2.6.1 The Main Argument

Bayesian orthodoxy about rational commitment is one example (the most plausible example) of a 'local' account of rational commitment. Local accounts say that for any type of comparative judgment $J_{X,Y}$ between $X$ and $Y$, there are certain marginal distributions, $f$ and $f'$, that encode the information relevant for determining whether $p$ commits its bearer to making the judgment $J_{X,Y}$. Moreover, that information is encoded locally, in the mean, or first moment of $f$ and $f'$, respectively. Local accounts of rational commitment, however, are implausible. So far, we have examined only *pro tanto* reason to think this. In a nutshell, the reason is: locality renders rational commitments insensitive to those features of priors (and posteriors) that encode information about weight (the higher moments of $f$ and $f'$). But, the Gärdenfors and Sahlin case, as well as the Ellsberg case, seem to suggest that this information is important for determining whether and which comparative/qualitative judgments you are committed to making.

The more definitive reason to reject the orthodox account of rational commitment (and any other local account) is this:

1. The orthodox account of rational commitment yields a particular inductive policy $\mathscr{I}$, or plan for making comparative and qualitative judgments in response to new data.

2. No plausible account of rational commitment $R$ yields an inductive policy $\mathscr{I}$ that is *strongly dominated* by another policy $\mathscr{I}^*$, in the sense that (i) for any prior $p$ and context $C$, $\mathscr{I}^*$'s expected epistemic utility in $C$, relative to $p$, is at least as great as $\mathscr{I}$'s, and (ii) for some prior $p'$ and context $C'$, $\mathscr{I}^*$'s expected epistemic utility in $C'$, relative to $p'$, is strictly greater than $\mathscr{I}$'s.

3. The orthodox policy $\mathscr{I}$ is strongly dominated by another policy $\mathscr{I}^*$. (In fact, any local policy is strongly dominated by $\mathscr{I}^*$.)

$C$ The orthodox account of rational commitment is implausible.

Section 2.6 defends premise 2. Sections 2.7-2.8 defend premise 3.

### 2.6.2 Expected Epistemic Utility of Inductive Policies

Every account of rational commitment $R$ corresponds to an inductive policy of the following form: an agent who adopts a prior $p$ in context $C$ and receives new data $D$ should make exactly the comparative and qualitative judgments that $p_D$ commits her to making in $C$, according to $R$ (no more, no less). For example, the orthodox account corresponds to the following policies for making comparative probability judgments and judgments of incremental support, respectively:

$$\mathscr{I}_{\preceq}(p, C, D, X, Y) = \begin{cases} X \succ Y \text{ if } p_D(X) > p_D(Y) \\ X \preceq Y \text{ otherwise} \end{cases}$$

- In words: if you adopt prior $p$ in context $C$ and receive new data $D$, judge that $X$ is more probable than $Y$ if $p_D(X)$ is greater than $p_D(Y)$; judge that $X$ is no more probable than $Y$ otherwise.

$$\mathscr{I}_{Confirmation}(p, C, D, H) = \begin{cases} D \textit{ incrementally confirms } H \text{ if } p_D(H) > p(H) \\ D \textit{ incrementally disconfirms } H \text{ if } p_D(H) < p(H) \\ D \textit{ is irrelevant to } H \text{ otherwise} \end{cases}$$

- In words: if you adopt prior $p$ in context $C$, judge that new data $D$ provides positive (negative) incremental support for $H$ if $p_D(H)$ is greater (less) than $p(H)$; judge that $D$ is irrelevant otherwise.

Inductive policies have *epistemic consequences*, and so are evaluable in terms of expected epistemic utility. Consider, for example, a policy $\mathscr{I}_{Accept}$ for accepting/rejecting theoretical hypotheses. Imagine that a doctor orders a test, in order to adjudicate between hypotheses about the disorder underlying a patient's symptoms (blindness in her left eye, perhaps). On the basis of her data, the doctor accepts hypothesis $H$ (that her patient has an autoimmune disorder) while rejecting $H'$ (that she has a viral infection), in accordance with $\mathscr{I}_{Accept}$. Typically, then, she will order certain kinds of follow-up tests. And these tests will put her in a better or worse position vis-á-vis securing *epistemically valuable* opinions regarding the patient's exact disorder (multiple sclerosis, lupus, etc.) and related issues (which treatment will be most effective). The epistemic value or utility of a state of opinion somehow summarizes all of its epistemically laudable qualities: accuracy, explanatory power, simplicity, and more. As Joyce (2009) notes, *accuracy* — which is a matter of how close the state of opinion is to the truth — is central to the notion of epistemic value. "Accuracy is the one epistemic value about which there can be no serious dispute: it

must be reflected in any plausible epistemic scoring rule," or epistemic utility function (Joyce 2009, 267).[26]

States of acceptance or rejection are plausibly evaluable *directly* in terms of epistemic utility (not simply indirectly, in terms of the downstream effects that they have on different types of opinions). Take, for example, a naïve, but illustrative version of a more sophisticated proposal from Briggs *et al.* (2013). An agent either accepts or rejects some hypotheses $H_1, ..., H_n$. We represent her acceptance/rejection state by a sequence $s$ of 0s and 1s ('0' for reject, '1' for accept). Then we can measure the inaccuracy of her acceptance/rejection state via the Hamming distance $d(s, s')$ between this sequence $s$ and the 'perfectly vindicated' sequence $s'$ (the sequence of 0s and 1s that give the truth-values of $H_1, ..., H_n$). This amounts to measuring inaccuracy by counting up the number of mistakes she makes (where she makes a mistake by either accepting a false hypothesis or rejecting a true hypothesis).[27] In contexts in which accuracy is paramount, then, $d(s, s')$ gives a rough measure of the epistemic utility of her acceptance/rejection state.

Given that acceptance/rejection states have epistemic utility scores relative to different worlds (and contexts, perhaps), different policies for accepting/rejecting theoretical hypotheses will have different *expected* epistemic utilities. If an agent adopts some prior $p$ over theoretical hypotheses $H_1, ..., H_n$ in some context $C$, the expected epistemic utility of a policy $\mathscr{I}$ for accepting/rejecting those hypotheses is:

$$\begin{aligned} Exp_p(eu(\mathscr{I})) &= \textstyle\sum_i \sum_j p(H_i \& D_j) \cdot eu(\mathscr{I}(p, D_j), H_i) \\ &= \textstyle\sum_i p(H_i) \sum_j p(D_j|H_i) \cdot eu(\mathscr{I}(p, D_j), H_i) \end{aligned}$$

where $\mathscr{I}(p, D_j)$ is the set of acceptance/rejection judgments that $\mathscr{I}$ advises our agent to make if she receives new data $D_j$ in context $C$, and $eu(\mathscr{I}(p, D_j), H_i)$ is the epistemic utility of making those judgments in $C$ given that $H_i$ is true.[28] Imagine,

for example, that a researcher adopts the uniform prior $p$ over competing theoretical hypotheses, $H$ and $H'$, so that $p(H) = p(H') = 1/2$. She performs an experiment to adjudicate between $H$ and $H'$, which can yield one of two data items, $D$ and $D'$. The objective probabilities for receiving any datum are given by: $p(D|H) = 0.7$, $p(D'|H) = 0.3$, $p(D|H') = 0.3$ and $p(D'|H') = 0.7$. Finally, the epistemic utility of accepting (rejecting) when $H$ is true (false) is as follows, let's suppose:

Table 2.4: Epistemic payoff of accepting/rejecting/abstaining.

| | *H true* | *H′ true* |
|---|---|---|
| *Accept H & Reject H′* | 1 | −5 |
| *Accept H′ & Reject H* | −5 | 1 |
| *Abstain from judgment* | −0.5 | −0.5 |

Then the expected epistemic utility of the most sensible 'total' policy $\mathscr{I}$, which says accept $H$/reject $H'$ if you receive $D$, and accept $H'$/reject $H$ if you receive $D'$, is $Exp_p(eu(\mathscr{I})) = (0.7) \cdot (1) + (0.3) \cdot (-5) = -0.8$. (So, it turns out, the most sensible total policy is not very sensible at all. Abstaining come what may has higher expected epistemic utility, relative to $p$.)

Other sorts of comparative and qualitative judgments have epistemic consequences as well. Consider comparative probability judgments, for example, which will be our focus from here on out. Suppose that a doctor orders a test, in order to adjudicate between hypotheses about a patient's disorder. On the basis of her data, she judges that $X$ is more probable than $Y$ (*e.g.*, that significant optic nerve demyelination/vision loss is more probable than minor demyelination/vision loss). Typically, this will affect how she structures subsequent inquiry. Perhaps she will not order the exact same suite of follow-up tests that she would if she outright *accepted* $X$ and *rejected* $Y$. Still, she will likely *focus* her inquiry by ordering more tests aimed at probing hypotheses that render $X$ probable, and fewer tests aimed at probing hypotheses that render $Y$

probable (*e.g.*, by ordering a lumbar puncture, which tests for autoimmune diseases, rather than blood work which tests for viral infections). This will put her in a better or worse position vis-á-vis arriving at valuable (accurate, etc.) opinions regarding the patient's exact disorder (multiple sclerosis, lupus, etc.) and related issues (which treatment will be most effective).

Comparative probability judgments are plausibly evaluable directly in terms of epistemic utility as well (not simply indirectly, in terms of their downstream epistemic effects). Here, for example, is a naïve version of a proposal from Fitelson (2013), which is similar in many respects to that of Briggs *et al.* (2013). An agent makes various comparative probability judgments between hypotheses $H_1, ..., H_n$. We represent her comparative probability ordering $\preceq$ by an adjacency matrix $m$; '1' at the $\langle H_i, H_j \rangle$ node of the matrix indicates that she judges $H_i \preceq H_j$ and '0' indicates that she does not. Then we can measure the inaccuracy of her comparative probability ordering via the Kemeny distance $d(m, m')$ (analog of Hamming distance for adjacency matrices) between this matrix $m$ and the 'perfectly vindicated' matrix $m'$, which has a '0' at the $\langle H_i, H_j \rangle$ node of the matrix if $H_i$ is true and $H_j$ false, and a '1' otherwise. This amounts to measuring inaccuracy by counting up the number of mistakes she makes (where she makes a mistake if she judges $H_i \preceq H_j$ with $H_i$ true/$H_j$ false, or fails to judge $H_i \preceq H_j$ in any other case). In contexts in which accuracy is paramount, then, $d(m, m')$ gives a rough measure of the epistemic utility of her comparative probability ordering. (In chapter 3, I discuss a way of measuring the epistemic value of imprecise credal states, which provides an alternative approach to measuring the epistemic value of comparative probability orderings.)

| | $H_1$ | ... | $H_n$ |
|---|---|---|---|
| $H_1$ | 1 | ... | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $H_n$ | 1 | ... | 1 |

Figure 2.8: Adjacency matrix representing $\preceq$.

Given that comparative probability orderings have epistemic utility scores relative to different worlds (and contexts), different policies for making comparative probability judgments will have different *expected* epistemic utilities. Suppose, for example, that we make a simplifying assumption of the following sort: the epistemic utility of judging $X$ more probable than $Y$ is given by:

Table 2.5: Epistemic payoff of judging $X \preceq Y$, or abstaining.

| | $X\&Y$ | $X\&\neg Y$ | $\neg X\&Y$ | $\neg X\&\neg Y$ |
|---|---|---|---|---|
| $X \preceq Y$ | $eu_1$ | $eu_2$ | $eu_3$ | $eu_4$ |
| *Abstain from judgment* | $eu_5$ | $eu_5$ | $eu_5$ | $eu_5$ |

or perhaps:

Table 2.6: Epistemic payoff of judging $X \preceq Y$, or abstaining.

| | $ch(X) \leq ch(Y)$ | $ch(X) > ch(Y)$ |
|---|---|---|
| $X \preceq Y$ | $eu_1^*$ | $eu_3^*$ |
| *Abstain from judgment* | $eu_2^*$ | $eu_2^*$ |

with $eu_1^* > eu_2^* > eu_3^*$. It does not matter much which simplifying assumption we make. The results that we obtain in §2.7-2.8 are fairly robust. But it will be helpful to have some numbers to work with. I will opt for the latter, simpler assumption.

In contexts in which the epistemic utilities are as described in table 2.6, the orthodox inductive policy $\mathscr{I}$ for making comparative probability judgments, namely:

$$\mathscr{I}(p, D) = \begin{cases} X \succ Y \text{ if } p_D(X) > p_D(Y) \\ X \preceq Y \text{ otherwise} \end{cases}$$

will typically have lower expected epistemic utility than various other policies $\mathscr{I}^*$, from the perspective of a range of different priors (*mutatis mutandis* for any other 'local' policy).

To illustrate, suppose that a bookie hands you a coin and offers you a bet. You have no prior evidence about the coin's bias. But the bookie allows you to flip the coin

for awhile — 25 times, for example — prior to deciding whether or not to take the bet. To take account of your prior information (*viz.*, none) in your decision problem, you decide to adopt the maximum entropy prior $u$ over hypotheses $B = x$ about the coin's bias. Then the expected epistemic utility of $\mathscr{I}$ from the perspective of $u$ is:

$$Exp_u(eu(\mathscr{I})) = \int_0^1 \sum_{k=0}^{25} \begin{pmatrix} 25 \\ k \end{pmatrix} \cdot x^k \cdot (1-x)^{25-k} \cdot eu(\mathscr{I}(u, H^k T^{25-k}), B = x)\, dx$$

where $D = H^k T^{25-k}$ is any sequence of $k$ heads and 25-$k$ tails.[29] For concreteness, suppose that the epistemic utilities are given by the following table (see the appendix for discussion of this particular assignment of epistemic utilities):

Table 2.7: Epistemic payoff of judging *Heads* $\preceq$ *Tails*, *Heads* $\succ$ *Tails*, or abstaining.

| | *ch*(*Heads*) $\leq$ *ch*(*Tails*) | *ch*(*Heads*) $>$ *ch*(*Tails*) |
|---|---|---|
| *Heads* $\preceq$ *Tails* | 1 | $-5$ |
| *Heads* $\succ$ *Tails* | $-5$ | 1 |
| *Abstain from judgment* | $-0.5$ | $-0.5$ |

Then the expected epistemic utility of $\mathscr{I}$ is $Exp_u(eu(\mathscr{I})) = 0.535057$. Now note that $\mathscr{I}$ prescribes judging that heads on the next toss is more probable than tails whenever $k \geq 13$ (whenever you flip more heads than tails on your first 25 tosses), and prescribes judging the opposite — that tails is more probable than heads — whenever $k < 13$. Compare $\mathscr{I}$ with the policy $\mathscr{I}^*$ that prescribes (i) judging that heads is more probable than tails if $k \geq 15$, (ii) abstaining from judgment if $11 \leq k \leq 14$, and (iii) judging that tails is more probable than heads if $k \leq 10$. One might expect $\mathscr{I}^*$ to have a higher expected epistemic utility than $\mathscr{I}$, since it directs you to 'hedge your epistemic bets' by abstaining from judgment when there's significant risk (from $u$'s perspective) of making a judgment with deleterious epistemic consequences. And indeed $Exp_u(eu(\mathscr{I}^*)) = 0.627876$.

This is the first step in seeing that the orthodox policy $\mathscr{I}$ is *strongly dominated* by another policy. If some other policy both (i) weakly dominates $\mathscr{I}$ (has at least as great an expected epistemic utility in every context, relative to every prior), and (ii) yields the same verdicts in cases like this, then that policy strongly dominates $\mathscr{I}$. In the next section, I identify exactly such a policy.

We ought to pause here, before proceeding, to say why accounts of rational commitment that yield strongly dominated inductive policies are implausible. The reason is: it ought to be possible to (i) make every comparative and qualitative judgment that you are rationally committed to making (in virtue of adopting a prior $p$ and receiving new data $D$ in context $C$), (ii) abstain from judgment when you are not so committed, and (iii) not contravene your best estimates (expectations), *i.e.*, not make comparative and qualitative judgments that are, according to $p$'s *best estimates* (expectations), epistemically inferior to some other set of judgments that you might have made. Whatever else is true about what you are *positively* committed to (in terms of making comparative/qualitative judgments), in virtue of adopting $p$, it is *open* to you (rationally permissible), one would think, to make exactly the judgments that you are committed to making (whatever they happen to be) without flouting your own best estimates. This is a bedrock fact about epistemic rationality, I posit. Any account of rational commitment that says otherwise — *e.g.*, any account that yields a strongly dominated inductive policy — is mistaken.[30]

## 2.7 Proceeding Without Locality

### 2.7.1 Globalism

Let's take stock. Imprecise Bayesians say that less than maximally specific evidence at least permits (and perhaps positively requires) less than maximally specific (merely

partial) opinions, or comparative/qualitative judgments. But adopting a precise prior commits you to a *total* comparative probability ordering, *total* preference ordering, etc. So when your evidence is unspecific, you should *not* adopt a precise prior. You should adopt an imprecise prior instead. Imprecise priors, unlike precise priors, permit merely partial comparative and qualitative judgments.

We will see, however, that adopting a precise prior only *seems* to commit you to maximally specific opinions, or comparative/qualitative judgments, because of Bayesian orthodoxy about rational commitment, *viz.*, that an agent who adopts a precise prior $p$ is committed to judging:

- ...that $X$ is more probable than $Y$ if $p(X) > p(Y)$.
- ...that data $D$ provides positive incremental support for $H$ if $p_D(H) > p(H)$.
- ...that action $A$ is preferable to $B$ if the expected utility of $A$ (relative to $p$) is greater than the expected utility of $B$ (relative to $p$).

This bit of orthodoxy is implausible. I examined some *pro tanto* reason to think this in §2.3-2.5. The more definitive reason, though, to reject the orthodox account of rational commitment is that it encodes an inductive policy $\mathscr{I}$ that is strongly dominated by another policy $\mathscr{I}^*$.

The final example in §2.6 shows that $\mathscr{I}$ has suboptimal expected epistemic utility in certain contexts (those in which table 2.7 describes the relevant epistemic utilities), relative to certain priors (the MaxEnt prior). But it does not show that there is a plausible, non-orthodox account of rational commitment which yields a better inductive policy $\mathscr{I}^*$, in the following sense: (i) $\mathscr{I}^*$ does at least as well (in terms of expected epistemic utility) as $\mathscr{I}$ in every context, relative to every prior, and (ii) $\mathscr{I}^*$ does strictly better in some contexts, relative to some priors. It does not show that $\mathscr{I}$ is strongly dominated.

My aim now is to identify such an account. I plan to construct a theory of rational commitment that takes seriously the moral from §2.5: even when you adopt a precise prior, the weight of your evidence is relevant for determining whether and which comparative/qualitative judgments you are rationally committed to making. I will then show that this account yields an inductive policy $\mathscr{I}^*$ that strongly dominates the orthodox policy $\mathscr{I}$.

To see how we might construct such a theory, recall that on the orthodox account, whether $p$ carries a commitment to judging $X \succ Y$ depends exclusively on $p(X)$ and $p(Y)$. These probabilities, however, are merely the first moments of certain marginal distributions (they are expected values, *viz.*, the expected objective probabilities of $X$ and $Y$, respectively); $p(X) = \sum_i p(X|H_i) \cdot p(H_i) = \sum_x x \cdot p(Q_X = x)$, where $Q_X = x$ if and only if some theoretical hypothesis $H_i$ with $p(X|H_i) = x$ is true. When considering uncountably many theoretical hypotheses, $p(X) = \int_0^1 x \cdot f_{Q_X}(x)\,dx$, where $f_{Q_X}$ is the density that defines the marginal distribution of $Q_X$.

The characteristic effect of weight is to cause probabilities to concentrate more and more heavily on increasingly smaller, typically connected subsets of theoretical hypotheses. Weighty evidence for $X$ causes $f_{Q_X}$ to become increasingly 'peaked'; likewise, weighty evidence for $Y$ causes $f_{Q_Y}$ to become increasingly 'peaked'. As a result, relative to any reasonable distance function $d$ on the space of probability densities, the distance between $f_{Q_X}$ and $f_{Q_Y}$, $d(f_{Q_X}, f_{Q_Y})$, approaches the distance between $p(X)$ and $p(Y)$. Accordingly, $d(f_{Q_X}, f_{Q_Y})/\ |p(X) - p(Y)|$ approaches 1.

One way to render rational commitments sensitive to those features of priors (and posteriors) that encode information about weight is to tie them quantities like $d(f_{Q_X}, f_{Q_Y})/|p(X) - p(Y)|$. This quantity is determined by all of the higher moments of $f_{Q_X}$ and $f_{Q_Y}$, which is where information about weight lives. This is the option we will explore here. In particular, the proposal is this: an agent who adopts a precise

prior $p$ is committed to to judging that $X$ is more plausible than $Y$ if (i) $p(X) > p(Y)$ and (ii) $1-\epsilon < d(f_{Q_X}, f_{Q_Y})/|p(X)-p(Y)| < 1+\epsilon$, for some contextually determined threshold $\epsilon > 0$ (for discussion of $\epsilon$, see §2.8).[31] (Or a bit more generally, an agent who adopts a precise prior $p$ is committed to to judging that $X$ is more plausible than $Y$ if (i) $p(X) > p(Y)$ and (ii) $\langle d(f_{Q_X}, f_{Q_Y}), |p(X)-p(Y)|\rangle$ satisfies a contextually determined constraint $\mathscr{C}$, where $\mathscr{C}$ is either of the form $1-\epsilon < d(f_{Q_X}, f_{Q_Y})/|p(X)-p(Y)| < 1+\epsilon$, or is the limit of a series of such constraints, *e.g.*, the trivial constraint, satisfied by all pairs $\langle x, y\rangle$. This generalization will be important later.) Call this *the globalist thesis*.

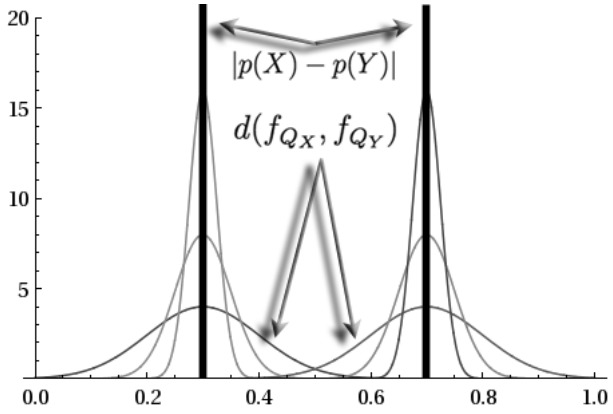


Figure 2.9: A range of distributions which concentrate probability more and more heavily on smaller and smaller subsets of theoretical hypotheses.

Like Bayesian orthodoxy, this view holds that for any type of comparative judgment $J_{X,Y}$ between $X$ and $Y$, there are certain marginal distributions, $f$ and $f'$, that encode the information relevant for determining whether $p$ commits its bearer to making the judgment $J_{X,Y}$. But unlike Bayesian orthodoxy, this view holds that this information is encoded *globally*, across all of the moments of $f$ and $f'$ (mean, variance, skewness, etc.).

### 2.7.2 An Illustration: Resolving the Preclusion Problem

Imagine that you are at a horse race. A bookie offers you one of three bets. You can either bet that Goldencents will beat a certain time $T$, or that Itsmyluckday

will beat time $T$, or that Pataky Kid will beat time $T$. At the outset, you have no relevant information about the three horses. To take account of your prior information (*viz.*, none) in your decision problem, you decide to adopt the maximum entropy prior $u$ over hypotheses $ch(\textit{Horse } X \textit{ beats time } T) = x$ about the chance that Goldencents/Itsmyluckyday/Pataky kid beats time $T$.

You then ask a friend to pull some strings. She gets you into some practice sessions. You attend $N$ independent runs for each horse, and observe that Goldencents ($GC$) beats time $T$ a total of $a$ times, Itsmyluckyday ($IMLD$) beats it $b$ times, and Pataky kid ($PK$) beats it $c$ times. Call this new data '$D$'. Next week, the bookie offers you the three bets again. To incorporate your new data in your decision problem, you condition $u$ on $D$. The result:

Table 2.8: Posterior probabilities that GC/IMLD/PK beats time $T$.

| | | $u_D(GC \textit{ beats } T)$ | $u_D(IMLD \textit{ beats } T)$ | $u_D(PK \textit{ beats } T)$ |
|---|---|---|---|---|
| *Case 1*: | $N = 5$, $a = 4$, $b = 3$, $c = 1$ | 0.714 | 0.571 | 0.286 |
| *Case 2*: | $N = 47$, $a = 34$, $b = 27$, $c = 13$ | 0.714 | 0.571 | 0.286 |

According to Bayesian orthodoxy, you are committed to the same total comparative probability ordering $\preceq$ in either case. You are committed to judging (i) that Goldencents is more likely to beat time $T$ than Itsmyluckyday ($IMLD \prec GC$), (ii) that Goldencents is more likely to beat $T$ than Pataky kid ($PK \prec GC$), (iii) that Itsmyluckyday is more likely to beat $T$ than Pataky kid ($PK \prec IMLD$), and so on.

| | | |
|---|---|---|
| $PK \prec IMLD$ | $IMLD \prec GC$ | $PK \prec GC$ |
| $\neg GC \prec GC$ | $\neg IMLD \prec IMLD$ | $PK \prec \neg PK$ |

According to the globalist thesis, however, $u_D$ only commits you to making a comparative probability judgment between $X$ and $Y$ when it reflects weighty enough

evidence for $X$ and $Y$ to make it the case that $1 - \epsilon < d(f_{Q_X}, f_{Q_Y})/|p(X) - p(Y)| < 1 + \epsilon$. (Again, for more on $\epsilon$, see §2.8.) When it does not reflect sufficiently weighty evidence, $u_D$ simply does not commit you to making a comparative probability judgments between $X$ and $Y$ one way or the other. It permits you to not take a stand on the matter.

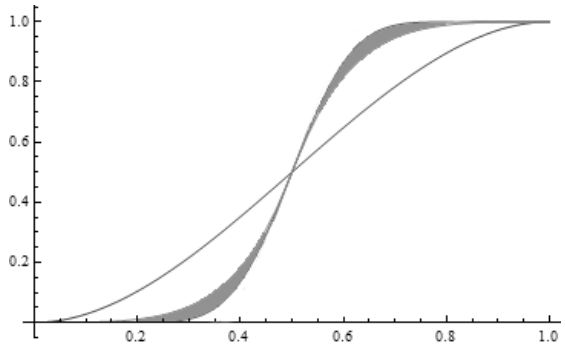


Figure 2.10: Cramer-von Mises distance.

Deza and Deza (2009) survey a wide range of distance functions on the space of probability densities. For concreteness, I focus on one in particular. I let $d(f_{Q_X}, f_{Q_Y})$ be the *Cramer-von Mises distance* between $f_{Q_X}$ and $f_{Q_Y}$, which we denote $\mathfrak{C}(f_{Q_X}, f_{Q_Y})$.

$$\mathfrak{C}(f, g) = \int_0^1 |F(x) - G(x)|^2 dx$$

$\mathfrak{C}$ specifies the distance between densities $f$ and $g$ as a function of the area between their corresponding cumulative distribution functions, $F$ and $G$ (counting regions of smaller divergence for less and regions of greater divergence for more; pictured left).[32] (It is the squared $L_2$ metric between $F$ and $G$.) It is attractive because (i) it is an analogue of squared Euclidean distance on the space of probability densities, and (ii) it yields the correct verdict about comparative closeness in those cases where obviously correct answers are to be had.[33]

In cases 1 and 2, you arrive at the same posterior truth-value estimates for the various propositions of interest: that Goldencents/Itsmyluckyday/Pataky kid will beat time $T$. But you have much weightier evidence undergirding your estimates in case 2. You have a firmer basis for making comparative and quali-

tative judgments, both doxastic and practical. This is reflected in the value of the quantity $d(f_{X|D}, f_{Y|D})/|u_D(X) - u_D(Y)|$ (where $f_{X|D}$ is the marginal density of $u_D$ over hypotheses $ch(\textit{Horse } X \textit{ beats time } T) = x$ about the chance that Goldencents/Itsmyluckyday/Pataky beats time $T$). For illustrative purposes, let $\epsilon = 0.52$. So $u_D$ only commits you to judging that horse $X$ is likelier to beat time $T$ than horse $Y$ (or vice versa) when it reflects weighty enough evidence for the propositions *Horse X will beat T* and *Horse Y will beat T*, respectively, to make it the case that $d(f_{X|D}, f_{Y|D})/|u_D(X) - u_D(Y)| \in [0.48, 1.52]$. Now compare:

Table 2.9: The effect of weight on $d(f_{X|D}, f_{Y|D})/|u_D(X) - u_D(Y)|$.

| | | $\frac{d(f_{GC\|D}, f_{IMLD\|D})}{\|u_D(GC) - u_D(IMLD)\|}$ | $\frac{d(f_{GC\|D}, f_{PK\|D})}{\|u_D(GC) - u_D(PK)\|}$ | $\frac{d(f_{IMLD\|D}, f_{PK\|D})}{\|u_D(IMLD) - u_D(PK)\|}$ |
|---|---|---|---|---|
| *Case 1*: | $N = 5$, $a = 4$, $b = 3$, $c = 1$ | $= 0.245 \notin [.48, 1.52]$ | $= 0.6 \in [.48, 1.52]$ | $= 0.443 \notin [.48, 1.52]$ |
| *Case 2*: | $N = 47$, $a = 34$, $b = 27$, $c = 13$ | $= 0.511 \in [.48, 1.52]$ | $= 0.832 \in [.48, 1.52]$ | $= 0.736 \in [.48, 1.52]$ |

In case 1, your posterior $u_D$ reflects insufficiently weighty evidence to commit you to a total comparative probability ordering. It commits you to a merely partial order. It permits you to not take a stand on some matters, to not make a judgment about the comparative probability of $X$ and $Y$, for certain $X$ and $Y$. It *does* commit you to making the following judgments:

$$\boxed{PK \prec GC,\ \neg GC \prec GC,\ PK \prec \neg PK}$$

But it does *not* commit you to making other judgments: that Goldcents is definitely likelier to beat time $T$ than Itsmyluckyday (or vice versa); that Itsmyluckyday is definitely likelier to beat $T$ than Pataky kid (or vice versa); that Itsmyluckyday is likelier than not to beat $T$ (or vice versa).


In case 2, however, your posterior $u_D$ reflects sufficiently weighty evidence to commit you to a total comparative probability ordering. In case 2, you are committed to making exactly the same judgments, on both the globalist and localist (orthodox) accounts.

This highlights how flexible precise priors are. If the globalist account of rational commitment is right, then precise priors do indeed allow for partial comparative probability orderings (as well as partial preference orderings, etc.). In that case, we ought to rethink the central epistemic motivation for imprecise Bayesianism. Imprecise Bayesians say, "You ought to adopt imprecise priors, in certain circumstances — in particular, when your prior evidence is unspecific or equivocal — because they allow for unspecific opinions; they allow for partial comparative and qualitative judgments. Precise priors do not." But precise priors *do* allow for unspecific and equivocal opinions in a wide range of evidential circumstances, on the globalist account; they *do* allow for partial comparative and qualitative judgments.

## 2.8 A Rationale for Globalism

The globalist inductive policy strongly dominates the orthodox policy (and any other 'local' policy), I hope to show. This provides good epistemic reason to reject Bayesian orthodoxy about rational commitment (and any other 'local' account).

To illustrate, imagine one more time that a bookie hands you a coin. She offers you a bet. You have no prior evidence about the coin's bias. The bookie allows you to flip the coin 25 times before deciding whether or not to take the bet. You adopt the maximum entropy prior $u$ over hypotheses $B = x$ about the coin's bias. The

epistemic utilities are given by:

Table 2.10: Epistemic payoff of judging *Heads* $\preceq$ *Tails*, *Heads* $\succ$ *Tails*, or abstaining.

| | $ch(Heads) \leq ch(Tails)$ | $ch(Heads) > ch(Tails)$ |
|---|---|---|
| *Heads* $\preceq$ *Tails* | 1 | $-5$ |
| *Heads* $\succ$ *Tails* | $-5$ | 1 |
| *Abstain from judgment* | $-0.5$ | $-0.5$ |

The globalist inductive policy is built out of policies $\mathscr{I}_x$ of the form: if you adopt a prior $p$ and receive new data $D$, make exactly the comparative and qualitative judgments that $p_D$ commits you to making, given $\epsilon = x$ (according to globalism). In particular, $\mathscr{I}_x$ says: judge that $X$ is more plausible than $Y$, iff (i) $p_D(X) > p_D(Y)$ and (ii) $1 - x < d(f_{V_X|D}, f_{V_Y|D})/|p_D(X) - p_D(Y)| < 1 + x$. (Recall, $V_X = x$ if and only if some theoretical hypothesis $H_i$ with $p_D(X|H_i) = x$ is true; $f_{V_X|D}$ is the density that defines the marginal distribution of $V_X$ conditional on $D$.) In our coin flipping case, then, the expected epistemic utility of $\mathscr{I}_x$ is:

Table 2.11: Expected epistemic utility of $\mathscr{I}_x$ from the perspective of $u$.

| | $x = 0$ | $x = 0.2$ | $x = 0.4$ | $x = 0.6$ | $x = 0.8$ | $x = 1$ |
|---|---|---|---|---|---|---|
| $Exp_u(eu(\mathscr{I}_x))$ | $-0.5$ | 0.305 | 0.588 | 0.628 | 0.615 | 0.535 |

The globalist inductive policy $\mathscr{I}^*$ says: if you adopt prior $p$ in context of inquiry $C$, and receive new data $D$, make the comparative/qualitative judgments that $\mathscr{I}_x$ prescribes making, for whichever $x$ maximizes expected epistemic utility $Exp_u(eu(\mathscr{I}_x))$ in $C$. In the case at hand (coin flipping, epistemic utilities as per above), $Exp_u(eu(\mathscr{I}_x))$

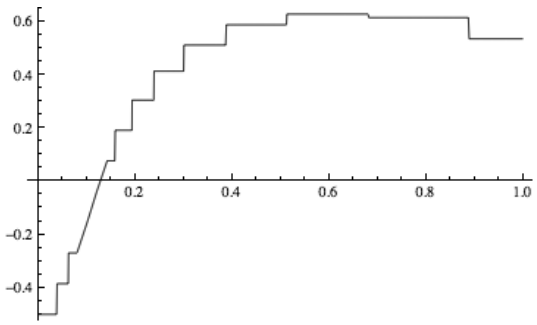


Figure 2.11: $Exp_u(eu(\mathscr{I}_x))$.

takes a maximum at $x = 0.52$ (right).[34] So the globalist inductive policy $\mathscr{I}^*$ is just $\mathscr{I}_{0.52}$ in this context. And $\mathscr{I}_{0.52}$ recommends the following:

$$\mathscr{I}_{0.52}(u, D, \textit{Heads}, \textit{Tails}) = \begin{cases} \textit{Heads} \succ \textit{Tails if } D = H^k T^{25-k} \textit{ for some } k > 14 \\ \textit{Heads} \prec \textit{Tails if } D = H^k T^{25-k} \textit{ for some } k < 11 \\ \textit{Abstain from judgment otherwise} \end{cases}$$

By construction, the globalist inductive policy $\mathscr{I}^*$ has at least as great an expected epistemic utility (from $u$'s perspective) as any policy $\mathscr{I}_x$. But there are, of course, various policies not of the form $\mathscr{I}_x$. An inductive policy is just a function:

$$\mathscr{I}(p, D, X, Y) = \begin{cases} X \preceq Y \textit{ if condition } C_1 \textit{ obtains} \\ X \succ Y \textit{ if condition } C_2 \textit{ obtains} \\ \textit{Abstain from judgment otherwise} \end{cases}$$

The policy that directs you to judge that heads is more probable than tails come what may, for example, is not equivalent to any $\mathscr{I}_x$. (No $\mathscr{I}_x$ advises an agent who adopts the uniform prior $u$ to judge that heads is more probable than tails in response to $D = H^0T^{25}$.)

It would be nice if $\mathscr{I}^*$ maximized expected epistemic utility in any context of inquiry, relative to *all* other inductive policies (not just policies of the form $\mathscr{I}_x$). Fortunately, it is easy to check that this is so, at least in simple enough contexts.[35] For example, when the epistemic utilities are given by our usual table:

Table 2.12: Old 'conservative' payoff matrix.

| | $ch(\textit{Heads}) \leq ch(\textit{Tails})$ | $ch(\textit{Heads}) > ch(\textit{Tails})$ |
|---|---|---|
| $\textit{Heads} \preceq \textit{Tails}$ | 1 | $-5$ |
| $\textit{Heads} \succ \textit{Tails}$ | $-5$ | 1 |
| *Abstain from judgment* | $-0.5$ | $-0.5$ |

and $n = 8$ (you flip the coin a total of 8 times), then $Exp_u(eu(\mathscr{I}^*)) = 0.349609$, which standard optimization techniques show to be a global maximum (maximum relative

to *all* inductive policies). Similarly, if $n = 10$, then $Exp_u(eu(\mathscr{I}^*)) = 0.414062$, which again is a global maximum.

If we consider a different context, in which the epistemic utilities are given by:

Table 2.13: New 'liberal' payoff matrix.

| | $ch(Heads) \leq ch(Tails)$ | $ch(Heads) > ch(Tails)$ |
|---|---|---|
| $Heads \leq Tails$ | 5 | $-1$ |
| $Heads > Tails$ | $-1$ | 5 |
| *Abstain from judgment* | $-0.5$ | $-0.5$ |

then $Exp_u(eu(\mathscr{I}_x))$ takes a maximum at $x = 1$. So the globalist inductive policy $\mathscr{I}^*$ is just $\mathscr{I}_1$ in this context. And $\mathscr{I}_1$ recommends the following:

$$\mathscr{I}_1(u, D, Heads, Tails) = \begin{cases} Heads \succeq Tails \text{ if } D = H^k T^{n-k} \text{ for some } k \geq n/2 \\ Heads \prec Tails \text{ if } D = H^k T^{n-k} \text{ for some } k < n/2 \end{cases}$$

So, in this context, the globalist policy and local orthodox policy make the same recommendations. Once more, it is easy to check that $\mathscr{I}^*$ has at least as great an expected epistemic utility (from $u$'s perspective) as *any* other inductive policy. If $n = 8$, then $Exp_u(eu(\mathscr{I}^*)) = 4.17969$. Standard optimization techniques show that this is a global maximum. No other inductive policy has a higher expected epistemic utility in this context. Similarly, if $n = 10$, then $Exp_u(eu(\mathscr{I}^*)) = 0.426172$, which again is a global maximum.

These examples by no means show that globalism is *the* uniquely plausible account of rational commitment. They do not show that there is no other principled, non-orthodox account which yields a better inductive policy, a policy $\mathscr{I}^{**}$ that strongly dominates $\mathscr{I}^*$. For all I have said, there could be another policy $\mathscr{I}^{**}$ that agrees with $\mathscr{I}^*$ in the two contexts considered here, but does strictly better in various other contexts.

My aim, however, is not to show that globalism is the uniquely plausible account of rational commitment. Rather, my aim is to identify good epistemic reason to reject the orthodox (localist) account of rational commitment, and to leverage this to resolve the preclusion problem.

These illustrative examples suffice for this purpose. The reason: by construction, $\mathscr{I}^*$ weakly dominates $\mathscr{I}$. Proof: the orthodox policy $\mathscr{I}$ is just the degenerate policy $\mathscr{I}_{trivial}$. That is, $\mathscr{I}$ says: judge that $X$ is more plausible than $Y$ if and only if (i) $p_D(X) > p_D(Y)$ and (ii) $\langle d(f_{V_X|D}, f_{V_Y|D}), |p_D(X) - p_D(Y)| \rangle$ satisfies the trivial constraint, *i.e.*, the constraint satisfied by all pairs $\langle x, y \rangle$. And in any context $C$, $\mathscr{I}^*$ is just $\mathscr{I}_x$, for whatever $\mathscr{I}_x$ is such that $Exp_u(eu(\cdot))$ takes a maximum at $\mathscr{I}_x$ in $C$. So, in any context in which $\mathscr{I}$ maximizes expected epistemic utility, $\mathscr{I}^*$ just is $\mathscr{I}_{trivial}$, which just is $\mathscr{I}$. By construction, then, there is no context in which $\mathscr{I}$ has higher expected epistemic utility (from $u$'s perspective) than $\mathscr{I}^*$. In addition, our illustrative examples show that there are contexts in which $\mathscr{I}^*$ has strictly higher expected epistemic utility (from $u$'s perspective) than $\mathscr{I}$. Hence, $\mathscr{I}^*$ strongly dominates $\mathscr{I}$.

This provides good epistemic reason to reject Bayesian orthodoxy about rational commitment (and any other 'local' account). Moreover, the contexts $C$ described above are ones in which *no other inductive policy* has greater expected epistemic utility (from $u$'s perspective) than $\mathscr{I}^*$. So the true account of rational commitment $R$, then — whatever it turns out to be — will agree with the globalist account in $C$, and hence make rational commitments sensitive to those features of priors (and posteriors) that encode information about weight (higher moments) in $C$ (though perhaps not in exactly the way the globalist account proposes). Plausibly, then, $R$ will make rational commitments sensitive to those features more generally (unless the correct account of rational commitment is a gerrymandered mess).

If this is right, then the big take-home lesson is this: there is plausibly no preclusion

problem for precise Bayesianism. Adopting a precise prior does *not* commit you to having perfectly precise opinions. It does not invariably commit you to making comparative and qualitative judgments that form total orders. The true account of rational commitment — whatever it turns out to be — will say: a wide range of precise priors carry commitments to merely partial comparative probability orderings, preference orderings, etc., in a wide range of contexts.

Of course, this does *not* mean that there are no good epistemic reasons to employ imprecise priors. It just means that we should not look to the preclusion problem to furnish those reasons. In the final chapter of my dissertation, I search for new epistemic reasons to employ imprecise priors, for a new epistemic foundation for imprecise Bayesianism.

## 2.9 Conclusion

I have argued that adopting a precise prior does not invariably commit you to a total comparative probability ordering, preference ordering, etc. A wide range of precise priors carry commitments to merely partial comparative probability orderings, etc., in a wide range of contexts. The proper motivation for introducing imprecise priors, then, is *not* they are required in order to avoid overly specific posterior states of opinion. To recap, the main argument goes as follows:

1. The globalist account of rational commitment, or something similar, is plausibly true. The orthodox account is implausible.

2. If the globalist account of rational commitment, or something similar, is true, then a wide range of precise priors carry commitments to *merely partial* comparative probability orderings (and preference orderings, etc.).

3. If a wide range of precise priors carry commitments to merely partial orderings, then they do not invariably capture improper responses to unspecific evidence.

C. Plausibly, then, precise priors do not invariably capture improper responses to unspecific evidence.

The defense of premise 1 comes in two parts. The first part goes as follows:

1′ Every account of rational commitment $R$ corresponds to an inductive policy of the form: an agent who adopts a prior $p$ in context $C$ and receives new data $D$ should make exactly the comparative and qualitative judgments that $p_D$ commits her to making in $C$, according to $R$.

2′ No plausible account of rational commitment $R$ yields an inductive policy $\mathscr{I}$ that is *strongly dominated* by another policy $\mathscr{I}^*$, in the sense that (i) for any prior $p$ and context $C$, $\mathscr{I}^*$'s expected epistemic utility in $C$, relative to $p$, is at least as great as $\mathscr{I}$'s, and (ii) for some prior $p'$ and context $C'$, $\mathscr{I}^*$'s expected epistemic utility in $C'$, relative to $p'$, is strictly greater than $\mathscr{I}$'s.

3′ The orthodox account of rational commitment yields an inductive policy $\mathscr{I}$ that is strongly dominated by the globalist policy $\mathscr{I}^*$.

$C'$ The orthodox account of rational commitment is implausible.

The second part:

1″ For some priors $p$ and contexts $C$, the globalist policy's $\mathscr{I}^*$'s expected epistemic utility in $C$, relative to $p$, is at least as great as any other policy's expected epistemic utility.

$2''$ The true account of rational commitment $R$, then, will agree with the globalist account in contexts $C$, and hence make rational commitments sensitive to those features of priors (and posteriors) that encode information about weight (higher moments) in $C$ (though perhaps not in exactly the way the globalist account proposes).

$3''$ Plausibly, then, $R$ will make rational commitments sensitive to those features more generally (unless the correct account of rational commitment is a gerrymandered mess).

$C''$ So the globalist account of rational commitment, or something similar, is plausibly true.

If correct, this seems to undercut the central epistemic motivation for imprecise Bayesianism. At a minimum, it provides impetus to search for new epistemic foundations for imprecise Bayesianism. I conclude by raising a few additional questions to be addressed in future research.

- We specified the globalist thesis using one particular distance function on the space of probability densities, *viz.*, Cramer-von Mises distance. Are our results robust across a range of metrics, *e.g.*, the Lévy metric? the $L_p$ metrics?

- We suggested that we might be able to measure the inaccuracy of an agent's comparative probability ordering $\preceq$ over hypotheses $H_1, ..., H_n$ at a world $w$ by representing $\preceq$ as an adjacency matrix $m$, and taking the Kemeny distance $d(m, m')$ between this matrix $m$ and the 'perfectly vindicated' matrix $m'$ at $w$. In contexts in which accuracy is paramount, then, $d(m, m')$ might provide a rough measure of the epistemic utility of $\preceq$. But is this really the right way to think about the epistemic utility of a comparative probability ordering at a

world? What are the alternatives? What exactly are the relevant desiderata for deciding between alternatives? I address some of these questions in the final chapter of my dissertation.

- The globalist account ties rational commitments vis-á-vis comparative probability to a particular quantity, $d(f_{Q_X}, f_{Q_Y})/|p(X) - p(Y)|$, which reflects information about the weight of the evidence that the prior $p$ summarizes. What other quantities might an alternative account tie such commitments to? What reasons are there to prefer the globalist quantity to these other quantities, or vice versa?

- How far can standard optimization techniques take us toward proving that the globalist inductive policy $\mathscr{I}^*$ is non-undermining, in the following sense: for any prior $p$ and context $C$, $\mathscr{I}^*$'s expected epistemic utility in $C$, relative to $p$, is at least as great as that of any other inductive policy?

# Notes

[18]Bayesians, of course, agree about much more than this. They agree, for example, that any reasonable distributions $p$ and $q$ agree on the likelihoods or direct inference probabilities that $H_1, ..., H_n$ specify for the potential experimental data sequences $D$ ($p(D|H_i) = q(D|H_i)$, though $p$ and $q$ might disagree on the unconditional probability of various $H_i$ ($p(H_i) \neq q(H_i)$; *cf.* Hawthorne 1994). They also disagree certain issues, *e.g.*, which quantitative confirmation judgments an agent who adopts a prior $p$ is committed to making.

[19]If, for example, an agent's comparative probability judgments between competing theoretical hypotheses satisfy Scott's axiom (read '$X \leq Y$' as hypothesis $Y$ is at least as plausible as hypothesis $X$), and are also rich enough to satisfy some additional structural axioms (completeness: $X \leq Y$ or $X > Y$ for all $X$ and $Y$; non-atomicity: for any $X$ such that $X > X\&\neg X$, there is a $Y$ such that $X\&Y > X\&\neg X$ and $X\&\neg Y > X\&\neg X$), then those comparative judgments 'pin down' a precise

prior, in the following sense: there is a unique probability distribution $p$ that represents them, *i.e.*, is such that $H_i \leq H_j$ only if $p(H_i) \leq p(H_j)$ (*cf*, Scott 1964, p. 246; Joyce 2011, p. 285).

> *Scott's Axiom.* If $\langle X_1, ..., X_n \rangle$ and $\langle Y_1, ..., Y_n \rangle$ contain the same number of truths as a matter of logic, so that $\sum_i w(X_i) = \sum_i w(Y_i)$ for any world $w$, then it is not true that $X_i \leq Y_i$ for all $i$ while $X_j < Y_j$ for some $j$.

In that case, according to the subjective Bayesian, she ought to adopt this unique probability $p$ as her prior (use $p$ to facilitate inference and decision-making).

[20]See Kyburg 1996, p. 326.

[21]Non-atomicity: for any $X$ such that $X > X\&\neg X$, there is a $Y$ such that $X\&Y > X\&\neg X$ and $X\&\neg Y > X\&\neg X$.

[22]Having or being representable by a set of probability distributions $S$ is a matter of having opinions that make it impermissible to estimate truth-values via any $p$ not in $S$.

[23]When considering uncountably many theoretical hypotheses, $p_D(X) = \int_0^1 x \cdot f_{V_X|D}(x)dz$, where $f_{V_X|D}$ is the density that defines the marginal distribution of $V_X$ conditional on $D$.

[24]When considering uncountably many theoretical hypotheses, $Exp_p(A) = \int_0^1 x \cdot f_{u|A}(x)dz$, where $f_{u|A}$ is the density that defines the marginal distribution of $u$ conditional on $A$.

[25]Your prior probability is $u(H) = \int_0^1 f(H|B = x) \cdot f(B = x)dx = \int_0^1 x\,dx = 1/2$ (where $f$ is the uniform density that defines $u$). Your posterior probability is

$$u_D(H) = \int_0^1 f_D(H|B = x) \cdot f_D(B = x)dx = \int_0^1 x \cdot 24.4 \cdot \sum_{k=480}^{520} \left( \begin{array}{c} 1000 \\ k \end{array} \right) (x)^k(1-x)^{1000-k} = 1/2.$$

[26]For discussion about features of epistemic utility beyond accuracy, see Maher 1993, ch. 9, and Joyce 1998, 2009.

[27]See Deza and Deza 2009 for a catalog of distance functions that one might employ in constructing accuracy measures for acceptance/rejection states.

[28]Note that the expected epistemic utility of $\mathscr{I}$, from the perspective of $p$, is just $p$'s best estimate of the objective expected utility of $\mathscr{I}$.

[29]I assume that $u$ treats any sequence of outcomes as exchangeable.

[30]For related discussion, see Gibbard 2008, p. 4, and Joyce 2009, p. 277.

[31]This proposal extends straightforwardly to other comparative judgments, *e.g.*, judgments of comparative preferability, and qualitative judgments, *e.g.*, judgments of incremental support.

[32]The cumulative distribution function $P$ corresponding to a distribution $f$ over chance hypotheses (defined by density $f$) is defined by $P(ch(X) \leq x) = \int_0^x f(y)dy$, and specifies the probability that the chance of $X$ is less than or equal to $x$.

[33]For example, for any beta densities $f$, $g$ and $h$, if they all have the same mean but increasing variance, then $f$ is closer to $g$ than to $h$. Similarly, if they all have the same variance but larger and larger means, then $f$ is closer to $g$ than to $h$.

[34]Of course, $Exp_u(eu(\mathscr{I}_x))$ does not take a *unique* minimum at $x = 0.52$. But this is to be expected, since in the relevant context, $\mathscr{I}_x = \mathscr{I}_y$ for any $x, y \in [0.52, 0.61]$.

[35]When an experiment has a small enough outcome space, we can simply examine the expected epistemic utility of every possible policy for responding to the experimental data, and check that the globalist policy attains a maximum. When the outcome space becomes too large, this is no longer feasible.

# Chapter 3

## Cliffordian Conservatism & Imprecise Prior Probabilities

When a doctor, or an engineer, or a scientist performs an experiment or test to adjudicate between competing theoretical hypotheses, or firm up her grounds for decision-making, she typically comes to the table with a great deal of relevant prior information. Any competent neurologist who is trying to diagnose a patient's disease, and settle on an appropriate treatment plan, not only has newly acquired clinical data — the results of blood tests, a lumbar puncture, etc. — but also an enormous amount of prior data: information about which symptoms correlate with which diseases, how those symptoms are caused, which treatments are most effective for which purposes, and so on. Obviously, it is imperative to take such prior information into account when making an inference or decision. To fail to do so is, as Jaynes says, "to commit the most obvious inconsistency of reasoning and may lead to absurd or dangerously misleading results" (Jaynes 1968, 1).

Unfortunately, finding a well-motivated, practically useful method for taking prior information into account is difficult. Prior information tends to be incredibly multifarious and complex. Precise Bayesians argue that the best method for incorporating prior evidence $E$ in decision and inference problems is to specify a 'prior' probability distribution $p$ over the competing hypotheses $H_1, ..., H_n$ which somehow summarizes

the information in $E$. We can think of these probabilities as estimates of the truth-values of $H_1, ..., H_n$ which (i) satisfy constraints imposed by $E$ while intuitively (ii) going no further than those constraints require. Imprecise Bayesians, such as Richard Jeffrey, Mark Kaplan, Peter Walley and Jim Joyce agree that an agent ought to take her prior evidence into account by adopting a 'prior' which summarizes it. But they disagree that priors should, in all circumstances, take the form of a single, precise probability distribution. Certain circumstances, they say, call for *imprecise priors*. In certain circumstances, you ought to use a *set* of distributions over $H_1, ..., H_n$ to incorporate your prior information. Like precise distributions, sets of distributions encode information about your prior evidence $E$. But a set of distributions encodes less information than any distribution in that set. It encodes only the information that is invariant across all elements of the set. The determinate properties of an imprecise prior (or posterior) $S$ are just the properties that all of the distributions in that set $S$ share in common.

Certain *subjectivist* proponents of imprecise probabilities say that an agent ought to look to her own opinions to furnish priors. And in many circumstances, an agent's actual actual prior opinions fail to pin down a single truth-value estimate for each of the theoretical hypotheses $H_1, ..., H_n$ under investigation (*cf.* Kyburg and Pittarelli 1996, 325). When they do, on the subjectivist view, she ought to adopt the set $S$ of distributions $p$ over $H_1, ..., H_n$ that are consistent with her prior opinions.

Joyce, Walley and others argue that there are more compelling reasons to adopt imprecise priors. Joyce contends that precise priors fail to adequately summarize certain kinds of evidence — in particular, unspecific and equivocal evidence. We need imprecise priors to summarize such evidence. Here is Joyce:

> ...the proper response to symmetrically ambiguous or incomplete evidence

> is not to assign probabilities symmetrically, but to refrain from assigning precise probabilities at all... Imprecise credences have a clear epistemological motivation: they are the proper response to unspecific evidence. (Joyce 2005, 171)

Suppose, for example, that you have a coin, but very little information about its bias (*cf.* Joyce 2010, 284). Perhaps you flip it over to check that it is not double sided. But nothing more. A bookie then offers you a bet. She kindly allows you to flip the coin a few times before deciding whether or not to accept or reject. You decide to adopt the uniform distribution $u$ over hypotheses $B = x$ about the coin's bias (with $0 < x < 1$ perhaps), to take account of your prior evidence $E$ (*viz.*, next to nothing). This might seem like an appropriate prior to adopt, since you have very little prior information, and $u$ is minimally informative (amongst precise priors, when measuring informativeness by Shannon entropy). But despite this fact, Joyce argues, the uniform distribution does a poor job summarizing your prior information. To see this, note that adopting $u$ commits you to making the following judgments:

- Rolling an ace with a fair 6-sided die is definitely less probable ($u(Ace) = 0.166$) than having the coin come up fewer than 17 times in 100 independent tosses ($u(Heads < 17) = 0.168$).
- It would be definite mistake to let \$100 ride on a rolling an ace than to let it ride on the coin coming up fewer than 17 times in 100 independent tosses.

Your prior evidence, however, is simply too unspecific to be this demanding. It might be specific enough to commit you to making *certain* comparative and qualitative judgments, *e.g.*, "It is more probable that the coin will come up heads than it is that the sun will suddenly expand and engulf the Earth." (You did, after all, see that

it is not double sided; it has a non-zero chance of coming up heads.) But it does *not* commit you to making specific judgments like the ones above. So any *prior* that commits you to making such judgments does a bad job summarizing your evidence. And precise priors as a class are bad in this respect. The moral is this: when you have unspecific and equivocal prior evidence, you should avoid precise priors altogether. You should adopt an imprecise prior instead. Imprecise priors, unlike precise priors, permit you to abstain from judgment on various issues. So they do not, as a class, capture improper responses to unspecific evidence, in the way that precise priors do.

The aim of this chapter is to point toward a *new* motivation for employing imprecise priors. You might hope for new reasons to employ imprecise priors because you find extent motivations less than fully compelling. (Perhaps you doubt that precise priors invariably capture improper responses to unspecific evidence, as I argue in chapter 2.) Or you might simply be interested in identifying the full range of reasons favoring imprecise Bayesianism. My plan is to highlight, for any interested parties, two new kinds of reasons for employing imprecise priors. We ought to adopt imprecise priors in certain contexts because they *put us in an unequivocally better position to secure epistemically valuable posterior beliefs* than precise priors do. We ought to adopt imprecise priors in various other contexts because they minimize our need for *epistemic luck* in securing such posteriors.

In §3.1, I investigate the theoretical role of priors, to illuminate what a compelling reason for adopting imprecise priors might look like. I suggest that the central role of priors is to help us secure *epistemically valuable* posterior beliefs, and to minimize our need for *epistemic luck* in securing those beliefs. In §3.2, I outline an argument that imprecise priors are sometimes best suited to play this role. In §3.3-3.6, I fill in this outline. In §3.3, I sketch an accuracy-centered approach to measuring the all-things-consider epistemic value or worth of imprecise priors and posteriors. Certain

of these measures, I argue, reflect *Jamesian liberalism*, while others reflect *Cliffordian conservatism*. In §3.4, I provide examples of conservative contexts in which imprecise priors put you in an unequivocally better position to secure epistemically valuable posteriors than precise priors do. This is the first new kind of reason to employ imprecise priors. In §3.5-3.6, I distinguish different types of epistemic luck, and illustrate how one prior might do more to ameliorate our dependence on luck than another. Lastly, in §3.7, I provide examples of conservative contexts in which imprecise priors do more to ameliorate dependence on epistemic luck than precise priors do. This is the second new kind of reason to employ imprecise priors.

## 3.1 The Theoretical Role of Priors

When we ask, "Why should we adopt imprecise priors?" we are asking for a certain kind of reason in response. If imprecise priors somehow made our knees less achy, or our jokes funnier, or our wallets fatter, that would be one reason to adopt them. But our question demands an *epistemic* answer, not a *pragmatic* one. Indeed, it demands a certain *kind* of epistemic answer. It demands reasons that speak to *the primary theoretical role of priors*. A *proper* answer to our question takes the form: we ought to adopt imprecise priors in certain contexts because they are best suited to play the relevant theoretical role (whatever that may be).

We must, then, be clear about what this theoretical role *is*. Some traditional, objective Bayesians, such as Edwin Jaynes, assume that the primary role of priors is *representational*. Jaynes prescribes adopting the maximum entropy prior for the "positive reason that it is... maximally noncommittal with regard to missing information" (Jaynes 1957, 623); the maximum entropy prior best *reflects* or *represents* the informational content of our prior evidence.

> *Informational Account.* The primary theoretical role of prior probabilities is to accurately reflect the informational content of the agent's prior evidence.

Certain subjective Bayesians agree that the primary role of priors is representational, but insist that Jaynes and others ought not restrict their attention to evidence. Prior probabilities ought to represent an agent's all-things-considered prior judgments about the plausibility of hypotheses, which might depend not only on her prior evidence, but also on her assessment of their intrinsic plausibility, her personal inductive quirks, etc.

> *Subjectivist Account.* The primary theoretical role of priors is to accurately represent the agent's prior opinions about the plausibility of hypotheses.

Neither of these accounts are quite right. The reason: *evidence* is important to our epistemic lives, at bottom, exactly because it helps us secure epistemically valuable (accurate, justified, sensitive, etc.) posterior beliefs in a luck-minimizing fashion. So *priors* — statistical tools for taking prior evidence into account — are plausibly important exactly to the extent that they help us achieve this end. They are important exactly to the extent that they put us in a position to secure valuable, minimally luck-dependent posterior beliefs by updating on new evidence.

> *Instrumental Account.* The primary theoretical role of priors is to put us in a position to secure epistemically valuable, minimally luck-dependent posteriors by updating on new data.

Imagine an objector who denies this. When the various roles listed above conflict, she will give precedence to one of the former ones, rather than the last one. Suppose,

for example, that a scientist has scant prior evidence about the causal mechanism under investigation (a particular virus' infection mechanism, perhaps). She *does*, however, find one particular hypothesis extremely intrinsically plausible. But she does not find it plausible for any good reason. Her hunch reflects no particular *skill* at assessing intrinsic plausibility. She simply 'feels it in her bones'. Our objector, if she favors the subjectivist account, will nevertheless advise her to adopt a prior that reflects this hunch, by concentrating probability on her favorite hypothesis. But this would be absurd. It would result in her discounting new data that she really ought to be more sensitive to (in much the way that a conspiracy theorist discounts data that tells against her favorite hypothesis, *e.g.*, that an alien spacecraft crashed near Roswell, New Mexico in 1947).

Alternatively, our researcher might have quite a lot of prior information, but find herself in an odd context of inquiry. For example, it might be much, much more epistemically important to avoid determinate error, in her context, than it is to get determinately close to the truth. Maybe *all* that matters is avoiding determinate error. (Some philosophers argue that the standards of evaluation operative in a contexts can depend on pragmatic factors. If this is correct, and there are much more serious negative consequences for getting it determinately wrong than there are positive consequences for getting it determinately right, then she might be in such a context.) Our objector, if she favors the informational account, will advise the researcher to adopt a prior that reflects the informativeness her evidence, presumably by concentrating probability on some hypothesis or other. But, if getting determinately close to the truth is *really* of no independent value — if avoiding determinate error is really all that matters in this context, from the epistemic perspective — then this is absurd. It is absurd in the way that gambling is absurd, if all that you care about is not losing money. You should simply not take the risk of gambling if all you

care about is not losing money. Similarly, our researcher should also not risk error by 'gambling' on some hypothesis or other (by concentrating probability on it). Instead, she should adopt a prior that encodes no opinion whatsoever about the virus' true infection mechanism, which no 'concentrated' prior does. She should adopt such a prior even though it does a rather poor job reflecting the informational content of her evidence.

This illustrates what should be clear: whichever prior best enables evidence to play *its* theoretical role is *ipso facto* best suited to play the theoretical role of priors. It is worth noting that the instrumental account does, in fact, enjoy a certain measure of support in the literature. James Berger (2006), for example, justifies the use objective Bayesian methods for constructing priors on the grounds that "objective Bayes intervals [95% confidence intervals] are, on average, smaller than the classically derived intervals," and have "better performance" in terms of average accuracy ("whether the interval contains the true $\theta$ or misses to the left or right") over a large number of independent trials (Berger 2006, 390-1). This is an appropriate justification, one might think, because the *job* of a prior is to put us in a position to secure accurate posteriors by updating on new data (and, plausibly, to do so in a way that minimizes our need for luck).

Similarly, Patrick Suppes says, "It is of fundamental importance to any deep appreciation of the Bayesian viewpoint to realize that the particular form of the prior distribution expressing beliefs held before the experiment is conducted is not a crucial matter... The well-designed experiment is one that will swamp divergent prior distributions with the clarity and sharpness of its results" (Suppes 1966, 204). The reason that it is not a crucial matter exactly which form the prior distribution takes is that, in a 'well-designed' experiment, where the experimental data is fairly 'weighty', a range of priors will converge on the true theoretical hypothesis (with high objec-

tive probability).[36] As a result, those priors are all likely to yield fairly accurate — and minimally luck-dependent — posterior distributions. Hence, they all play the primary theoretical role of priors close to equally well. And they do so even though some priors do a rather poor job representing, for example, the agent's prior opinions about the plausibility of hypotheses. This latter fact is — or at least ought to be — "not a crucial matter" from the Bayesian viewpoint.

## 3.2 Main Argument

The remainder of this chapter outlines two new kinds of reasons for employing imprecise priors. In broad strokes, the idea is this:

1. In any context of inquiry, you ought to adopt whichever prior is best suited to play the primary theoretical role of priors in that context, if there is one.

2. The primary role of priors is to help you secure epistemically valuable posterior beliefs, and to minimize your need for epistemic luck in securing those beliefs.

3. In certain contexts, imprecise priors put you in a better position to secure epistemically valuable posteriors than precise priors do.

4. In other contexts, no imprecise prior puts you in a better position to secure valuable posteriors than every precise prior, or vice versa. But imprecise priors minimize your need for epistemic luck.

C. In some contexts, you ought to adopt imprecise probabilities to incorporate your prior information.

If correct, this points the way toward a new, potentially promising foundation for imprecise Bayesianism. The reasons outlined here are not like those that concern

Joyce and others. They have nothing to do with whether imprecise priors are required to summarize certain kinds of evidence (*cf.* Joyce 2005, 2011). They are not like those that concern Walley, in many places, *viz.*, whether imprecise priors alone satisfy a range of nice symmetry and invariance principles (*cf.* Walley 1996). Instead, the reasons outlined here go straight to the heart of *what priors are for*; imprecise models are required because they often are best suited to play the primary theoretical role of priors.

To be clear, I will not offer a complete, systematic defense of this thesis. My aim here is limited. My aim is merely to *gesture toward* two new kinds of reasons for employing imprecise priors. To do this, I will simply provide examples of contexts in which (i) some imprecise prior puts you in a better position to secure valuable posteriors than precise priors do, and (ii) some imprecise does more to ameliorate dependence on luck than precise priors do.

## 3.3 Epistemic Value

### 3.3.1 General Remarks

Priors and posteriors often have a range epistemically laudable qualities. They are *accurate*, for example. The truth-value estimates they encode are close to the actual truth-values of the target hypotheses. (Imprecise priors and posteriors are determinately accurate when the family of truth-value assignments they encode are *all* close to the truth.) They are *well calibrated*. The relative frequency estimates they encode are close to the actual relative frequencies. They are *refined*.[37] They sort hypotheses into classes that are (more or less) uniformly true or false. They are *justified*. They capture appropriate responses to the available evidence. They are *informative*. They encode truth-value estimates for hypotheses that paint a rich, detailed picture

of the world, and so on.[38] For almost any property of qualitative beliefs that traditional epistemologists focus on — being reliably produced, sensitive, etc. — there are analogous properties of priors and posteriors.

Epistemic utility functions provide a measure of a prior (or posterior) $p$'s all-things-considered epistemic value or worth at a world $w$, which it has in virtue of having these laudable qualities to a greater or lesser extent at $w$. As Joyce (1998, 2009) stresses, while there is room for reasonable dispute about the relative importance of certain qualities, any reasonable epistemic utility function must reflect an overriding concern for *accuracy*. Ceteris paribus, priors and posteriors are all-things-considered better, from the epistemic perspective, the more accurate they are.[39]

What we might call *pluralist approaches* to theorizing about epistemic value treat various of these laudable qualities — accuracy, justification, informativeness, etc. — as making an independent contribution to all-things-consider epistemic worth. *Accuracy-centered approaches*, in contrast, treat 'auxiliary' virtues — everything but accuracy — as relevant to the epistemic value of a prior or posterior only to the extent that they are reflected in its accuracy (*cf.* Joyce 2013). On this view, our carefully considered judgments regarding justification, and so on, may well influence how we value 'closeness to the truth', how we measure $p$'s accuracy at $w$.[40] But this is the only route by which they affect epistemic value. (This is a bit too narrow of a characterization, but will do for now.) In what follows, I will sketch an accuracy-centered approach to theorizing about the epistemic value of both precise and imprecise priors.

How exactly should we think about the accuracy of a prior (or posterior) at a world? Following Joyce (1998, 2009), Predd *et al.* (2009), and Leitgeb and Pettigrew (2010), we will measure the accuracy of *precise* priors/posteriors by an *epistemic scoring rule* or *inaccuracy score*. An inaccuracy score is a function $\mathscr{I}$, which maps probability distributions $p$ and worlds $w$ to non-negative real numbers, $\mathscr{I}(p, w)$. $\mathscr{I}(p, w)$

measures how inaccurate $p$ is if $w$ is actual. If $\mathscr{I}(p, w)$ equals zero, then $p$ is minimally inaccurate (maximally accurate) at $w$. Inaccuracy increases as $\mathscr{I}(p, w)$ grows larger. As in Joyce (2009, p. 269, 280), we assume that any reasonable inaccuracy score satisfies the following two conditions:

> **Truth-Directedness**. Moving a prior $p$'s probabilities, or truth-value estimates closer to the actual truth-values always improves accuracy. If $p$ and $q$ differ only in that $p$ assigns higher (lower) probabilities than $q$ does to some propositions that are true (false) at $w$, then $\mathscr{I}(p, w) < \mathscr{I}(q, w)$, *i.e.*, $p$ is less inaccurate (more accurate) than $q$ at $w$.[41]
>
> **Coherent Admissibility**. No probabilistically coherent prior $p$ is less accurate than an incoherent prior $q$ in *every* world. More carefully, it is not the case that (i) $\mathscr{I}(p, w) \geq \mathscr{I}(q, w)$, for all $w$, and (ii) $\mathscr{I}(p, w') > \mathscr{I}(q, w')$, for some $w'$.

**Truth-Directedness** guarantees that moving prior probabilities (truth-value estimates) closer to the truth always has a net positive impact on accuracy (*cf.* Joyce 2013, 3). (The positive effect associated with getting closer to the truth always outweighs the negative effect associated with becoming less well-calibrated, less justified, etc.) **Coherent Admissibility** guarantees that reasonable inaccuracy scores cohere with our most robust intuitions about which priors are appropriate to adopt in which evidential circumstances. For any probabilistically coherent prior $p$, we can find evidential circumstances in which $p$ seems very clearly to be the right prior to adopt (*cf.* Joyce 2009, 279). If, however, $p$ is less accurate than some other $q$ in *every* world according to an inaccuracy function $\mathscr{I}$, then $p$ is epistemically defective from $\mathscr{I}$'s perspective. You should not adopt $p$ in any circumstances, according to $\mathscr{I}$. Coherent admissibility sees this flouting of our most robust intuitions as a mark of $\mathscr{I}$'s

unreasonableness.

These constraints rule out many inaccuracy scores as unreasonable.[42] Nevertheless, a number of scores satisfy both Truth-Directedness and Coherent Admissibility. The Brier score, for example, satisfies both constraints. The *Brier score*, $\mathscr{I}$, measures the inaccuracy of a prior $p$, defined over a partition $\langle H_1, ..., H_n \rangle$, at a world $w$, by the average squared Euclidean distance between its truth-value estimates, $p(H_i)$, and the actual truth-values at $w$, $w(H_i)$. That is, $\mathscr{I}(p, w) = (1/n) \sum_i (p(H_i) - w(H_i))^2$. The *logarithmic score*, which measures the inaccuracy of $p$ at $w$ by $(1/n) \sum_i -ln[|1 - w(H_i) - p(H_i)|]$, also satisfies both constraints. So do various other scores, *e.g.*, the power score, spherical score, any other proper scoring rule.[43,44] We will measure inaccuracy by the Brier score in what follows.

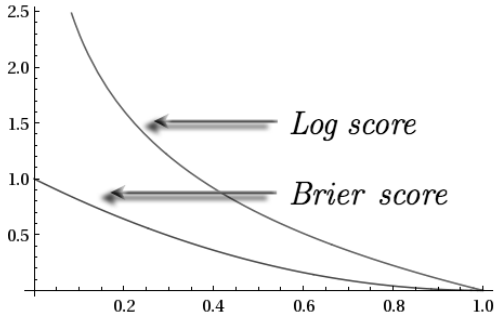


Figure 3.1: Inaccuracy of $p(H_i) = x$ when $H_i$ is true, relative to Brier and log scores, respectively.

Unlike precise priors, imprecise priors and posteriors are typically not accurate to any determinate degree. The determinate properties of an imprecise prior (or posterior) $S$ are just the properties that all of the distributions in $S$ share in common. So, for example:

- An imprecise prior encodes a determine probability $x$ for a hypothesis $H$ only if every element of that set agrees that the probability of $H$ is $x$.

- An imprecise prior carries a commitment to making a qualitative or comparative judgment, *e.g.*, that $X$ is more probable than $Y$, or that $X$ is independent of $Y$, only if every element of that set carries that commitment.

- An imprecise prior is accurate to a determinate degree $y$ only if every element of that set is accurate to degree $y$.

Except in very special circumstances, the elements $p$ of an imprecise prior $S$ will vary in their accuracy at a world $w$. So typically imprecise priors (and posteriors) $S$ will not be accurate to any determinate degree at $w$.

This does *not* mean, however, that there is nothing determinate to say about the accuracy of imprecise priors. Often there is *quite a lot to say.* For example, suppose that you are wondering about the amount of rain $R$ that Bristol received yesterday. A friend informs you that the best estimate of $R$ is 2.5mm. Assume that $R$ can only take one of three values: $R = 1$ in world $w_1$, $R = 2$ in world $w_2$, and $R = 3$ in world $w_3$. To incorporate your prior information, you adopt an imprecise prior: the set $S$ of probability distributions $p$ over $\langle w_1, w_2, w_3\rangle$ consistent with the constraint imposed by your prior information, *viz.*, $p(w_1) + 2p(w_2) + 3p(w_3) = 2.5$. Then your prior $S$ is not accurate to any determinate degree, whatever world is actual. To see this, suppose that $w_1$ is actual (Bristol actually received 1mm of rain). Then there are distributions $p$ and $q$ in $S$ such that $\mathscr{I}(p, w_1) = 0.42 \neq 0.445 = \mathscr{I}(q, w_1)$.[45] Nevertheless, there *are* quite strong determinate facts about $S$'s accuracy at $w_1$. For example, we can say that $S$ does at least this poorly, *i.e.*, is at least this inaccurate: 0.375. (Every element $p$ of $S$ is such that $\mathscr{I}(p, w_1) \geq 0.375$.) Similarly, we can say that $S$ does at most this poorly: 0.5. (Every element $p$ of $S$ is such that $\mathscr{I}(p, w_1) \leq 0.5$.) This grounds a rather rich set of determinate facts about $S$'s comparative accuracy. We can say, for example, that $S$ is determinately *more* accurate than the precise prior $r$ which is such that $r(w_1) = 0.01$, $r(w_2) = 0.01$, $r(w_1) = 0.98$ and determinately *less* accurate than the prior $t$ with $t(w_1) = 0.5$, $r(w_2) = 0.25$, $r(w_1) = 0.25$. The inaccuracy of $r$ and $t$ at $w_1$ is $\mathscr{I}(r, w_1) = 0.647$ and $\mathscr{I}(t, w_1) = 0.125$, respectively.

The accuracy-centered theorist might propose that these determinate facts are sufficient to determine precise degrees of all-things-considered epistemic value or worth. Imprecise priors and posteriors $S$ are *epistemically valuable* to a determinate degree at worlds $w$, even though they are not *accurate* to a determinate degree at $w$. Their epistemic value supervenes on what we will call their *lower* and *upper-inaccuracy scores*.

- $l(S, w) = inf\{\mathscr{I}(p, w)|p \in S\}$
- $u(S, w) = sup\{\mathscr{I}(p, w)|p \in S\}$

The first quantity, $l(S, w)$, is $S$'s *lower-inaccuracy*. For any prior $S$ and world $w$, $S$ does at least this poorly at $w$, *i.e.*, is at least this inaccurate: $l(S, w)$. (Every element $p$ of $S$ is such that $\mathscr{I}(p, w) \geq l(S, w)$.) The second quantity, $u(S, w)$, is $S$'s *upper-inaccuracy*. $S$ does at most this poorly: $u(S, w)$. (Every element $p$ of $S$ is such that $\mathscr{I}(p, w) \leq u(S, w)$.) As noted above, these quantities ground a rich set of facts about $S$'s comparative accuracy. If $l(S, w)$ and $u(S, w)$ are both greater than $l(S', w)$ and $u(S', w)$, then $S$ is determinately more inaccurate (less accurate) than $S'$ at $w$. If $l(S, w)$ and $u(S, w)$ are both less than $l(S', w)$ and $u(S', w)$, then $S$ is determinately less inaccurate (more accurate) than $S'$ at $w$. If neither is true, then there is no determinate fact about the comparative inaccuracy of $S$ and $S'$ at $w$.

The reason that a prior $S$'s upper and lower-inaccuracy scores ground a precise degree of epistemic value, an accuracy-centered theorist might say, is this: they allow us to say how $S$ fares with respect our two "great commandments as would-be knowers", *viz.*, Believe truth! Shun error! (James 1896, §VII). At bottom, she might continue, epistemic value is a matter of obeying these two commands. (This is in keeping with the spirit of the accuracy-centered approach.) And imprecise priors, in virtue of having upper and lower-inaccuracy scores, obey these two commands to a

determinate degree. A prior $S$'s lower-inaccuracy score, $l(S, w)$, provides a measure of the extent to which $S$ avoids determinate error at $w$. Its upper-inaccuracy score, $u(S, w)$, provides a measure of the extent to which $S$ determinately converges on the truth at $w$. So any imprecise prior $S$ ought to count as epistemically valuable to a determinate degree at any world $w$.

We will measure the all-things-considered epistemic value or worth of a prior (or posterior) at a world by an epistemic disutility score. An epistemic disutility score is a function $\mathscr{D}$, which maps priors, or sets $S$ of probability distributions (treating precise priors as singleton sets), and worlds $w$ to non-negative real numbers, $\mathscr{D}(S, w)$. $\mathscr{D}(S, w)$ measures how much disutility or disvalue $S$ has if $w$ is actual, from the epistemic perspective. If $\mathscr{D}(S, w)$ equals zero, then $S$ has minimal epistemic disutility (maximal utility) at $w$. Epistemic disutility increases as $\mathscr{D}(S, w)$ grows larger. The accuracy-centered theorist assumes that any reasonable epistemic disutility score satisfies at least the following four conditions:

> **Extensionality**. The epistemic disutility of a prior $S$ at a world $w$ is solely a function of $S$'s upper and lower-inaccuracy scores at $w$.
>
> **Continuity**. Epistemic disutility scores are continuous.
>
> **Upper/Lower Dominance**. Moving a prior $S$'s upper and lower inaccuracy scores uniformly downward always improves epistemic utility. If $u(S, w)$ and $l(S, w)$ are both less than $u(S', w)$ and $l(S', w)$, then $\mathscr{D}(S, w) < \mathscr{D}(S', w)$.
>
> **Normalization**. When a prior $S$ has a determinate degree of inaccuracy at $w$, its epistemic disutility at $w$ just is that degree of inaccuracy. If $\mathscr{I}(p, w)=x$, for all $p$ in $S$, so that $u(S, w)=l(S, w)=x$, then $\mathscr{D}(S, w)=x$.

**Extensionality** guarantees that accuracy is the cardinal 'epistemic good'. 'Auxiliary' goods — calibration, justification, etc. — impact the all-things-considered epistemic value or worth of a prior/posterior $S$ at a world $w$ by impacting how we value 'closeness to the truth' (how we measure accuracy), or by impacting how we balance off different determinate facts about accuracy (upper and lower-inaccuracy scores) against one another to arrive at an all-things-considered epistemic (dis)utility score. **Continuity** guarantees that small changes to facts about accuracy do not result in excessively large changes in facts about epistemic utility. **Upper/Lower Dominance** guarantees that determinate improvements in accuracy always result in determinate improvements in epistemic utility. Finally, **Normalization** guarantees that when a prior is informationally rich enough to pin down a precise degree of accuracy, nothing else matters to its epistemic utility. Its degree of epistemic (dis)utility *just is* its degree of (in)accuracy.

## 3.4 Cliffordian Conservatism and Jamesian Liberalism

William James emphasizes the first of our two "great commandments as would-be knowers," *viz.*, Believe truth! The risk of being in error, he says, is "a very small matter when compared with the blessings of real knowledge." W. K. Clifford, on the other hand, emphasizes the second, Shun error! "It is wrong always, everywhere, and for anyone," he says, "to believe anything upon insufficient evidence" (Clifford 1877).

All inaccuracy scores strike some balance between these two commandments, and in this way take some stand in the Clifford/James debate (*cf.* Joyce 2009, 281). The more *convex* an inaccuracy score $\mathscr{I}$ is, the more it emphasizes avoiding error; the more it reflects Cliffordian conservatism. An inaccuracy score $\mathscr{I}$ is convex at a world $w$ if $(1/2)\mathscr{I}(p,w)+(1/2)\mathscr{I}(q,w) \geq \mathscr{I}((1/2)p+(1/2)q,w)$, for any prior distributions

$p$ and $q$.[46] The more concave it is, the more it emphasizes the pursuit of truth; the more it reflects Jamesian liberalism.

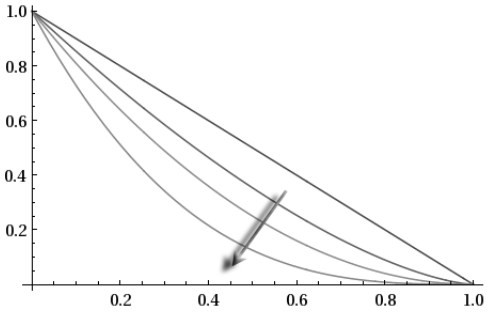


Figure 3.2: Inaccuracy of $p(H_i) = x$ when $H_i$ is true, according to increasingly convex exponential scores.

To illustrate, compare the Brier score, $\mathscr{I}(p, w) = (1/n)\sum_i |p(H_i) - w(H_i)|^2$, which is convex, with the power score ($z = 8$), $\mathscr{I}^*(p, w) = (1/n)\sum_i 7p(H_i)^8 + w(H_i) \cdot (1 - 8p(H_i)^7)$, which is almost everywhere concave. Suppose that I have an urn containing black, green and yellow balls mixed in some unknown proportion. (See Joyce 2009, p. 283 for a similar example.) I decide to adopt the uniform prior $u$ over hypotheses $H$ regarding the chance of drawing a black, green or yellow ball, respectively. So my prior probability (truth-value estimate) for observing a black ball on next draw is 1/3 (similarly for green and yellow). Now imagine that *you* draw a ball and observe that it is black. You decide not to tell me the outcome of your draw outright. But you have a pill that you can give me, which will randomly raise or lower my prior probability for *Black* (the proposition that the selected ball is black), with equal chance, by 1/3, while leaving the rest of my prior probabilities the same. What should you do?

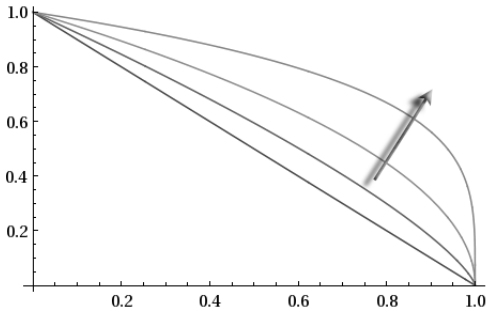


Figure 3.3: Inaccuracy of $p(H_i) = x$ when $H_i$ is true, according to increasingly concave exponential scores.

If all you care about is the accuracy of my prior, and you measure inaccuracy by the (mostly) concave power score $\mathscr{I}^*$, then you

should opt for the more aggressive, truth-seeking option. You should give me the pill. Because $\mathscr{I}^*$ is concave, the benefits of getting closer to the truth significantly outweigh the costs of getting further from it. This is reflected in your best estimates of the inaccuracy of my prior conditional on my taking the pill, on the one hand, and standing pat, on the other. You are sure that if I stand pat, then my prior probabilities for *Black*, *Green* and *Yellow*, respectively, are inaccurate to degree 0.333. Your best estimate of their inaccuracy if I take the pill, however, is 0.302. You expect me to improve, in terms of inaccuracy, if I take the pill.

If you measure inaccuracy by the convex Brier score $\mathscr{I}$, however, then you should opt for the more conservative option. You should tell me to stand pat. Because $\mathscr{I}$ is convex, the benefits of getting closer to the truth pale in comparison to the costs of getting further from it. This, again, is reflected in your best estimates of the inaccuracy of my prior, conditional on my taking the pill/standing pat. You are sure that if I stand pat, my prior probabilities are inaccurate to degree 0.222. Your best estimate of their inaccuracy if I take the pill, however, is 0.259. You expect me to do worse, in terms of inaccuracy, if I take the pill.

The moral is this: the convexity/concavity properties of inaccuracy scores reflect some way of balancing our two "great commandments as would-be knowers", *viz.*, Believe truth! Shun error! Concave scores place more of an emphasis on believing the truth. Convex scores place more of an emphasis on avoiding error.

On the accuracy-centered view, every reasonable epistemic disutility score $\mathscr{D}$ is a function of some inaccuracy score $\mathscr{I}$. The epistemic disutility of $S$ at $w$, $\mathscr{D}(S, w)$, is determined by $S$'s upper and lower-inaccuracy scores at $w$:

- $l(S, w) = inf\{\mathscr{I}(p, w) | p \in S\}$
- $u(S, w) = sup\{\mathscr{I}(p, w) | p \in S\}$

Epistemic disutility scores strike a balance between our 'two great commandments', then, by featuring inaccuracy scores which strike a particular balance. Some scores $\mathscr{D}$ feature the Brier score, while others $\mathscr{D}'$ feature the logarithmic score, while still other $\mathscr{D}''$ feature a power score, and so on. In virtue of the different balances that these inaccuracy scores strike, $\mathscr{D}$, $\mathscr{D}'$ and $\mathscr{D}''$ count as doing so as well, as taking different positions in the Clifford/James debate. But epistemic disutility scores strike a balance between our 'two great commandments' in another way too. Different disutility scores afford upper and lower-inaccuracies, $u(S, w)$ and $l(S, w)$, different degrees of relative importance. Depending on which dictum you are inclined to place more emphasis on — Believe truth! or Shun error! — you will see certain weightings as more reasonable than others.

Jamesians will see disutility scores $\mathscr{D}$ that treat $u(S, w)$ as more important than $l(S, w)$ as capturing a more reasonable view about how to balance off all of the determinate facts about $S$'s accuracy at $w$ to arrive at an all-things-considered judgment about epistemic worth. "Avoiding determinate error," they will say, "is a very small matter when compared with the blessings of getting determinately close to the truth." Avoiding determinate error is a matter of having a low *lower*-inaccuracy score, $l(S, w)$. Getting determinately close to the truth, in contrast, is a matter of having a low *upper*-inaccuracy score, $u(S, w)$. The upshot: any reasonable measure of epistemic disutility, or all-things-considered epistemic disvalue, according to the Jamesian, will count $u(S, w)$ as much more important than $l(S, w)$. It will reward a prior $S$ more for having $u(S, w)$ close to zero than for having $l(S, w)$ close to zero.

Cliffordians, on the other hand, will see disutility scores $\mathscr{D}$ that treat $l(S, w)$ as more important than $u(S, w)$ as capturing a more reasonable view about all-things-considered epistemic worth. "The sin of being in determinate error," they will say, "is a much greater offense than the sin of failing to get determinately close to the truth."

So, any reasonable measure of epistemic disutility, or all-things-considered epistemic disvalue, according to the Cliffordian, will count $l(S, w)$ as much more important than $u(S, w)$. It will penalize a prior $S$ much more for having a high lower-inaccuracy score (*i.e.*, for being in determinate error) than it will for having a high upper-inaccuracy score (*i.e.*, for failing to get determinately close to the truth).

In what follows, I will consider only the simplest epistemic disutility scores, 'linear scores' of the form:

$$\mathscr{D}_\lambda(S, w) = \lambda \cdot l(S, w) + (1 - \lambda) \cdot u(S, w).$$

Linear scores $\mathscr{D}_\lambda$ with $\lambda > 1/2$ treat lower inaccuracy, $l(S, w)$, as more important than the upper inaccuracy, $u(S, w)$. We call these *Cliffordian disutility scores*. Linear scores $\mathscr{D}_\lambda$ with $\lambda < 1/2$ treat $u(S, w)$ as more important than $l(S, w)$. We call these *Jamesian disutility scores*.

Linear disutility scores are 'reasonable', in the sense described in §3.3.1. They satisfy **Extensionality**, and so guarantee that accuracy, in some sense, is the cardinal 'epistemic good'. They satisfy **Continuity**, and so guarantee that small changes to facts about accuracy do not result in large changes in facts about epistemic utility. They satisfy **Upper/Lower Dominance**, and so guarantee that determinate improvements in accuracy always result in determinate improvements in epistemic utility. Finally, they satisfy **Normalization**, and so guarantee that when a prior is informationally rich enough to pin down a precise degree of accuracy, nothing else matters to its epistemic utility.

Many other epistemic disutility scores satisfy these constraints as well. I focus on linear disutility scores because the aims of this chapter are limited. I only hope to gesture toward some new kinds of reasons for employing imprecise priors. And for this end, it is sufficient to simply provide examples of contexts in those reasons are

extremely salient. Linear disutility scores are particularly useful for furnishing such examples.

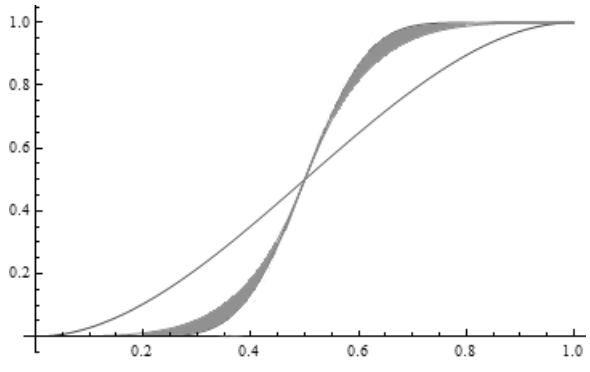


Figure 3.4: The Cramer-von Mises distance between two beta distributions, given by a function of the area between their respective cumulative distribution functions.

For concreteness, I will also restrict my attention to disutility functions that feature the Brier score, at least when considering discrete priors (priors defined over countably many theoretical hypotheses). When we consider continuous priors (priors defined over uncountably many theoretical hypotheses, *e.g.*, chance hypotheses), we will measure inaccuracy by *Cramer-von Mises distance*, which is a natural extension of squared Euclidean distance ('Brier distance') to the space of continuous distributions.[47] The Cramer-von Mises distance between continuous distributions $p$ and $q$, $\mathfrak{C}(p,q) = \int_0^1 |P(x) - Q(x)|^2 dx$, is just the squared $L_2$ metric between their respective cumulative density functions, $P$ and $Q$ (Deza and Deza 2009, 245).[48] $\mathfrak{C}(p,q)$ specifies the distance between $p$ and $q$ as a function of the area between the CDFs, $P$ and $Q$, counting regions of smaller divergence for less and regions of greater divergence for more (left, previous page). The proposal, a bit more carefully then, is to measure the inaccuracy of a continuous prior $p$ at a world in which $H$ is true by the Cramer-von Mises distance between $p$ and the indicator distribution $\iota_H$, which is defined by the Dirac density that centers all of its probability mass on $H$. Again, restricting our attention in this way will allow us to provide specific examples of contexts in which the reasons for employing imprecise priors are extremely salient.

## 3.5 A Dominance Argument for Imprecise Priors

We ought to adopt imprecise priors in certain contexts because they *put us in an unequivocally better position to secure epistemically valuable posterior beliefs* than precise priors do. In such contexts, imprecise models are better suited to play the primary theoretical role of priors than precise models are.

What exactly does it take, though, for one prior to put you in a better position to secure epistemically valuable posteriors than another prior? We can get a handle on this question, I suggest, by comparing the *objective expected posterior epistemic value* of different priors $S$ across all relevant theoretical hypotheses. In any experimental context aimed at adjudicating between hypotheses $H_1, ..., H_n$, we can ask: how (objectively) likely is it that the experiment will yield any particular data item $D_1, ..., D_m$ if hypothesis $H_i$ is true? We can also ask: to what extent will $S$ converge on $H_i$ when conditioned on $D_j$? How epistemically valuable will the posterior, $S_{D_j}$, be as a result? Finally, we can ask what the (objectively) best estimate of $S$'s posterior epistemic value is if hypothesis $H_i$ is true. What is $Exp_{H_i}(\mathscr{D}(S_D, H_i))$? In certain cases, I claim, facts about these (objective) best estimates settle our question definitively; they settle the matter of whether one prior $S$ puts you in a better position than another prior $S^*$ to secure epistemically valuable posteriors.

Consider a concrete case. A scientist is going to perform an experiment to adjudicate between competing theoretical hypotheses $H_1, ..., H_n$ about whether (and how) over expression of a certain gene causes chromosomal instability in breast tumors. She has a great deal of relevant prior evidence $E$: information about the levels of different genes expressed in past patients, as well as their various clinical symptoms, recurrence rates, etc.; information about the broader causal mechanisms that give rise to breast cancer, and so on. If two priors, $S$ and $S'$, both satisfy the constraints

imposed by $E$, but $S$'s objective expected posterior epistemic disutility is lower than $S^*$'s *relative to every theoretical hypothesis* $H_i$, then $S$ *must* put her in a better position than $S^*$ to secure epistemically valuable posteriors. Whatever else is true about "putting oneself in a good position" with respect to some goal, it must be the case that if one option gives you a better chance of achieving the goal than another, however the world happens to be (whatever the true chance hypothesis is), then that option puts you in a better position with respect to that goal.

So we have a sufficient condition for one prior $S$ to put you in a better position than another $S^*$, in terms of securing epistemically valuable posteriors:

> (★) $S$ puts you in a better position than $S^*$ to secure epistemically valuable posteriors if $S$'s objective expected posterior epistemic disutility is lower than $S^*$'s relative to all $H_i$.

The goal now is to provide examples of contexts in which imprecise priors put you in an unequivocally better position to secure epistemically valuable posterior beliefs than precise priors do.

In some contexts, certain *precise* priors put you in a better position to secure epistemically valuable posteriors than imprecise priors do. Imagine that a bookie hands you a coin and offers you a bet. You have no prior evidence about the coin's bias. But the bookie allows you to flip the coin for awhile — 5 times, for example — prior to deciding whether or not to take the bet. Consider two options that you have for taking your prior

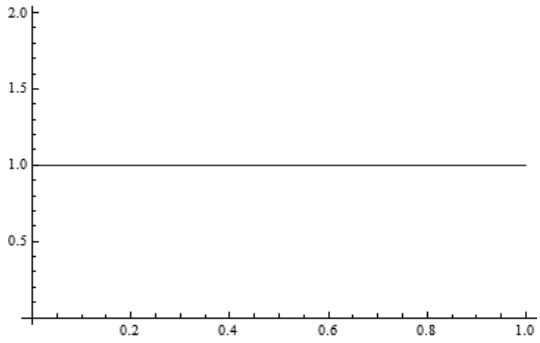


Figure 3.5: MaxEnt prior $u$ over hypotheses $B = x$.

information (*viz.*, none) in your decision problem. (Of course, these are not the only two options.) Option 1: adopt the (precise) maximum entropy (uniform) prior $u$ over hypotheses $B = x$ about the coin's bias. Option 2: adopt an *imprecise beta-binomial model*, with some level of concentration $s$, *e.g.*, $s = 10$ (see Walley 1991, §5.3, for a more detailed exposition). Beta distributions $b$ are parameterized by two quantities, $\alpha$ and $\beta$. These 'shape parameters' determine which hypotheses $B = x$ the distribution $b$ focuses its probability mass on. The concentration parameter, $s = \alpha + \beta$, corresponds roughly to how 'peaked' $b$ is around its mean. The imprecise beta model with concentration 10 is the set $\mathscr{M}_{10}$ of all beta distributions $b$ with $s = 10$ (examples of such distributions pictured left). *Nota bene*: I focus on beta priors in what follows because (i) they are very rich; any prior distribution can be approximated by a finite mixture of beta distributions; (ii) they are mathematically tractable; they generate beta posterior distributions. (*cf.* Walley 1996, 9).

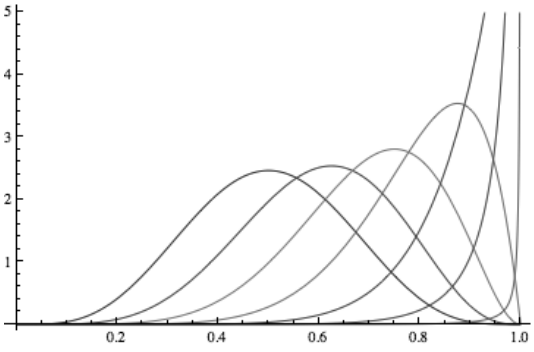


Figure 3.6: Beta distributions with concentration $s = 10$.

Suppose that, given the standards of evaluation operative in your context of inquiry, the appropriate measure of epistemic disutility $\mathscr{D}$ is *completely Jamesian*:

$$\mathscr{D}(S, w) = \mathscr{D}_0(S, w) = 0 \cdot l(S, w) + 1 \cdot u(S, w) = sup\{\mathscr{I}(p, w) | p \in S\}$$

Such a disutility function $\mathscr{D}$ reflects an unmitigated commitment to getting determinately close to the truth. It sees *no* independent value in avoiding determinate error. Posteriors that get determinately close to the truth (have low upper-inaccuracy scores), of course, will also avoid determinate error (have low lower-inaccuracy scores).

But error avoidance is not something to be sought for its own sake, on this view.

The important observation is this: given a completely Jamesian measure $\mathscr{D}$, the MaxEnt prior $u$'s objective expected posterior epistemic disutility is lower than $\mathscr{M}_{10}$'s relative to *every* chance hypothesis $B = x$. The reason: for any chance hypotheses $B = x$ and any data sequence $D = H^k T^{5-k}$, there is some distribution $b$ in $\mathscr{M}_{10}$ that converges on $B = x$ *less* than $u$ does, when conditioned on $D$. (This is just a consequence of the 'inclusiveness' of the imprecise beta model $\mathscr{M}_{10}$.) This ensures that the upper-inaccuracy of $u$ is lower than the upper-inaccuracy of $\mathscr{M}_{10}$, whichever data sequence you observe, and whatever the true chance hypothesis happens to be. And *that* guarantees that $u$'s objective *expected* posterior epistemic disutility is lower than $\mathscr{M}_{10}$'s, come what may. It guarantees that $u$ dominates $\mathscr{M}_{10}$, in terms of objective expected disutility.

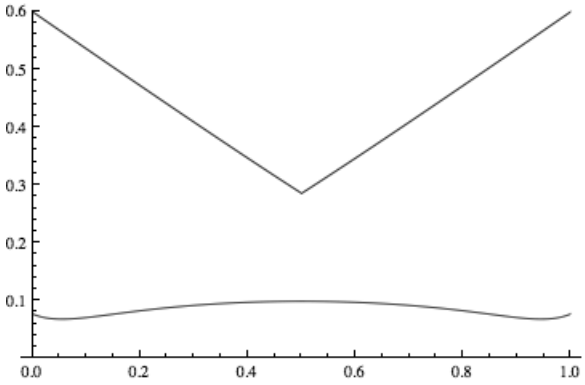


Figure 3.7: Top curve: $\mathscr{M}_{10}$'s objective expected (Jamesian) disutility, relative to chance hypotheses $B = x$. Bottom: $u$'s objective expected disutility, relative to $B = x$.

This shows that there are contexts in which certain precise priors (*viz.*, the MaxEnt prior) put you in a better position to secure epistemically valuable posteriors than certain imprecise priors (*viz.*, the imprecise beta prior with concentration $s = 10$). We will now show that the converse occurs as well. There are contexts in which certain *imprecise* priors put you in a better position than various precise priors to secure epistemically valuable posteriors. In fact, there are contexts in which they put you in a better position than *any* reasonable precise prior. *This provides decisive epistemic*

*reason to employ imprecise priors in those contexts.*

Suppose once more that you are considering different options for incorporating your prior information about the coin's bais (*viz.*, none) in your decision problem. Now, however, the operative standards of evaluation yield a measure of epistemic disutility $\mathscr{D}$ that is *completely Cliffordian*, rather than Jamesian.

$$\mathscr{D}(S,w) = \mathscr{D}_1(S,w) = 1 \cdot l(S,w) + 0 \cdot u(S,w) = inf\{\mathscr{I}(p,w) | p \in S\}$$

Such a disutility function $\mathscr{D}$ reflects unadulterated concern for avoiding error. It sees no independent value in getting determinately close to the truth. Of course, getting determinately close to the truth (have a low upper-inaccuracy score) is instrumentally valuable; it guarantees avoidance of determinate error (it guarantees a low lower-inaccuracy score). But getting determinately close to the truth is not worth pursuing for its own sake, on this view.

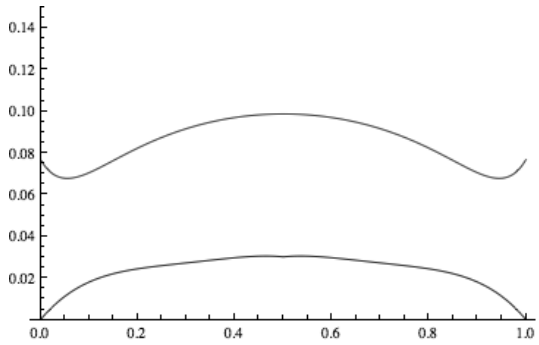


Figure 3.8: Top curve: $u$'s objective expected (Cliffordian) disutility, relative to chance hypotheses $B{=}x$. Bottom: $\mathscr{M}_{10}$'s objective expected disutility, relative to $B{=}x$.

Given a completely Cliffordian measure $\mathscr{D}$, the imprecise beta model, $\mathscr{M}_{10}$, dominates the MaxEnt prior $u$, in terms of objective expected posterior epistemic disutility. The reason: for any chance hypotheses $B = x$ and any data sequence $D = H^k T^{5-k}$, there is some distribution $b$ in $\mathscr{M}_{10}$ that converges on $B = x$ *more* than $u$ does when conditioned on $D$. (Again, this is just a consequence of the 'inclusiveness' of the imprecise beta model $\mathscr{M}_{10}$.) This is sufficient to guarantee that the lower-inaccuracy of $\mathscr{M}_{10}$ is less than the lower-inaccuracy of $u$, whichever data sequence you observe, and what-

ever the true chance hypothesis happens to be. This, in turn, guarantees that $\mathscr{M}_{10}$'s objective *expected* posterior epistemic disutility is lower than $u$'s, come what may.

This phenomenon — objective expected disutility domination — is rare. But it is not *entirely* restricted to contexts in which the appropriate measure of all-things-considered epistemic value or worth is given by a maximally Jamesian or Cliffordian disutility score. Compare, for example, the objective expected disutilities of the imprecise beta model, $\mathscr{M}_{10}$, and the Max-Ent prior, $u$, when the appropriate disutility score is $\mathscr{D}_\lambda(S,w) = \lambda \cdot l(S,w) + (1-\lambda) \cdot u(S,w)$, with $\lambda > 0.9$. Such scores are Cliffordian, but not *maximally* Cliffordian. In any such context, $\mathscr{M}_{10}$ dominates $u$, in terms of objective expected disutility (right).

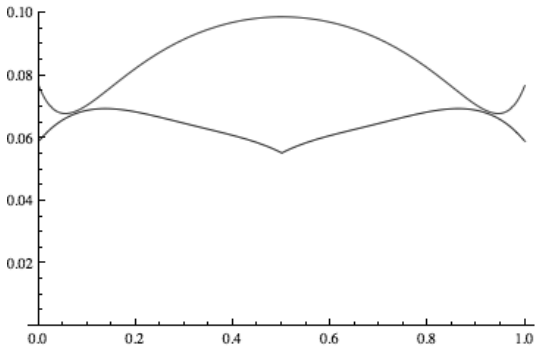


Figure 3.9: Top curve: $u$'s objective expected disutility, relative to chance hypotheses $B{=}x$, measuring disutility by $\mathscr{D}_{0.9}$. Bottom: $\mathscr{M}_{10}$'s objective expected disutility, relative to $B{=}x$.

Or consider a context in which the appropriate disutility score is $\mathscr{D}_{0.95}(S,w) = 0.95{\cdot}l(S,w){+}0.05{\cdot}u(S,w)$. Again, this score is Cliffordian, but not maximally so. Relative to *this* score, $\mathscr{M}_{10}$ dominates *all beta priors* $b$, in terms of objective expected disutility, other than those with excessively low entropy.[49] It is not difficult to see why, either. For every relatively high entropy beta prior $b$, every chance hypothesis $B = x$, and every data sequence $D$, there is a lower entropy $b'$ in $\mathscr{M}_{10}$ that converges more on $B = x$ when conditioned on $D$ than $b$ does. Only if $b$ has particularly low entropy, and is centered on $B = x$, will it tend to converge on $B = x$ more than any prior in $\mathscr{M}_{10}$.

The consequence is that the imprecise beta prior $\mathscr{M}_{10}$ puts you in a better position

to secure epistemically valuable posteriors than any *prima facie* reasonable precise beta prior. Why? Because excessively low entropy priors are (at least) *prima facie* unreasonable, in this context. Given that you have no relevant prior evidence about the coin's bias, low entropy priors will depend significantly on *epistemic luck* for success (posterior epistemic value), in a sense to be made precise in §3.6.

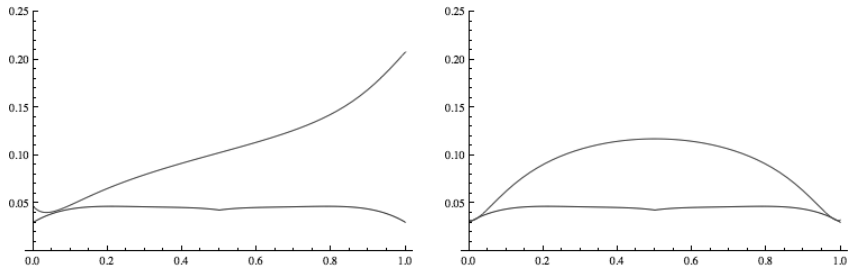


Figure 3.10: Left: $\mathscr{M}_{10}$'s objective expected disutility across hypotheses $B{=}x$, compared to the beta distribution $p$ with $\alpha = 0.8$ and $\beta = 2.4$ (entropy: -0.425). Right: $\mathscr{M}_{10}$'s objective expected disutility compared to $q$ with $\alpha = \beta = 0.5$ (entropy: -0.242).

This provides good epistemic reason to employ imprecise priors in contexts like the ones considered here, contexts in which all-things-considered epistemic value is best measured by a severely conservative disutility score. Still, such contexts are (plausibly) too exotic to be central to the foundations of imprecise Bayesianism. So we turn to another new motivation for employing imprecise priors. This second motivation provides good reason to employ imprecise priors in a much wider range of contexts than the first.

## 3.6 Epistemic Luck

Even when imprecise priors do not *dominate* precise priors, in terms of objective expected disutility, there is often good epistemic reason to adopt one, rather than a precise prior. The primary role of priors is to help you secure epistemically valuable posteriors, *and to minimize your need for epistemic luck in securing them.* In a fairly

wide range of conservative contexts, imprecise priors minimize your need for luck.

There are various kinds of epistemic luck. If there could easily be an earthquake in Los Angeles right now, but it fails to materialize, and a talented artist in the midst of a project goes on to produce her most beautiful painting to date, then her success is subject to what virtue epistemologists such as Pritchard (2009) call *environmental luck*. This is the sort of luck that enables agents to exercise skill. Without it, our artist would not have managed to paint anything at all, and so would not have been successful (produced a beautiful painting). Even so, note: certain important contrastive facts about her success are explained primarily by her *skill*, an *internal* factor, *e.g.*, the fact that her subject's eyes glisten to just the right degree, rather than slightly more, or less.

In contrast, another sort of luck — *intervening* luck — severs this explanatory link. Suppose, for example, that our artist's arch nemesis tries to sabotage her. He covers her canvas with a chemical which, when mixed with oil-based paint, produces colors at random. Fortunately for our artist, this random process happens to return each stroke, time after time, to its original color. So she is successful. Her efforts yield a beautiful painting. But she is not successful *because* skillful (her performance is not *apt*, in Sosa's terminology; *cf.* Sosa 2007, 79). Her particular degree of success (the fact that her painting is nearly perfect, rather than marred by 1, or 2, or 100, or 1000 off-colored strokes) is not explained primarily by internal factors (the agent's skill). Rather, it is explained by external factors (fortuitous chemical reactions). We will take this to be the defining characteristic of intervening luck: it is in play when external factors are primarily responsible for explaining an agent's particular *degree* of success (why she achieved exactly *this* degree of success, rather than some other).

Priors are also subject to intervening epistemic luck, in the following sense: when you update a prior on evidence, it yields a posterior which is more or less epistemically

valuable (more or less successful). The fact that the posterior is valuable to exactly *this* degree, rather than some other degree, in turn, is *explained* more or less by two different kinds of factors. On the one hand, internal factors — facts about the prior's intrinsic properties, such as how resilient it is — might bear the bulk of the explanatory burden. On the other hand, external factors — facts about the prior's extrinsic properties, such as the proximity of a coin's true bias to the prior's expected bias — might end up shouldering a bigger part of this burdern.

Of course, no prior minimizes dependence on luck *tout court*. There are various kinds of both environmental and intervening luck that adopting a prior — any prior — will simply not mitigate. No prior mitigates the environmental luck in play when the ground underneath one's lab stays intact, rather than opening up and swallowing the building whole (as it easily could have, perhaps, if the conditions were right for an earthquake). No prior helps eliminate the luck involved in stumbling upon a particularly pertinent journal article. (No prior mitigates this sort of luck in receiving new evidence.) And no prior ameliorates the luck involved in avoiding misleading evidence, of the sort that a detective faces if the primary suspect in her investigation is being framed.

When evaluating the claim, then, that imprecise priors minimize your need for luck in a range of conservative contexts, we ought to focus our attention on a particular *kind* of luck, not epistemic luck *tout court*. We ought to focus attention on whatever kind of luck a well-constructed prior could plausibly mitigate. A good candidate: luck in having the true theoretical hypothesis (*e.g.*, the coin's true bias) fall close to one's prior estimate of the true theoretical hypothesis (*e.g.*, its prior expected bias). When we talk of epistemic luck from here on out, we will have this special kind of (intervening) luck in mind.

## 3.7 Ameliorating Dependence on Luck

A prior depends on this special sort of luck — luck in having the true theoretical hypothesis fall close to its prior estimate — for success (posterior epistemic value) to the extent that facts about the proximity of that estimate to the true hypothesis are relevant for *explaining* success. It depends on this special sort of luck to the extent that such facts are relevant for explaining why the posterior is epistemically valuable to some particular degree, rather than some other degree.

To show, then, that imprecise priors do more to ameliorate dependence on luck than precise priors, in certain contexts, we need to show that the relevant proximity-facts do less to *explain* their degrees of success, in those contexts, than they do to explain the success of precise priors. As a warm-up, let's first illustrate how fortuitous proximity-facts might do less to explain the success of one precise prior than another. Take our standard example: a bookie hands you a coin and offers you a bet; you have no prior evidence about the coin's bias. Consider two options for taking your prior information (*viz.*, none) into account. You could adopt the (precise) maximum entropy (uniform) prior $u$ over hypotheses $B = x$ about the coin's bias. Alternatively, you could adopt a more concentrated beta distribution $b$ (with $\alpha = 10$ and $\beta = 4$).

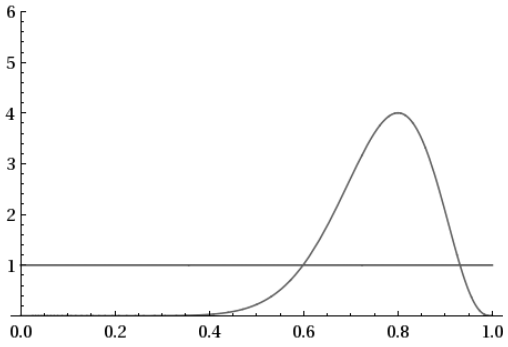


Figure 3.11: Uniform prior $u$ (bottom) and more concentrated beta prior $b$ (top).

Now imagine that you flip the coin 14 times. It comes up heads 10 times and tails 4 times. When you condition both priors on this data $D$, you arrive at the

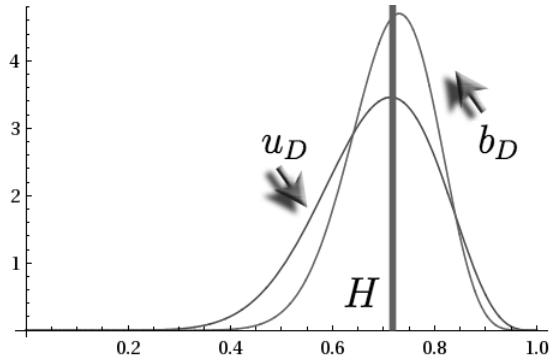


Figure 3.12: $u_D$ and $b_D$.

posteriors $u_D$ and $b_D$, respectively (pictured left, next page). Suppose that the true hypothesis $H$ about coin's bias is $B = 5/7$ (exactly the frequency of heads in your data sequence). Then $b_D$ is (determinately) more accurate, and hence (determinately) more epistemically valuable than $u_D$, on the accuracy-centered view. The former is inaccurate to degree $\mathfrak{C}(b_D, H) = 0.020$, while the latter is inaccurate to degree $\mathfrak{C}(u_D, H) = 0.028$ (measuring inaccuracy by Cramer-von Mises distance).

Though the concentrated beta prior $b$ is more successful (attains a higher degree of posterior epistemic value), the uniform prior $u$'s success depends less on luck. To see this, note: the explanation of the fact that $u_D$ has a particular degree of epistemic disutility ($\mathscr{D}(u_D, H) = \mathfrak{C}(u_D, H) = 0.028$), rather than some other degree (0.027, 0.026, etc.) is probabilistic. The most proximate explanatory factor is that, immediately prior to your experiment (flipping the coin), the true marginal chance distribution $p$ for $\mathscr{D}(u_D, H)$ had a particular character (pictured right). To explain why $u_D$ is valuable to the exact degree that it is, rather than something slightly higher or lower, we must cite not only probability mass that $p$ assigns to the hypothesis $\mathscr{D}(u_D, H) = 0.028$, but also the mass that $p$ assigns to $\mathscr{D}(u_D, H) = 0.027$,

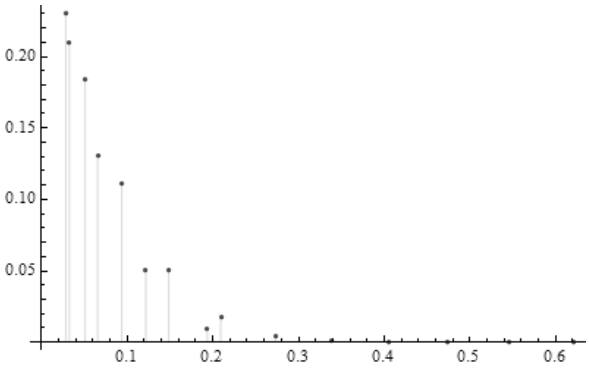


Figure 3.13: The marginal chance distribution $p$ for $\mathscr{D}(u_D, H)$, relative to the true hypothesis $H$ about the coin's bias, $B = 5/7$.

$\mathscr{D}(u_D, H) = 0.026$, etc.; the entire distribution is relevant. In addition, $p$ serves as an explanatory screen, it seems. Any other factor relevant for explaining why $u_D$ is valuable to exactly the degree that it is (0.028), rather than some other degree (0.027, 0.026, etc.), is only relevant in virtue of explaining why $p$ takes the exact form that it does.[50]

The key observation: $p$ is more or less invariant across hypotheses $H$ about the coin's bias. Whether the true bias is 5/7, 11/64 or 82/97, the marginal chance distribution $p$ for $\mathscr{D}(u_D, H)$ will look more or less the same.[51] This is reflected in the fact that $p$'s mean — $u$'s expected posterior epistemic disutility — stays fairly constant across hypotheses $H$ (see figure 13, p. 24). The upshot: the external factor in question — how close the coin's true bias happened to fall to $u$'s prior estimate — is not terribly relevant to explaining why $p$ takes the exact form that it does. Hence, it is also not terribly relevant to explaining why $u_D$ is valuable to exactly degree 0.028, rather than 0.027, 0.026, etc. The moral: $u$ depends fairly minimally on luck in having the true chances fall close to its prior estimates for success (posterior epistemic value).

To underscore this point, consider an analogy.

> *The Expert Archer.* A highly skilled archer faces a number of different targets $T$ arranged at varying distances. Given her expertise, the marginal chance distribution $p$ for $D$ (distance of her arrow from the center of the target) looks more or less the same, regardless of which target she takes aim at. Whether she aims at some target $T$ rather close by, or some $T'$ rather far away (within reasonable bounds, of course), $p$ assigns roughly the same (high) probability mass to the hypothesis $D = 0$ (hitting the target dead center), roughly the same (low) probability mass to the

hypothesis $D = 15$ (hitting 15cm off target), and so on.

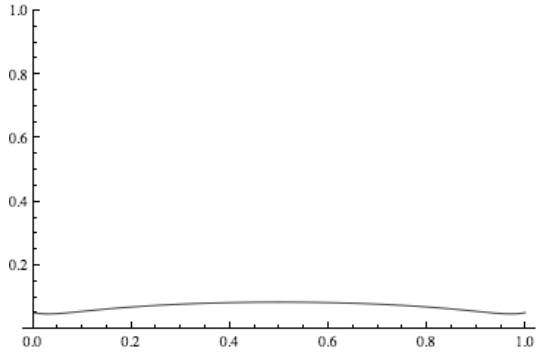


Figure 3.14: The objective expected posterior epistemic disvalue of $u$, relative to chance hypotheses $B = x$.

Because $p$ remains largely unaltered across targets $T$, the initial proximity of our archer to $T$ is plausibly (more or less) irrelevant for explaining why $p$ takes the exact form that it does. And because facts about the form that $p$ takes serve as an explanatory screen vis-á-vis $D$ — any other factor relevant for explaining why $D = 0$ (she hits the target dead center), rather than $D = 1$, $D = 2$, etc., is only relevant in virtue of explaining why $p$ takes the exact form that it does — that initial proximity is (more or less) irrelevant for explaining why our archer is successful to the exact degree that she is.

The uniform prior is much like this expert archer. Because $p$ remains largely unaltered across chance hypotheses $H$, the (initial) proximity of the uniform prior $u$ to $H$ is plausibly (more or less) irrelevant for explaining why $p$ takes the exact form that it does. And because facts about the form that $p$ takes serve as an explanatory screen vis-á-vis posterior epistemic disutility — any other factor relevant for explaining why $\mathscr{D}(u_D, H) = 0.028$, rather than $\mathscr{D}(u_D, H) = 0.027$, $\mathscr{D}(u_D, H) = 0.026$, etc., is only relevant in virtue of explaining why $p$ takes the exact form that it does — that initial proximity is next to irrelevant for explaining why the posterior $u_D$ is successful (epistemically valuable) to the exact degree that it is (0.028).

The more biased beta prior, however, is rather more like an unskilled archer. Such an archer might face targets $T$ arranged at varying distances. Suppose she aims at

a close one and hits the bullseye dead center. Unlike in the expert archer case, the marginal chance distribution $q$ for $D$ (distance of *her* arrow from the center of the target) varies significantly across $T$. If she aims at some target $T$ rather close by, the mean of $q$ (*i.e.*, the expected value of $D$) might be close to 0. There is a high chance of hitting the bullseye dead center, a lower chance of hitting 1cm off target, an even lower chance of hitting 2 cm off target, etc. But if, instead, she aims at some $T'$ far away, the mean of $q$ might be much higher. There is a much higher chance of missing the bullseye by quite a bit. The upshot: the unskilled archer's initial proximity to her target *is* relevant for explaining why $q$ takes the exact form that it does. In turn, it *is* relevant for explaining why she is successful to the exact degree that she is.

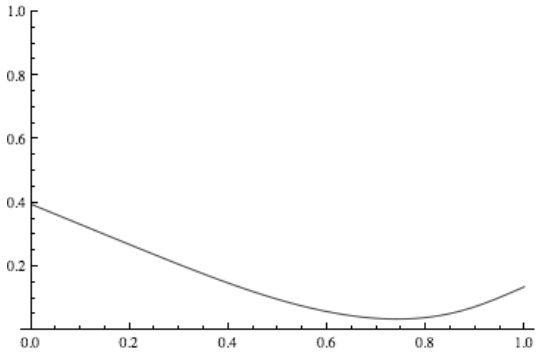


Figure 3.15: The objective expected posterior epistemic disvalue of $b$, relative to chance hypotheses $B = x$.

Similarly, the marginal chance distribution $q$ for $\mathscr{D}(b_D, H)$ varies rather significantly across chance hypotheses $H$. This is reflected in the fact that $q$'s mean — the expected posterior epistemic disutility of $b$ — varies significantly across $H$ (right, previous page). The upshot: the (initial) proximity of $b$ to the true chance hypothesis $H$ *is* relevant for explaining why $q$ takes the exact form that it does. In turn, it *is* relevant for explaining why the posterior $b_D$ is successful (epistemically valuable) to the exact degree that it is (0.020).

The moral of all of this is that certain priors (like certain archers) depend more on a special sort of luck — luck in having the truth theoretical hypothesis fall close to its prior estimate — for success (posterior epistemic value) than others. This fact, I hope to show, gives us good epistemic reason to employ imprecise priors, in a wide range

of conservative contexts. In these contexts, imprecise priors do more to ameliorate dependence on luck than precise priors do.

## 3.8 An Anti-Luck Argument for Imprecise Priors

### 3.8.1 Case 1: Imprecise Informationless Priors

In many contexts of inquiry, no precise prior puts you in an unequivocally better position to secure epistemically valuable posteriors than imprecise priors do. And vice versa. No imprecise priors puts you in an unequivocally better position to secure valuable posteriors than precise priors do. Nevertheless, there is often still good epistemic reason to prefer an imprecise prior over a precise one. The primary role of priors is to help you secure epistemically valuable posteriors, *and to minimize your need for epistemic luck in securing them.* I will show that in many contexts — conservative contexts, in particular — there are imprecise priors whose objective expected disutility varies much less across chance hypotheses than any precise prior. And this, I have argued, shows that these special imprecise priors depend less on luck for success (posterior epistemic value) than precise priors. They do more to ameliorate dependence on luck. This is the second new motivation for employing imprecise priors.

To illustrate, suppose one last time that a bookie hands you a coin and offers you a bet. You have no prior evidence about the coin's bias. The bookie is going to allow you to flip the coin 5 times prior to deciding whether or not to take the bet. Given the standards of evaluation operative in your context of inquiry, the appropriate measure of epistemic disutility $\mathscr{D}$ is Cliffordian. It is not maximally Cliffordian, however. Perhaps $\mathscr{D}(S, w) = \mathscr{D}_{0.708}(S, w) = 0.708 \cdot l(S, w) + 0.292 \cdot u(S, w)$. Such a disutility function sees *some* independent value in getting determinately close to the truth, but

nevertheless, places a premium on avoiding determinate error.

Now consider two options that you have for taking your information (*viz.*, none) into account in your decision problem. (Of course, these are not the only two options. But they do have implications for nearly all options, as we will see.) Option 1: adopt the (precise) beta prior $b$ with $\alpha = \beta \approx 1.2$ over hypotheses $B = x$ about the coin's bias. Option 2: adopt the imprecise beta-binomial model, with concentration $s = 3$.

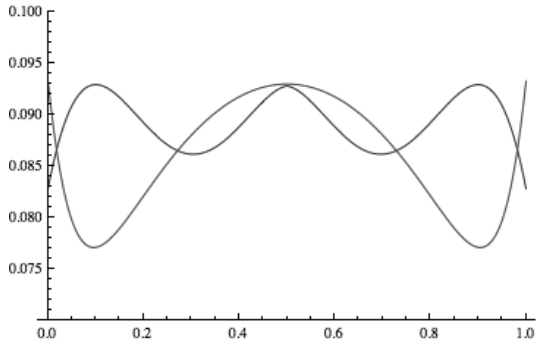


Figure 3.16: Top curve: $\mathscr{M}_3$'s objective expected disutility, relative to chance hypotheses $B{=}x$, measuring disutility by $\mathscr{D}_{0.708}$. Bottom: $b$'s objective expected disutility, relative to $B{=}x$.

In this context, I claim, $\mathscr{M}_3$ depends less on luck for success (posterior epistemic value) than $b$. The reason: $\mathscr{M}_3$'s objective expected posterior disutility varies less across chance hypotheses $B = x$ than $b$'s does (left). The difference between $b$'s maximum and minimum objective expected disutilities, $max_i\ Exp_{H_i}(\mathscr{D}_{0.708}(b_D, H_i))$ - $min_j\ Exp_{H_j}(\mathscr{D}_{0.708}(b_D, H_j))$, is 0.0162. The difference between $\mathscr{M}_3$'s maximum and minimum objective expected disutilities is less: 0.0102.

Not only does $\mathscr{M}_3$ depend less on luck for success (posterior epistemic value) than $b$, but it depends less on luck for success *than any precise beta prior*. The beta prior $b$ with $\alpha = \beta \approx 1.2$ is no arbitrarily chosen prior. It is what I call in chapter 1 the *maximally sensitive* or *MaxSen* beta prior. It depends less on luck for success than any other precise beta prior. (There is no other beta prior whose objective expected posterior disutility varies less across chance hypotheses.) So the fact that the imprecise beta model $\mathscr{M}_3$ depends less on luck for success than *it* means that $\mathscr{M}_3$ depends less on luck for success than *all* precise beta priors.

This seems to me to provide good epistemic reason to adopt an imprecise prior in conservative contexts such as the one outlined here. There is an imprecise model, *viz.*, $\mathscr{M}_3$, that is better suited to play the primary theoretical role of priors than any precise beta model. Further, given the flexibility of the class of beta distributions, one might expect $\mathscr{E}_{0.24}$ to be better suited than *any* precise distribution to play the primary role of priors. And in any context of inquiry, we ought to adopt whichever prior is best suited to play the primary theoretical role of priors in that context.

### 3.8.2 Case 2: Informative Imprecise Priors

Up to this point, we have focused on a particular class of imprecise priors, *viz.*, the imprecise beta models. Imprecise beta models are 'reference priors' or 'informationless priors', meant to be used when we lack any prior information relevant to the inference problem at hand. We have focused on this class of priors primary because (i) they form a rich, flexible class, (ii) they are mathematically tractable, and (iii) they have proved successful in a range of practical applications, *e.g.*, analyzing clinical data from randomized trials of medical treatments (*cf.* Burton *et al.* 1996).

But, in most inference problems, we come to the table with *a great deal of relevant prior information*. If the present, anti-luck motivation for employing imprecise priors is to be central to the foundations of imprecise Bayesianism, then it *must* say something about such cases. And it does. In many conservative contexts, the prior that both (i) satisfies the constraints imposed by one's prior evidence, and (ii) minimizes one's need for luck in securing epistemically valuable posteriors, is imprecise.

Imagine, for example, that a knowledgeable friend tells you that the best estimate of the coin's bias is approximately $1/2$ (perhaps she services the machine that made it, or something of the sort). Consider two options that you have for taking your evidence $E$ into account. (Again, not the only two options.) Option 1: adopt the

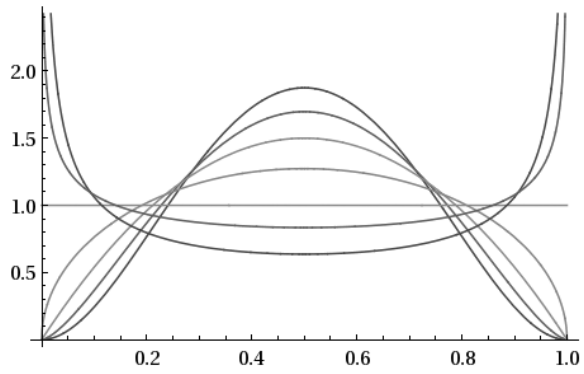


Figure 3.17: Beta distributions $p$ with entropy $H(p) \geq 0.24$ and $Exp_p(B) = 1/2$.

(precise) beta prior $b$ with $\alpha = \beta \approx 1.2$ over hypotheses $B = x$ about the coin's bias. Of course, since $\alpha = \beta \approx 1.2$, $p$ satisfies the constraint imposed by $E$, *viz.*, $Exp_b(B) = 1/2$ (as does any beta prior with $\alpha = \beta$). Option 2: adopt what we might call an *imprecise entropy model*. An imprecise entropy model is a set $\mathscr{E}_x$ of beta priors $p$ (with density functions $f$) such that (i) the entropy of $p$ is greater than or equal to $x$, *i.e.*, $H(p) = -\int_0^1 f(y) \cdot log(f(y))dy \geq x$, and (ii) $p$ satisfies the constraints imposed by $E$. (In the case at hand, $p$ satisfies the constraint imposed by $E$ just in case $\alpha = \beta = z$, for some $z$.) In particular, imagine that you adopt the imprecise entropy model with minimum entropy level $x = 0.24$, $\mathscr{E}_{0.24}$ (left).

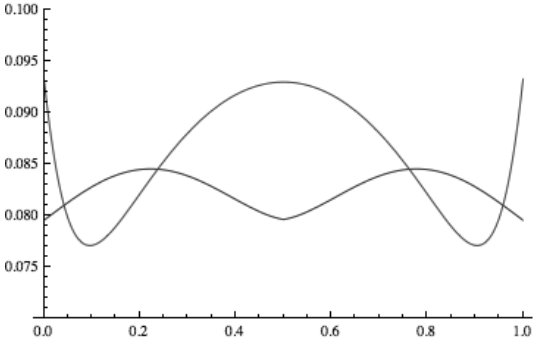


Figure 3.18: Top curve: $p$'s objective expected disutility, relative to chance hypotheses $B{=}x$, measuring disutility by $\mathscr{D}_{0.708}$. Bottom: $\mathscr{E}_{0.24}$'s objective expected disutility, relative to $B{=}x$.

Once more, the standards of evaluation operative in your context, we suppose, determine a Cliffordian measure of epistemic disutility, $\mathscr{D}_{0.708}(S, w) = 0.708 \cdot l(S, w) + 0.292 \cdot u(S, w)$. So, getting determinately close to the truth is of *some* independent value. But avoiding determinate error is much more epistemically important. In this context, I claim, $\mathscr{E}_{0.24}$ depends less on luck for success (posterior epistemic value) than $b$. The reason: $\mathscr{E}_{0.24}$'s objective ex-

pected posterior disutility varies less across chance hypotheses $B = x$ than $b$'s does (see figure 3.18). The difference between $b$'s maximum and minimum objective expected disutilities, $max_i\ Exp_{H_i}(\mathscr{D}_{0.708}(b_D, H_i))$ - $min_j\ Exp_{H_j}(\mathscr{D}_{0.708}(b_D, H_j))$, is 0.0162. The difference between $\mathscr{E}_{0.24}$'s maximum and minimum objective expected disutilities is 0.0056.

This means that $\mathscr{E}_{0.24}$ depends less on luck for success (posterior epistemic value) than any precise beta prior. The reason, again, is that beta prior $b$ with $\alpha = \beta \approx 1.2$ depends less on luck for success than any other precise beta prior. So the fact that the imprecise entropy model $\mathscr{E}_{0.24}$ depends less on luck for success than *it* means that $\mathscr{E}_{0.24}$ depends less on luck for success than *all* precise beta priors. $\mathscr{E}_{0.24}$ also depends less on luck for success than all other imprecise entropy models. There is no other imprecise entropy model whose objective expected posterior disutility varies less across chance hypotheses.

We have good epistemic reason, then, to adopt an imprecise prior in conservative contexts like this. There is an imprecise model, *viz.*, $\mathscr{E}_{0.24}$, that is better suited to play the primary theoretical role of priors than any precise beta model. Given the flexibility of the class of beta distributions, one might even expect $\mathscr{E}_{0.24}$ to be better suited than any precise distribution *tout court* to play the primary role of priors. And in any context of inquiry, we ought to adopt whichever prior is best suited to play the primary theoretical role of priors in that context.

We now have a range of illustrative examples at our disposal. We have examples of contexts in which some imprecise prior puts you in a better position to secure valuable posteriors than precise priors do (§3.4). We have examples of contexts in which some imprecise does more to ameliorate dependence on luck than precise priors do (§3.7). This is sufficient to achieve the rather limited aims of this chapter. It is sufficient to gesture toward new kinds of epistemic reasons for employing imprecise

priors.

## 3.9 Conclusion

My aim in this chapter was to illuminate two new kinds of reasons for employing imprecise priors. We ought to adopt imprecise priors in certain contexts because they *put us in an unequivocally better position to secure epistemically valuable posterior beliefs* than precise priors do. We ought to adopt imprecise priors in various other contexts because they minimize our need for *epistemic luck* in securing such posteriors. I illuminated these reasons by providing examples of the relevant sorts of contexts. This work points the way toward a new, potentially promising foundation for imprecise Bayesianism.

To recap, my main argument went as follows:

1. In any context of inquiry, you ought to adopt whichever prior is best suited to play the primary theoretical role of priors in that context, if there is one.

2. The primary role of priors is to help you secure epistemically valuable posterior beliefs, and to minimize your need for epistemic luck in securing those beliefs.

3. In certain contexts, *imprecise* priors put you in a better position to secure epistemically valuable posteriors than precise priors do.

4. In other contexts, no imprecise prior puts you in a better position to secure valuable posteriors than precise priors do, or vice versa. But *imprecise* priors minimize your need for epistemic luck.

C. In some contexts, you ought to adopt imprecise probabilities to incorporate your prior information.

I conclude by raising a few additional questions to be addressed in future research.

- I made a number of restrictive assumptions regarding the form of the epistemic disutility scores under consideration. I only considered 'linear scores', and only considered scores that measure inaccuracy by the Brier score (or Cramer-von Mises distance). How robust are our results across all reasonable epistemic disutility scores?

- The beta distributions are a very rich class of distributions. Any precise prior can be approximated by a finite mixture of beta distributions. But does this guarantee that if (i) an imprecise prior $S$'s objective expected disutility varies less than all beta priors across chance hypothesis, then (ii) $S$'s objective expected disutility varies less than all precise priors, *tout court*?

- Is it possible, for fairly general classes of evidential constraints, to specify exactly which conservative disutility scores call for imprecise priors, and to provide a tractable method for identifying which imprecise prior they call for?

# Notes

[36]See Savage 1972, pp. 46-50, for discussion of his 'washing out theorem'. See also Barron, Schervish and Wasserman (1999), or Hawthorne (1993) for discussion of conditions that guarantee convergence.

[37]Murphy (1973) shows that the Brier score, an eminently plausible measure of accuracy, decomposes into calibration and refinement components. DeGroot and Fienberg (1982, 1983) generalize this result, showing that any proper scoring rule can be separated into calibration and refinement components. See Blattenberg (1985) for additional discussion.

[38]For illuminating discussions of epistemic value, see Maher 1993, ch. 9 and Joyce 2009.

[39]See in particular Joyce 2009, §2 and §4.

[40]See Joyce 2013 for discussion of accuracy-centered approaches to theorizing about epistemic value.

[41]See Joyce 1998, p. 593; Joyce 2009, p.269; Joyce 2013, p. 3.

[42]Consider, for example, the *absolute value score*, $\mathscr{I}_{\alpha_1}$. Let $p$ be a prior defined over mutually exclusive and jointly exhaustive theoretical hypotheses $H_1, ..., H_n$. The absolute value score measures the inaccuracy of $p$ at a world $w$ by the average linear distance between $p$'s estimate of the $H_i$'s truth-value, $p(H_i)$, and $H_i$'s actual truth-value at $w$, $w(H_i)$. That is, $\mathscr{I}_{\alpha_1}(p, w) = (1/n)\sum_i |p(H_i) - w(H_i)|$. The absolute value score is ruled out as unreasonable because it violates Coherent Admissibility. To illustrate, let $p$ be the uniform distribution over mutually exclusive and jointly exhaustive hypotheses $H_1, H_2$, and $H_3$; $p(H_1) = p(H_2) = p(H_3) = 1/3$. The absolute value score of the probabilistically coherent prior $p$ at each world $w$ is $\mathscr{I}_{\alpha_1}(p, w) = (1/3)\cdot[2\cdot(1/3-0)+(1-1/3)] \approx 0.44$. In contrast, the absolute value score of the probabilistically incoherent prior $q$, which assigns 0 to each hypothesis, is 1/3 at each world $w$: $\mathscr{I}_{\alpha_1}(q, w) = (1/3) \cdot [2 \cdot (0 - 0) + (1 - 0)] \approx 0.33$. So, according to the absolute value score, there are probabilistically coherent priors which are *accuracy-dominated* by incoherent priors, *i.e.*, are at least as inaccurate at every world, and strictly more inaccurate at some worlds. This is precisely what Coherent Admissibility disallows. For a very similar, but more detailed discussion, see Joyce 2009, §9.

[43]An inaccuracy measure $\mathscr{I}$ is strictly proper just in case any probabilistically coherent prior $p$ minimizes expected inaccuracy, as measured by $\mathscr{I}$, from its own perspective, *i.e.*, $\sum_w p(w) \cdot \mathscr{I}(p, w) \leq \sum_w p(w) \cdot \mathscr{I}(q, w)$ for any other $q$.

[44]See Gneiting 2007 for discussion of the power score, spherical score, and other proper scoring rules.

[45]Specifically, if $p(w_1) = 0.1$, $p(w_2) = 0.3$, $p(w_1) = 0.6$, and $q(w_1) = 0.05$, $q(w_2) = 0.4$, $q(w_1) = 0.55$, respectively, then $\mathscr{I}(p, w_1) = 0.42$ and $\mathscr{I}(q, w_1)0.445$.

[46]An inaccuracy score $\mathscr{I}$ is concave at a world $w$ if $(1/2)\mathscr{I}(p, w) + (1/2)\mathscr{I}(q, w) \leq \mathscr{I}((1/2)p + (1/2)q, w)$, for any prior distributions $p$ and $q$.

[47]Cramer-von Mises distance also yields the correct verdict about comparative closeness in those cases where obviously correct answers are to be had. For example, for any beta densities $f$, $g$ and $h$ that have the same mean but increasing variance, $f$ is closer to $g$ than to $h$, in terms of Cramer-von Mises distance. Similarly, if $f$, $g$ and $h$ all have the same variance but larger and larger means, then $f$ is closer to $g$ than to $h$.

[48]The cumulative distribution function $F$ corresponding to a density function $f$ for a variable $V$ is defined by $F(x) = \int_{-\infty}^{x} f(y)dy$, and specifies the probability that $V$ takes a value less than or equal to $x$.

[49]$\mathscr{M}_{10}$ dominates all beta priors $b$ with differential entropy between roughly -0.25 and 0.

[50]Save, of course, for the fact that, at the end of the day, your experiment produced exactly the outcome that it did.

[51]Of course, when $B \approx 0$ or $B \approx 1$, this distribution will concentrate probability almost exclusively on one value for $\mathscr{D}(u_D, H)$.

# Appendix A

In §2.6.2, we considered the following case. You have a coin, but no information about its bias. In order to adjudicate between competing hypotheses about that bias, you plan to flip the coin 25 times. To take account of your prior information (*viz.*, none), you adopt the maximum entropy prior $u$.

We then asked: what is the expected epistemic utility of the orthodox inductive policy for making comparative probability judgments, *viz.*

$$\mathscr{I}(u, D) = \begin{cases} X \succ Y \text{ if } u_D(X) > u_D(Y) \\ X \preceq Y \text{ otherwise} \end{cases}$$

from $u$'s own perspective? We imagined that the relevant epistemic utilities are given by the following table:

Table A.1:

| | $ch(Heads) \leq ch(Tails)$ | $ch(Heads) > ch(Tails)$ |
|---|---|---|
| $Heads \preceq Tails$ | 1 | $-5$ |
| $Heads \succ Tails$ | $-5$ | 1 |
| *Abstain from judgment* | $-0.5$ | $-0.5$ |

If this is the epistemic payoff matrix, then the expected epistemic utility of $\mathscr{I}$ is $Exp_u(eu(\mathscr{I})) = 0.535057$. We then outlined an alternative inductive policy that $u$ expects to do better, *viz.*, the policy $\mathscr{I}^*$ that prescribes (i) judging that heads is more

probable than tails if $k \geq 15$, (ii) abstaining from judgment if $11 \leq k \leq 14$, and (iii) judging that tails is more probable than heads if $k \leq 10$. This policy has an expected epistemic utility of $Exp_u(eu(\mathscr{I}^*)) = 0.627876$.

A critical question: why think that any reasonable epistemic utility function yields a payoff matrix anything like this one? The answer, which I will only briefly sketch here, comes in three parts. First, the epistemic utility of a comparative probability ordering $\preceq$ is best identified with the epistemic utility of the set $S_{\preceq}$ of probabilities that represent it:

$$S_{\preceq} = \{p \mid X \preceq Y \text{ only if } p(X) \leq p(Y)\}.$$

Second, as I argued in chapter 3, there are a range of reasonable measures of the all-things-considered epistemic value or worth of a set of probabilities $S$. In particular, I argue, simple 'linear scores' of the form:

$$\mathscr{D}_\lambda(S, w) = \lambda \cdot l(S, w) + (1 - \lambda) \cdot u(S, w).$$

are *prima facie* reasonable. The quantities $l(S, w)$ and $u(S, w)$ are what I call the *lower and upper-inaccuracy scores* of $S$ at a world $w$.

- $l(S, w) = inf\{\mathscr{I}(p, w) | p \in S\}$
- $u(S, w) = sup\{\mathscr{I}(p, w) | p \in S\}$

$\mathscr{I}$ is some 'reasonable' inaccuracy function. (See, for example, Joyce 1998, 2009, Predd *et al.* 2009, and Leitgeb & Pettigrew 2010 for discussion of constraints on reasonable inaccuracy functions.)

Finally, certain linear disutility scores yield a payoff matrix like the one above. In particular, severely 'conservative scores' yield a similar payoff matrix, *e.g.*, $\mathscr{D}_{0.925}(S, w) = 0.925 \cdot l(S, w) + 0.075 \cdot u(S, w)$, where $l(S, w)$ and $u(S, w)$ are determined by the

Brier score (or in the continuous case, Cramer-von Mises distance). (See §2.3.2 for discussion of why such scores count as conservative.)

For illustrative purposes, consider an agent who judges that heads is at least as probable as tails ($\mathit{Heads} \succeq \mathit{Tails}$), and one who abstains from judgment on the matter. Identify their respective comparative probability orderings with the following sets of probabilities:

$$S_{\mathit{Abstain}} = \{b \mid \alpha + \beta \leq 10\}$$
$$S_{\mathit{Heads} \succeq \mathit{Tails}} = \{b \mid \alpha + \beta \leq 10 \,\&\, \alpha/(\alpha + \beta) \geq 1/2\}.$$

$S_{\mathit{Abstain}}$ is the beta-binomial model with concentration level 10 (see §3.4, and Walley 1991, §5.3), *i.e.*, the set of beta distributions $b$ (over hypotheses $B = x$ about the bias of the coin) with concentration $(\alpha + \beta)$ less than or equal to 10. The beta-binomial model is a popular 'imprecise ignorance prior', well-suited for modeling suspension of judgment. I use the beta-binomial model, rather than the full set of priors over hypotheses $B = x$ simply to reduce computational complexity. $S_{\mathit{Heads} \succeq \mathit{Tails}}$ is the set of beta distributions $b$ with (i) a concentration less than or equal to 10, and (ii) a mean, $\alpha/(\alpha + \beta)$, greater than or equal to $1/2$. It is the subset of $S_{\mathit{Abstain}}$ containing exactly the $b$ that assign at least as much probability to *Heads* as to *Tails*.

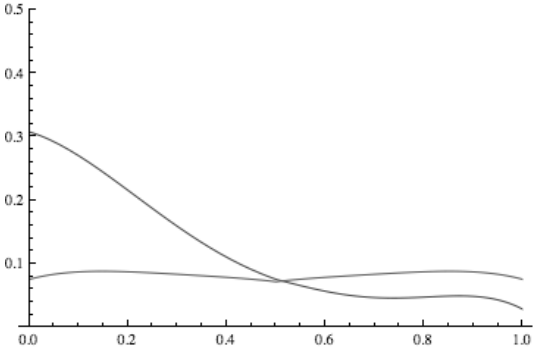


Figure A.1: Epistemic disutility of $S_{\mathit{Abstain}}$ (bottom) and $S_{\mathit{Heads} \succeq \mathit{Tails}}$ (top), respectively, as measured by the linear score $\mathscr{D}_{0.925}$.

Now consider the epistemic disutility of $S_{\mathit{Abstain}}$ and $S_{\mathit{Heads} \succeq \mathit{Tails}}$, respectively, across worlds in which $ch(\mathit{Heads}) = x$ (for all $x \in [0, 1]$), measuring disutility by

the 'conservative score' $\mathscr{D}_{0.925}(S, w) = 0.925 \cdot l(S, w) + 0.075 \cdot u(S, w)$ (pictured right, previous page). Note that $\mathscr{D}_{0.925}$ assigns exactly the same sorts of penalties that we observed in table 2.7. Judging *Heads* $\succeq$ *Tails* is slightly better, according to $\mathscr{D}_{0.925}$, than abstaining if $ch(\textit{Heads}) > ch(\textit{Tails})$. But it is *much* worse than abstaining if $ch(\textit{Heads}) \leq ch(\textit{Tails})$. The average disutilities in these two cases are as follows:

Table A.2:

| | $ch(\textit{Heads}) \leq ch(\textit{Tails})$ | $ch(\textit{Heads}) > ch(\textit{Tails})$ |
|---|---|---|
| *Heads* $\succeq$ *Tails* | 0.189 | 0.051 |
| *Abstain from judgment* | 0.083 | 0.083 |

The lesson is this: table 2.7 is simply a coarse-grained representation of the sort of epistemic disutility assignment furnished by severely conservative disutility scores, *e.g.*, $\mathscr{D}_{0.925}$. The upshot: *prima facie* reasonable epistemic utility functions yield payoff matrices like those hypothesized in table 2.7. Table 2.7 is not an ad hoc assignment of epistemic utilities.

# Bibliography


Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402.

Briggs, R., Cariani, F., Easwaran, K., and Fitelson, B. (2013). Individual coherence and group coherence. *Ms.*

Clifford, W. K. (1877). The ethics of belief. *Contemporary Review*.

Dalkey, N. (1985). Inductive inference and the representation of uncertainty. In *1st Workshop on Uncertainty in Artificial Intelligence*, pages 109–116.

de Finetti, B. (1972). *Probability Induction and Statistics*. Wiley, New York.

Deza, M. M. and Deza, E. (2009). *Encyclopedia of Distances*. Springer-Verlag, Berlin Heidelberg.

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Research*, 70(3):193–242.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368.

Fitelson, B. (2013). Coherence: Comparative confidence iii. *Ms.*

Furrer, F., Aberg, J., and Renner, R. (2011). Min- and max-entropy in infinite dimensions. *Communications in Mathematical Physics*, 306(1):165–186.

Gardenfors, P. and Sahlin, N. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese*, 53(3):361–386.

Gibbard, A. (2008). Rational credence and the value of truth. In Szabo Gendler, T. and Hawthorne, J., editors, *Oxford Studies in Epistemology*, volume 2. Oxford University Press.

Hawthorne, J. (1993). Bayesian induction is eliminative induction. *Philosophical Topics*, 21(1):99–138.

Hawthorne, J. (1994). On the nature of bayesian convergence. *Proceedings of the Biennial Meeting of the Philosophy of SCience Association*, pages 241–249.

James, W. (1897). The will to believe. In *The Will to Believe and Other Essays in Popular Philosophy*. Longmans Green and Co.

Jaynes, E. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630.

Jaynes, E. (1968). Prior probabilities. *IEEE Transactions On Systems Science and Cybernetics*, 4(3):227–241.

Jaynes, E. (1973). The well-posed problem. *Foundations of Physics*, 3:477–493.

Jaynes, E. (1976). Confidence intervals vs bayesian intervals. In Harper and Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume II. D. Reidel Publishing Company, Dordrecht-Holland.

Jeffrey, R. (1965). *The Logic of Decision*. University of Chicago Press.

Jeffrey, R. (1983). Bayesianism with a human face. *Testing Scientific Theories, Minnesota Studies in the Philosophy of Science*, 10:133–156.

Jeffrey, R. (1986). Probabilism and induction. *Topoi*, 5(1):51–58.

Jeffrey, R. (1987). Indefinite probability judgment: A reply to levi. *Philosophy of Science*, 54(4):586–591.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65(4):575–603.

Joyce, J. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, 19:153–178.

Joyce, J. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In Huber, F. and Schmidt-Petri, C., editors, *Degrees of Belief*. Springer.

Joyce, J. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, 24:281–323.

Joyce, J. (2013). Why evidentialists need not worry about the accuracy argument for probabilism. *Ms.*, pages 1–29.

Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.

Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan and Co., London.

Keynes, J. M. (1937). The general theory of employment. *The Quarterly Journal of Economic*, XIV:109–123.

Kyburg, H. and Pittarelli, M. (1996). Set-based bayesianism. *IEEE Transactions on Systems, Man, and Cybernetics*, 26:324–339.

Laplace, P. S. (1774). Memoir on the probability of the causes of events. *Mémoires de l'Académie royale des sciences de Paris*, pages 364–378.

Leitgeb, H. and Pettigrew, R. (2010a). An objective justification of bayesianism i: Measuring inaccuracy. *Philosophy of Science*, 77(2):201–235.

Leitgeb, H. and Pettigrew, R. (2010b). An objective justification of bayesianism ii: The consequences of minimizing inaccuracy. *Philosophy of Science*, 77(2):236–273.

Levi, I. (1980). *The Enterprise of Knowledge*. MIT Press, Cambridge, MA.

Machina, M. and Rothschild, M. (1990). Risk. In *The New Palgrave: Utility and Probability*.

Maher, P. (1993). *Betting on Theories*. Cambridge University Press.

Neal, R. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Orbanz, P. and Teh, Y. (2010). *Encyclopedia of Machine Learning*, chapter Bayesian Nonparametric Models. Springer.

Pearson, E. (1962). *The Foundations of Statistical Inference: A Discussion*, pages 53–58. John Wiley and Sons.

Popper, K. (1959). The propensity interpretation of probability. *Philosophy of Science*, 10(37):25–42.

Predd, J., Seiringer, R., Lieb, E., Osherson, D., Poor, H. V., , and Kulkarni, S. R. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transaction on Information Theory*, 55(10):4786–4792.

Pritchard, D. (2009). Apt performance and epistemic value. *Philosophical Studies*, 143:407–416.

Savage, L. (1972). *The Foundations of Statistics*. Dover, New York.

Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology*, 1(2):233–247.

Skyrms, B. (1977). Resiliency, propensities, and causal necessity. *The Journal of Philosophy*, 74(11):704–713.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford University Press, Oxford.

Steel, D. (2003). A bayesian way to make stopping rules matter. *Synthese*, 58:213–227.

Steele, K. (2012). Persistent experimenters, stopping rules, and statistical inference. *Erkenntnis*.

Suppes, P. (1966). A bayesian approach to the paradoxes of confirmation. In J. Hintikka, P. S., editor, *Aspects of Inductive Logic*, pages 198–207. North Holland Pub. Co.

Venn, J. (1866). *The Logic of Chance*. Macmillan, London.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):3–57.

Walley, P., Gurrin, L., and Burton, P. (1996). Analysis of clinical data using imprecise prior probabilities. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45(4):457–485.

Williamson, J. (2007). Motivating objective bayesianism: from empirical constraints to objective probabilities. In Harper, W. and Wheeler, G., editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg Jr.* College Publications, London.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford University Press.