Information Theory and Biased Beliefs

Andrew T. Little*

February 2022

Preliminary and incomplete, please do not circulate.
Note to Rochester PE Seminar readers: I will use this to give a sense of the broader research agenda I am working on, but also spend a good chunk of the talk going through the application in another working paper "Detecting Motivated reasoning", available at `https://osf.io/b8tvk`

People make a wide variety of mistakes when forming beliefs about the world. A common lament in the study of such biases is that it can resemble a laundry list of deviations from the rational benchmark without a common unifying force. There is only one way to apply Bayes rule (for a given prior belief, and likelihood function, and realization of information), but there are an infinite number of ways to not be Bayesian. And experimental evidence indicates that, in practice, many kinds of deviations seem to be common (see Tversky and Kahneman, 1974; Rabin, 1998; Benjamin, 2019, for overviews).

Here is a partial list of prominent biases. In some settings, subjects under-weight new information ("conservatism", or, if only for some signals, "confirmation bias"), under-weight prior information ("base rate neglect"), over-weight possibilities they like believing ("motivated reasoning" or "wishful thinking"), or place too much weight on salient ("availability heuristic") or focal hypotheses ("overprecision"). More generally, people may misunderstand the data generating process which produces the information they see ("misspecified models", "selection neglect", "correlation neglect").

This paper a synthetic microfoundation for these biases. The common element is that they are a solution to maximization problems using measures from information theory. These concepts have been applied to study attention costs (e.g., Sims, 2003), where beliefs converge if agents place probability zero on the truth (e.g., Esponda and Pouzo, 2016), and quantifying uncertainty and the value of information for general classes of decision problems (e.g., Frankel and Kamenica, 2019).

*Associate Professor of Political Science, UC Berkeley.

In a series of papers, Zellner (1998, 2002) suggests that that information theory can be used to microfound Bayesian updating and biased updating. The aim here is to flesh out this idea and show that it applies beyond particular kinds of biased updating.

One benefit to this organization is that it suggests a useful categorization of different types of belief biases, based on whether they are driven by (1) an incorrect prior, (2) an incorrect likelihood function, or (3) an incorrect combination of prior and likelihood.

In this note, I first provide a quick overview of the key information theory quantities which will be used, with an eye towards how they can be interpreted in the context of biased beliefs (section 1). Next, I discuss some "static" belief biases, which can be thought of as "incorrect prior" biases, though without reference to an updating problem (section 2). Finally, I discuss how this approach can be used to model biased updating (section 3).

1 Preliminaries

1.1 Entropy

One of the most widely used properties of a probability distribution in the natural sciences is *entropy*, which serves as the key building block for information theory (Shannon, 1948).

Consider a random variable ω that takes on values $\Omega = \{1, 2, ..., n\}$. Let P be the set of probability distributions on Ω, and write an individual distribution $p = (p_1, ..., p_n) \in P$, where $p_i = Pr(\omega = i)$. The entropy of distribution p is defined as:

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) = \sum_{i=1}^{n} p_i \log(1/p_i)$$

A standard interpretation of this is the average "surprise" upon learning the true value of ω, where $\log(1/p_i)$ quantifies the surprise of $\omega = i$. If p_i is very small, then $\log(1/p_i)$ will be very large. As $p_i \to 1$, the surprise upon learning $\omega = i$ approaches zero. If the expected surprise is low, we can say there is little uncertainty in a distribution. Higher entropy corresponds to more uncertainty.

A natural question to ask is what distribution(s) maximize or minimize entropy. Since entropy must be positive, and using the convention that $0 \log 0 = 0$,[1] it is immediate that any distribution that puts probability 1 on some element of Ω has the minimal entropy of 0. If we know what value the random variable takes on, there is no uncertainty.

[1] This is standard, and typically justified from the fact that $\lim_{x \to 0} x \log x = 0$.

Conversely, the solution to:

$$\tilde{p} = \max_{p' \in P} H(p') \tag{1}$$

is a uniform distribution where $\tilde{p}_i = 1/n$.

1.2 Cross Entropy

To study biased or incorrect beliefs, we will need measures which include multiple distributions: a *subjective distribution* p which may be biased with respect to an *objective distribution* q.

The entropy formula uses p_i twice, so a natural thing to do is to swap in a q_i for one of of these. Define:

$$H(q,p) = -\sum_{j=1}^{n} q_i \log(p_i).$$

Fortunately this is also a standard information theory object: the cross entropy of p relative to q. One way to think about cross-entropy is "if the real distribution is q but someone thinks the distribution is p, how surprised will they be by the revelation of the truth (on average)?" If q is an objective probability distribution, then a p which has a higher cross entropy can be thought of as a a "worse" subjective belief.

By this standard, we can make a subjective belief p arbitrarily bad by putting very small weight on some p_i such that $q_i > 0$. This will leave open the possibility of being "infinitely surprised."

What is the best subjective belief to hold for a fixed objective belief, in terms of minimizing $H(q,p)$? To cast this as a maximization problem, this will be:

$$\tilde{p} = \max_{p \in P} -H(p,q) \tag{2}$$

This is an optimization problem with the constraint that $\sum_{i=1} p_i = 1$. The Lagrangian can be written:

$$L(p) = \sum_{i=1}^{n} q_i \log(p_i) + \lambda \quad 1 - \sum_{i=1}^{n} p_i$$

and so the FOC for each p_i is:

$$\frac{\partial L}{\partial p_i} = q_i/p_i - \lambda = 0 \leftrightarrow p_i = q_i/\lambda$$

From the constraint that $\sum p_i = 1$ it immediately follows that the optimal subjective belief is $\tilde{p}_i = q_i$, i.e., the objective belief.

An interpretation of this result is that surprise is minimized when holding a correct belief about the likelihood of different possibilities. Conversely, one will tend to be more surprised when holding "more incorrect" beliefs.

We may also be interested in the cross-entropy of q relative to p:

$$H(p, q) = -\sum_{j=1}^{n} p_i \log(q_i)$$

$H(p, q)$ can be interpreted as "if someone subjectively believes that the distribution is p while also keeping in the back of their mind that the true distribution is q, how surprised do they expect to be?" This may seem a bit less natural at first but will prove useful later.

Treating q as fixed and maximizing/minimizing with respect to p, this cross entropy is linear in each p_i. As a result, if there is a unique $\omega^* = \arg\max_i q_i$, the solution to:

$$\tilde{p} = \max_{p' \in P} -H(p, q) \tag{3}$$

is $\tilde{p}_i = \mathbf{1}_{i=\omega^*}$. (If there are multiple states which maximize the objective likelihood, any distribution which places all probability on these is equally good at minimizing cross-entropy).

1.3 Kullback-Leibler Divergence

A key quantity for all that follows builds on the idea of entropy and cross entropy. The *Kullback-Leibler (KL) Divergence* is just the difference between the two. As with entropy, we can write this two ways depending on what we take as the "reference distribution", or what we average over. One version takes the objective belief as the reference distribution. This is often called the KL divergence *from* the subjective belief p *to* the objective belief q:

$$D_{\text{KL}}(q \| p) = H(q, p) - H(q) = \sum_{i=1}^{n} q_i \log(q_i / p_i)$$

Recall we can think of $H(q)$ as the average surprise with the correct belief, and $H(q, p)$ as the average surprise if expecting p when the real distribution is q. So, this KL divergence can be read as "if the real distribution is q, how much more surprise would be experienced by a subject with (potentially incorrect) belief p."

As with cross-entropy, this can be made arbitrarily large by picking a p which puts a vanish-

ingly small weight on a p_i that happens with positive probability. A good belief may be one that minimizes KL divergence, or maximizes:

$$\tilde{p} = \max_{p \in P} -D_{\text{KL}}(q\|p) = H(q) - H(q,p) \tag{4}$$

Since the $H(q)$ terms drop out when maximizing with respect to p, this is the same as minimizing cross-entropy, and $\tilde{p} = q$.

A nice feature of thinking of the optimal belief as minimizing KL divergence is that, at the optimal belief, $D_{\text{KL}}(\tilde{p}\|q) = 0$. That is, KL divergence measures something like the "distance between p and q", though importantly it is not a distance metric as it is not symmetric.

Unlike cross entropy, if we flip the arguments and treat the subjective belief as the reference distribution, we get the same solution when maximizing $-D_{\text{KL}}(p\|q)$. Formally, if we maximize:

$$\tilde{p} = \max_{p \in P} -D_{\text{KL}}(p\|q), \tag{5}$$

the Langrangian is:

$$L(p) = -\sum_{i=1}^{n} p_i \log(p_i/q_i) + \lambda \quad 1 - \sum_{j=1}^{n} p_i \Bigg) \Bigg($$

And so the FOC on p_i are:

$$-1 - \log(p_i) + \log(q_i) - \lambda = 0 \leftrightarrow p_i = q_i e^{-1-\lambda}$$

and the constraint that the p_i's sum to 1 ensures that $\tilde{p}_i = q_i$.

A standard way to think about $D_{\text{KL}}(q\|p)$ is as the "information gain" going from a belief of p to q, if q is in fact the truth. In some applications the reference distribution (first argument) is a "real" distribution and the second argument is a model or an approximation of the real distribution. So it may be tempting to use this as a way to measure how "bad" the subjective belief is.

However, if the subjects we are studying are going to make decisions using their subjective belief, then it is the real distribution to them. As a result, we will use $D_{\text{KL}}(p\|q)$ rather that $D_{\text{KL}}(q\|p)$ to measure "how close is the subjective belief to the objective belief." We can then interpret $D_{\text{KL}}(p\|q)$ as the "false information gain" going from the objective belief to the subjective belief, if one sincerely believes their subjective belief.

To give a bit more justification for this choice, in many applications using KL divergence, the

reference distribution is the choice which is maximized/minimized.[2] It is also generally the "after" distribution when looking at a change in beliefs, which in this context is going from objective (q) to subjective (p) belief.

Finally, a cheap answer is that it will generally lead to much nicer math as we add more components to our optimization problems. As a stark example, if using $D_{\text{KL}}(q\|p)$, it is impossible to capture the idea that the subjective belief may "not consider" or place probability 0 event which is objectively possible, since for any $q_i > 0$, $\lim_{p_i \to 0} D_{\text{KL}}(q\|p) = \infty$, but $\lim_{p_i \to 0} D_{\text{KL}}(p\|q)$ is finite.

2 Static Belief Biases

Now that we have a benchmark of what constitutes the ideal belief via maximization problems, we can start to use related problems to capture biased beliefs.

Many if not most belief distortions relate to how new information is processed. We will get to that in section 3. First, we will study "static" biases that distort beliefs without reference to any new information to incorporate.

The general approach to this will be setting up maximization problems where one goal is minimizing KL divergence, and there are either other goals or constraints on the subjective belief.

2.1 Motivated Beliefs

One of the most widely studied biases in the social sciences is *motivated reasoning* (e.g., Kunda, 1990; Epley and Gilovich, 2016; Bénabou and Tirole, 2016). In an influential formulation, Kunda (1990) emphasizes that all reasoning is motivated, and the key is to distinguish between *accuracy* motives and *directional* motives. As discussed above, negative KL divergence is a natural way to capture the accuracy motive, as it generates a penalty for moving further from the objective distribution. (Or, somewhat more precisely, penalizes "false subjective information gain.")

Following Little (2021), we can modify the maximization problem by adding a term which captures the directional motive:

$$\tilde{p} = \max_{p' \in P} -D_{\text{KL}}(p\|q) + \sum_{j=1}^{n} v_i p_i \tag{6}$$

where v_i represents how much the subject likes believing the state is equal to i (see also Bracha and Brown, 2012).

[2] Need some cites here. Jaynes? Kullback?

Theorem 1 in Little (2021) states that this has a unique solution given by:

$$\tilde{p}_i = \frac{q_i e^{v_i}}{\sum_{j=1}^{n} q_j e^{v_j}}$$

That is, each possibility gets weighted by an increasing function of how pleasant the subject finds the possibility (e^{v_i}), and then normalized to remain a proper probability distribution.[3]

While there are challenges to detecting motivated reasoning of this form (Little, 2021), some research designs show that manipulating directional motives (the v_i's) can lead to different reported beliefs, even with financial incentives for accuracy (see Epley and Gilovich, 2016, for an overview). For example, one strand of this literature documents self-serving biases by exploring random assignment into roles which lead to different desired beliefs, in the context of pretrial bargaining (Babcock et al., 1995; Babcock and Loewenstein, 1997) or debate tournaments Schwardmann, Tripodi and Van der Weele (2021).

2.2 Overestimation, wishful thinking, anchoring

While motivated by motivated beliefs, this formalization can also capture several other common biases. If each state of the world is associated with a utility u_i, and $v_i = g(u_i)$ for some increasing function g, then we can think of this as capturing "wishful thinking"; see Bracha and Brown (2012) and Mayraz (2019) for related models and discussion of empirical evidence.

Another one of the most widely studied biases is overconfidence (e.g., Ortoleva and Snowberg, 2015). One variety of overconfidence is overestimation, where people think they are better than they really are in some way. If ω is a measure of one's ability, and if v is increasing, then the subject will exhibit overestimation in the sense that their subjective belief monotone likelihood ratio dominates the objective belief.

Yet another heavily studied bias which can be captured with this formalization is anchoring (Tversky and Kahneman, 1974; Epley and Gilovich, 2006): if an experimental treatment gives an anchor of i^*, and v_i is increasing up to i^* and then decreasing, then more weight will be placed on values close to i^*.

2.3 Overprecision, version 1

Another variety of overconfidence, which is arguably the hardest to correct, is overprecision (Moore, Tenney and Haran, 2015). This calls for a different formalization, since it is not about

[3] Mayraz (2019) provides an alternative microfoundation for "wishful thinking" which leads to essentially the same result.

liking certain states, but about wanting to have more "certain" beliefs in general. One natural way to capture this idea is by adding a term which penalizes beliefs with higher entropy. Formally, suppose

$$\tilde{p} = \max_{p' \in P} -D_{\text{KL}}(p\|q) - aH(p) \tag{7}$$

Setting up the Lagrangian as before gives:

$$L(p) = -\sum_{j=1}^{n} p_j \log(p_j/q_j) + a\sum_{j=1}^{n} p_j \log(p_j) + \lambda \quad 1 - \sum_{j=1}^{n} p_j \tag{8}$$

Setting the derivative of the Lagrangian with respect to p'_j equal to zero gives:

$$0 = \frac{\partial L}{\partial p'_j} = -1 - \log(p_j) + \log(q_j) + a\log(p_j) + a - \lambda \tag{9}$$

As long as $a < 1$, the objective function is concave in each p_i, and is solved by:

$$\tilde{p}_i = \frac{q_i^{1/(1-a)}}{\sum_{j=1}^{n} q_i^{1/(1-a)}}$$

If $a = 0$, this nests the objective belief. As $a \to 1$, the exponent approaches infinity, which places all weight on whichever belief(s) are most likely. (If $a > 1$ the objective function is convex in each p_i, and the optimal subjective belief places probability 1 on the most likely state.)

One nice way to see how the a parameter affects beliefs is to consider the ratio of the subjective beliefs about two states i and j:

$$\frac{\tilde{p}_i}{\tilde{p}_j} = \left(\frac{q_i}{q_j}\right)^{1/(1-a)}$$

If $q_i > q_j$, then as a increases this ratio increases. I.e., when comparing two choices, overprecision via entropy aversion increases the relative probability assigned to the *ex ante* more likely choice.

An immediate implication of this is that for $a \in (0, 1)$, the subjective belief will put more probability on the most likely ("focal") possibility than the objective belief, consistent with a range of evidence (e.g., Moore, Tenney and Haran, 2015).

2.4 Ignorance priors and partition dependence

Other studies find evidence of a bias towards an "ignorance prior" (Fox and Clemen, 2005), which is effectively the opposite of overprecision as modeled here.[4] While a motive for more precision pushes towards a reduction of entropy in the distribution, we can model movement towards ignorance as injecting *more* entropy in the distribution. Formally, this can be captured with the same maximization problem as in (7) but with $a < 0$.

This will tend to decrease the probability assigned to the most likely options and increase the probability assigned to the least likely options, consistent with "favorite long-shot bias" (Snowberg and Wolfers, 2010). The idea of "adding entropy" to a subjective belief will also prove important when trying to match empirical results about belief updating.

Any $a \neq 0$ also introduces *partition dependence*, where lumping together (or dividing) events can change the probabilities assigned to other possibilities (Tversky and Koehler, 1994). To illustrate, consider what happens if the possible outcomes are divided into $\omega = 1$ or $\omega > 1$, i.e., all other possibilities. Let P' be the set of probability distributions on $\Omega = \{1, > 1\}$. By the same logic, for an entropy adjustment a, the subjective belief on $\omega = 1$ will be:

$$\tilde{p}'_1 = \frac{q_1^{1/(1-a)}}{q_1^{1/(1-a)} + \left(\sum_{j=2}^{n} q_j\right)^{1/(1-a)}}$$

this will lead to a strictly higher subjective estimate of $\omega = 1$ (and a lower estimate of $\omega > 1$) if and only if:

$$q_1^{1/(1-a)} + \left(\sum_{j=2}^{n} q_j\right)^{1/(1-a)} < q_1^{1/(1-a)} + \sum_{j=2}^{n} q_j^{1/(1-a)}$$

which holds $a < 0$. Put another way, if $a < 0$, then events are subadditive, meaning that when they are lumped together the total subjective probability is less than when they are "unpacked" Tversky and Koehler (1994).[5]

[4] See also Clemen and Ulu (2008).

[5] To map this to the notion of a "support function" developed by Tversky and Koehler (1994), for any $E \subseteq \Omega$ let:

$$s(E) = \left(\sum_{j \in E} p_j\right)^{1/(1-a)}$$

If $a \leq 0$, then events are subadditive, in the sense that for any E^1 and E^2 which are mutually exclusive and exhaustive subsets of E ($E^1 \cap E^2$ and $E^1 \cup E^2$), $s(E) \leq s(E^1) + s(E^2)$. If $a < 0$ then this inequality is strict. As discussed

A range of studies find evidnece consistent with this prediction (Tversky and Koehler, 1994; Fox and Clemen, 2005; Benjamin, 2019).

2.5 Overprecision, version 2

Given the fact that biased beliefs often seem consistent with entropy injection rather than entropy reduction (and we will see even more evidence of this when studying updating), we may want another way to capture overprecision. One mechanism is that it is often hard to think through all possibilities (Moore, 2022). If one does not even know all the possible values that ω can take on, this will tend to lead to overprecision relative to what the subject would believe if they could "think through" all possibilities.

Backus, Little and Moore (2021) capture this idea by assuming subjects only think through a subset $T \subseteq \Omega$, subject to the constraint that the ratios of subjective probabilities match the ratios of objective probabilities. We can arrive a the same formula they do by assuming subjects minimize the KL divergence from the objective belief to their subjective belief, subject to the constraint that the subjective distribution only places positive probability on elements in T. Formally, let P^T be the set of probability distributions such that $\sum_{j \in T} p_j = 1$. The subjective belief which solves:

$$\tilde{p} = \max_{p' \in P^T} -D_{\text{KL}}(p \| q) \tag{10}$$

is:

$$\tilde{p}_j = \frac{q_j}{\sum_{j \in T} q_j}$$

See López-Pérez, Rodriguez-Moral and Vorsatz (2021) and Moore (2022) and for experimental evidence consistent with this expression of overprecision, and Backus, Little and Moore (2021) for a discussion of other empirical results it can explain.

3 Biased Updating

Now let's consider how subjects respond to new information.

To capture biased updating with information theory maximization problems, we need to first show that the standard benchmark—Bayes' rule—can arise from this approach.

in Benjamin, Moore and Rabin (2017) and Benjamin (2019), this is related to the fact that $x^{1/(1-a)}$ is concave in x if and only if $a < 0$.

3.1 Bayes' Rule

Suppose a subject starts with an objective prior belief q, and then observes a signal s with likelihood function $f(s|\omega)$.

A simple trick to derive Bayes' rule with an information theory maximization problem is to see a parallel between the formula for motivated beliefs derived above – where the objective probability is multiplied by e^{v_i} and then normalized – and Bayes' rule, where the prior is multiplied by the likelihood and then normalized. So, motivated beliefs are like a Bayesian update from the objective prior where $e^{v_i} = f(s|\omega = i)$.

Rearranging, we can take the maximization problem for motivated beliefs and replace v_i with $\log(f(s|\omega = i))$:

$$\tilde{p} = \max_{p' \in P} -D_{\text{KL}}(p\|q) + \sum_{j=1}^{n} p_j \log(f(s|\omega = j)) \tag{11}$$

which is solved by:

$$\tilde{p}_i = \frac{p_i f(s|\omega = i)}{\sum_{j=1}^{n} p_j f(s|\omega = j)},$$

equivalent to the standard Bayesian update.

Zellner (1988) uses an equivalent entropy minimization problem to derive Bayes' rule. In doing so he describes the $\sum_{j=1}^{n} p_i \log(f(s|\omega = j))$ term as the "information" of the likelihood function, which sounds like something we would like to maximize. It is standard to describe the negative of entropy as information, and this is mathematically the "cross-entropy of the likelihood function relative to the subjective belief p," though this is a bit odd since the likelihood function is not a probability distribution.[6] So I'm a bit confused as to how exactly this should be motivated.[7] (And not alone, in commenting on Zellner (1988), Bernardo (1988) writes "I fail to appreciate why [this equation] should be a good measure of information (information about what)?")

3.2 Conservatism

Now that we have derived Bayes' rule in this fashion, it is natural to ask what happens if we solve more general versions of this maximization problem (Zellner, 2002). A natural variant is to

[6] Can we say this measures "once we learn the truth about ω, how surprised should we be that s turned out to be what we observed"?

[7] Ebrahimi, Soofi and Soyer (2010) has a discussion of this which looks useful.

place weights on the prior and likelihood function:

$$\tilde{p} = \max_{p \in P} -bD_{\text{KL}}(p\|q) + c\sum_{j=1}^{n} p_i \log(f(s|\omega = j)). \tag{12}$$

This is equivalent to maximizing

$$-D_{\text{KL}}(p\|q) + \sum_{j=1}^{n} p_i \log(f(s|\omega = j)^{c/b})$$

which by the same proof as for motivated beliefs gives:

$$\tilde{p}_i = \frac{q_i f(s|\omega = i)^{c/b}}{\sum_{j=1}^{n} q_j f(s|\omega = j)^{c/b}}$$

If $c < b$, i.e., "less weight" is put on the likelihood function, this will weaken the influence of the signal on the posterior belief. To see this, consider the log likelihood ratio of two possibilities i and j:

$$\log(\tilde{p}_i/\tilde{p}_j) = \log(q_i/q_j) + c/b \log(f(s|\omega = i)/f(s|\omega = j))$$

If $c = b$, this corresponds to the the standard Bayesian posterior ratio which puts equal weight on the log prior ratio and the log likelihood ratio. If $c < b$, less weight is placed on the likelihood.

For binary random variables, this is the logit transformation, and this is a common regression used to estimate whether new information is incorporated properly (see section 4 of Benjamin, 2019, for an overview). A common feature of this work is estimating coefficients on the log-likelihood ratio of less than one ($c/b < 1$), which is typically called "conservatism."

3.3 Confirmation Bias

One way to think about confirmation bias is putting less weight on signals which go against one's prior belief, at least on average (Rabin and Schrag, 1999). See also Little (2021, section 4) for a model where signals may be ignored/downweighted if they lead to less pleasant beliefs, consistent with the evidence that it is this motivation rather than confirming the prior *per se* that drives asymmetric incorporation of information (e.g., Tappin, van der Leer and McKay, 2017).

3.4 Base Rate Neglect

Another common feature in experiments is "base rate neglect", or not fully incorporating prior information. It is tempting to think that we can get at this by putting less weight on the KL divergence term, which in a sense makes it easier to move away from the prior. However, recall we already put a weight b on the prior, which only affected the relative weight put on the new information.

To see how we can downweight the prior, recall that KL divergence is the difference between cross entropy and entropy, and hence we can write the "standard" Bayes' rule maximand as:

$$-(H(p,q) - H(p)) + \sum_{j=1}^{n} p_i \log(f(s|\omega = j))$$

So, the b term before weights both the cross entropy term ($H(p,q)$) and the entropy term ($H(p)$). If we allow these to have separate weights the resulting maximization problem can be written as the previous but with an "entropy adjustment" term, which as in (7) we weight with an a:

$$\tilde{p} = \max_{p \in P} -bD_{\text{KL}}(p\|q) + c\sum_{j=1}^{n} p_i \log(f(s|\omega = j)) - aH(p) \tag{13}$$

This is maximized at:

$$\tilde{p}_i = \frac{q_i^{b/(b-a)} f(s|\omega = i)^{c/(b-a)}}{\sum_{j=1}^{n} q_j^{b/(b-a)} f(s|\omega = j)^{c/(b-a)}}$$

If $a = 0$ this is the same as above. If $a < 0$–meaning there is entropy "added" in the maximand—then the subject exhibits base rate neglect as the exponent on q_j is less than 1.

Again, it is helpful to look at the log odds ratio:

$$\log(\tilde{p}_i/\tilde{p}_j) = \frac{b}{b-a}\log\left(\frac{q_i}{q_j}\right) + \frac{c}{b-a}\log\left(\frac{f(s|\omega = i)}{f(s|\omega = j}\right)$$

If $a = 0$ and $b = c$, both coefficients are 1, corresponding to the Bayesian update. If $a < 0$ the coefficient on the objective log odds ratio is below 1, corresponding to base rate neglect. And if $b - a < c$ there will be conservatism. If $a < 0$ it is possible for the subject to underweight both the prior and new information, again consistent with a wide range of evidence reviewed by Benjamin (2019).

In the context of updating we can think of "adding entropy" as the subjects being confused about exactly how to process information, and tending towards relatively uniform distributions as

a result.

3.5 Wrong likelihood

Another broad class of biases can be captured by assuming the subject maximizes with the wrong likelihood function. Let $\tilde{f}(s|\omega)$ be a subjective likelihood function. If the subject maximimizes $\tilde{f}$

$$\tilde{p} = \max_{p' \in P} -D_{\text{KL}}(p\|q) + \sum_{j=1}^{n} \log(\tilde{f}(s|\omega = j))p_i \tag{14}$$

then the subjective posterior will be:

$$\tilde{p}_i = \frac{q_i \tilde{f}(s|\omega = i)}{\sum_{j=1}^{n} q_j \tilde{f}(s|\omega = h)}$$

While it is not particularly generative, this can capture a variety of prominent biases such as correlation neglect (Levy and Razin, 2015; Ortoleva and Snowberg, 2015), not accounting for the (strategic) selection problem in the information observed (Eyster and Rabin, 2005; Jehiel, 2018; Enke, 2020), or more generally updating based on a misspecified model of how the world works (e.g., Esponda and Pouzo, 2016).

4 Discussion

Things to elaborate on:

- What other biases/empirical patterns can be captured here?
- Compare long-run convergence to the truth with different kinds of biases. I suspect that under some circumstances "wrong prior" and maybe even "wrong weights" biases don't prevent convergence to the truth, but "wrong likelihood" biases clearly can.
- Connections to the wider "cross-entropy minimization" paradigm.

References

Babcock, Linda and George Loewenstein. 1997. "Explaining bargaining impasse: The role of self-serving biases." *Journal of Economic perspectives* 11(1):109–126.

Babcock, Linda, George Loewenstein, Samuel Issacharoff and Colin Camerer. 1995. "Biased judgments of fairness in bargaining." *The American Economic Review* 85(5):1337–1343.

Backus, Matthew, Andrew T Little and Don A Moore. 2021. "Constraints on Thinking Cause Overprecision.".
URL: *psyarxiv.com/evcx2*

Bénabou, Roland and Jean Tirole. 2016. "Mindful economics: The production, consumption, and value of beliefs." *Journal of Economic Perspectives* 30(3):141–64.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics: Applications and Foundations 1* 2:69–186.

Benjamin, Daniel J, Don A Moore and Matthew Rabin. 2017. Biased beliefs about random samples: Evidence from two integrated experiments. Technical report National Bureau of Economic Research.

Bernardo, José M. 1988. "Optimal information processing and Bayes's theorem; comment." *American Statistician* 42:282–282.

Bracha, Anat and Donald J Brown. 2012. "Affective decision making: A theory of optimism bias." *Games and Economic Behavior* 75(1):67–80.

Clemen, Robert T and Canan Ulu. 2008. "Interior additivity and subjective probability assessment of continuous variables." *Management Science* 54(4):835–851.

Ebrahimi, Nader, Ehsan S Soofi and Refik Soyer. 2010. "Information measures in perspective." *International Statistical Review* 78(3):383–412.

Enke, Benjamin. 2020. "What You See Is All There Is." *Quarterly Journal of Economics* 135(3):1363–1398.

Epley, Nicholas and Thomas Gilovich. 2006. "The anchoring-and-adjustment heuristic: Why the adjustments are insufficient." *Psychological science* 17(4):311–318.

Epley, Nicholas and Thomas Gilovich. 2016. "The Mechanics of Motivated Reasoning." *Journal of Economic Perspectives* 30(3):133–40.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.30.3.133*

Esponda, Ignacio and Demian Pouzo. 2016. "Berk–Nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica* 84(3):1093–1130.

Eyster, Erik and Matthew Rabin. 2005. "Cursed equilibrium." *Econometrica* 73(5):1623–1672.

Fox, Craig R. and Robert T. Clemen. 2005. "Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior." *Management Science* 51(9):1417–1432.
URL: *https://doi.org/10.1287/mnsc.1050.0409*

Frankel, Alexander and Emir Kamenica. 2019. "Quantifying information and uncertainty." *American Economic Review* 109(10):3650–80.

Jehiel, Philippe. 2018. "Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism." *American Economic Review* 108(6):1582–1597.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108(3):480.

Levy, Gilat and Ronny Razin. 2015. "Correlation neglect, voting behavior, and information aggregation." *American Economic Review* 105(4):1634–45.

Little, Andrew T. 2021. "Detecting Motivated Reasoning.".
URL: *osf.io/b8tvk*

López-Pérez, Raúl, Antonio Rodriguez-Moral and Marc Vorsatz. 2021. "Simplified mental representations as a cause of overprecision." *Journal of Behavioral and Experimental Economics* 92:101681.

Mayraz, Guy. 2019. "Priors and Desires: A Bayesian Model of Wishful Thinking and Cognitive Dissonance.". Manuscript. Available at http://mayraz.com/papers/PriorsAndDesires.pdf.

Moore, Don A. 2022. "Overprecision is a property of thinking systems.".
URL: *https://osf.io/kn8as/*

Moore, Don A, Elizabeth R Tenney and Uriel Haran. 2015. "Overprecision in judgment." *The Wiley Blackwell handbook of judgment and decision making* 2:182–209.

Ortoleva, Pietro and Erik Snowberg. 2015. "Overconfidence in political behavior." *American Economic Review* 105(2):504–535.

Rabin, Matthew. 1998. "Psychology and economics." *Journal of economic literature* 36(1):11–46.

Rabin, Matthew and Joel L Schrag. 1999. "First impressions matter: A model of confirmatory bias." *The quarterly journal of economics* 114(1):37–82.

Schwardmann, Peter, Egon Tripodi and Joël J Van der Weele. 2021. "Self-persuasion: Evidence from field experiments at two international debating competitions.".

Shannon, Claude Elwood. 1948. "A mathematical theory of communication." *The Bell system technical journal* 27(3):379–423.

Sims, Christopher A. 2003. "Implications of rational inattention." *Journal of monetary Economics* 50(3):665–690.

Snowberg, Erik and Justin Wolfers. 2010. "Explaining the favorite–long shot bias: Is it risk-love or misperceptions?" *Journal of Political Economy* 118(4):723–746.

Tappin, Ben M, Leslie van der Leer and Ryan T McKay. 2017. "The heart trumps the head: Desirability bias in political belief revision." *Journal of Experimental Psychology: General* 146(8):1143.

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

Tversky, Amos and Derek J Koehler. 1994. "Support theory: a nonextensional representation of subjective probability." *Psychological review* 101(4):547.

Zellner, Arnold. 1988. "Optimal information processing and Bayes's theorem." *The American Statistician* 42(4):278–280.

Zellner, Arnold. 2002. "Information processing and Bayesian analysis." *Journal of Econometrics* 107(1-2):41–50.