Modelling Biased Judgement with Weighted Updating

Zinn, Jesse

University of California, Santa Barbara

30 September 2013

Modelling Biased Judgement with Weighted Updating

Jesse Aaron Zinn*

University of California Santa Barbara

September 30, 2013

Abstract

The weighted updating model is a generalization of Bayesian updating that allows for biased beliefs by weighting the functions that constitute Bayes' rule with real exponents. I provide an axiomatic basis for this framework and show that weighting a distribution affects the information entropy of the resulting distribution. This result provides the interpretation that weighted updating models biases in which individuals mistake the information content of data. I augment the base model in two ways, allowing it to account for additional biases. The first augmentation allows for discrimination between data. The second allows the weights to vary over time. I also find a set of sufficient conditions for the uniqueness of parameter estimation through maximum likelihood, with log-concavity playing a key role. An application shows that self attribution bias can lead to optimism bias.

JEL Codes: C02, D03
Keywords: Bayesian Updating, Cognative Biases, Learning, Uncertainty

*I am grateful for valuable comments and suggestions from Ted Bergstrom, Javier Birchenall, Gary Charness, Zack Grossman, Jason Lepore, Dick Startz, and seminar discussants at Cal Poly San Luis Obispo and at the 2013 conference of the Society for the Avdancement of Behavior Economics.

1 Introduction

The last several decades have witnessed the accumulation of overwhelming evidence suggesting that we do not have rational expectations, nor do we always form beliefs rationally, according to Bayes' rule. Rather, we consistently and systematically exhibit a number of biases that tend to distort our perception of reality.[1] This paper presents an axiomatic development and analysis of the weighted updating model, which generalizes Bayesian updating by exponentially weighting the likelihood function(s) and prior distribution to allow for biased belief formation.

The weighted updating model has seen some use in the economics literature.[2] Grether (1980) and Grether (1992) provide empirical evidence for the representativeness heuristic by estimating the weights on the likelihood function and the prior distribution. In recent theoretical work, Benjamin, Rabin, and Raymond (2011) model "non-belief in the law of large numbers" using the weighted updating model. Palfrey and Wang (2012) use weighted updating to model investors who under- or overreact to public information regarding financial assets in a model with speculative pricing.

I expand the weighted updating literature in several ways. I strive for generality throughout, providing results that are clearly applicable to a variety of models. A major part of this involves studying general distributions, rather than particular families of distributions. This contrasts with the previous literature on weighted

[1] See Rabin (1996) and DellaVigna (2009) for surveys of the literature at the intersection of psychology and economics, including detailed discussion of many belief perturbing biases.

[2] There has been some work that utilizes models similar to weighted updating outside of economics. Ibrahim and Chen (2000) introduced *power priors*, a framework that allows the statistician to consider data from previous studies by finding a weight in $(0, 1)$ to put on that data while maintaining a weight of 1 on current data. This can be viewed as a case of weighted updating wherein the statistician rationally discriminates between different batches of data. In the logic literature, Van Benthem, Gerbrandy, and Kooi (2009) define a "weighted product updating rule" and show that Bayes' rule and the Jeffrey updating rule are both special cases.

updating, as Grether (1980), Grether (1992), Benjamin, Rabin, and Raymond (2011), and Palfrey and Wang (2012) each focus exclusively on Bernoulli random variables, thereby limiting their analyses to distributions from the beta-binomial family.

Section 2 presents a set of three axioms that apply to pairs of beliefs that differ only by how strong they are. In that section, I also show that the information entropy of a distribution that models such beliefs represents the ordering necessitated by these axioms. Section 3 establishes that weighting a distribution and then normalizing it results in a distribution that, along with the original distribution, satisfies the axioms from Section 2. It also presents additional axioms that necessitate transformation by weighting. Comparing the more stringent axioms from Section 3 to those from Section 2 illuminates the essential differences between weighting and other types of transformations. Section 4 introduces the model and discusses how it is usually the case that weights on both the likelihood function and the prior probability distribution are generally necessary for a full description of an agent's beliefs within the weighted updating framework.

Section 5 details two ways in which the base model can be expanded, allowing weighted updating to model several biases it otherwise could not. Previous studies that utilize the weighted updating model all implicitly assume that if non-prior information is mis-weighted then each datum is mis-weighted by the same factor. Subsection 5.1 discusses how it is possible to relax this restriction. In particular, this expanded version of weighted updating allows one to model those biases that involve discrimination between non-prior pieces of information as with, for

example, order effects[3] and self-attribution bias.[4] In subsection 5.2 I relax another implicit restriction that is present in the previous literature, namely that weights do not change over time.

Section 6 considers the problem of finding point estimates of distribution parameters by way of maximizing the weighted posterior distribution. The main result of this section provides a set of sufficient conditions for maximization that are satisfied by many workhorse distributions. As the implicit function theorem is utilized in this result, this result also provides comparative static conditions. To illustrate how the weighted updating model might be applied, Section 7 utilizes the main result from Section 6 to show how optimism bias can be a consequence of self-attribution bias. Section 8 provides concluding thoughts.

2 An Axiomatic Development

Throughout the paper, h_t denotes an ordered history of observations $(x_1, \ldots, x_t)$. A decision maker will consider h_t as an outcome from a stochastic process with density $f(h_t|\theta)$, where θ is an unknown parameter from parameter space Θ. Bayesian beliefs regarding the value of θ after observing h_t are completely described by the posterior distribution $\pi(\theta|h_t)$. Denote the likelihood function with $f(h_t|\theta)$ and the prior distribution with $\pi(\theta)$, then Bayes' rule states that

$$\pi(\theta|h_t) = \frac{f(h_t|\theta)\pi(\theta)}{\int_\Theta f(h_t|\theta)\pi(\theta)\, d\theta}.$$

[3]Order effects are when the order of data affects the beliefs those data are based upon. For example, the recency and primacy effect respectively describe cases where more or less recent data have more salience in belief formation.

[4]Self-attribution bias involves attributing desirable events to internal factors (such as ability) while attributing undesirable outcomes to bad luck or external factors.

Suppose we want to model an individual whose beliefs are biased by a likelihood function that is too strong, in the sense that the posterior distribution is affected too much by the data h_t. Or perhaps the likelihood function is too weak. Alternatively, the prior may be too strong or weak relative to the prior that a rational person would have (given the prior information they have seen, if any). The following three axioms shall be treated as necessary for modelling any such combination of weak or strong likelihood functions and prior distributions.

Since the axioms apply to both likelihood functions and prior distributions, we will use neutral notation that has not already been assigned to either type of distribution. Let g and Γ denote distributions that represent beliefs based on identical data,[5] where g is stronger than Γ. Also, define the transformation $T : [0,1] \to [0,1]$ by $T \circ g = \Gamma$.

Axiom 1 (Single-Valued)**.** T is a function (i.e. it is single-valued). Equivalently,

$$g(\omega_1) = g(\omega_2) \quad \Leftrightarrow \quad \Gamma(\omega_1) = \Gamma(\omega_2).$$

Axiom 2 (Monotone)**.** T is monotonically increasing. Equivalently,

$$g(\omega_1) > g(\omega_2) \quad \Leftrightarrow \quad \Gamma(\omega_1) > \Gamma(\omega_2).$$

Axiom 1 necessitate that T is a function while Axiom 2 ensures that this function is monotonically increasing. This monotonicity eliminates pairs of distributions that are opposite in the sense that the maximizer of one is the minimizer of the other (which Axiom 1 on its own does not preclude). In tandem, Axioms 1 and 2 ensure that the ordinal properties are identical. In other words, two agents

[5]The "data" can either be h_t or prior information.

with beliefs g and Γ that in tandem satisfy these first two axioms will agree on a rank ordering of the ω's according to their likelihoods as given by their respective beliefs. This is to be expected if the only substantive difference between the two distributions is that one is stronger than the other.

Technically, Axioms 1 and 2 partition the set of all distributions into equivalence classes. There remains the task of ordering the distributions within these equivalence classes according to how strong they are, which is the job of the next axiom. The question is: how should these distributions be ordered? To answer this question, we first need to precisely define what we mean by one set of beliefs being stronger or weaker than another set of beliefs. I submit that when we say that one set of beliefs is *stronger* than another set of beliefs that agree with each other (in that they satisfy Axioms 1 and 2) that we mean that the stronger beliefs have higher likelihood on parameter values that both agree are more likely and, correspondingly, lower likelihood on parameter values that they agree are less likely. That is, the stronger beliefs have "higher highs" and "lower lows". It will be shown below (in Theorem 1) that the following axiom guarantees this notion of strength.

Axiom 3 (Likelihood Concentration Order)**.** Any random variable with density g is larger, in the likelihood concentration order, than any random variable with density Γ. This ordering is defined by

$$g(\omega_1) > g(\omega_2) \quad \Rightarrow \quad \frac{g(\omega_1)}{g(\omega_2)} > \frac{\Gamma(\omega_1)}{\Gamma(\omega_2)}.$$

Axiom 3 introduces a new stochastic order that allows one to rank sets of distributions that satisfy the first two axioms according to the strength of the beliefs that the distributions represent. Note that the likelihood concentration

order defined in Axiom 3 is not equivalent to the likelihood ratio order, which is defined by the ratio of one likelihood to another increasing monotonically over the union of their supports. It is also different than all of the orders described in Shaked and Shanthikumar (2007), many of which necessitate restrictions on how the means and variances of ordered distributions relate, restrictions that are not necessary to the likelihood concentration order. A stochastic order that stipulates a relation regarding the variances of distributions is undesirable for our purposes because, for example, greater variance does not entail "weaker" beliefs for all distributions, e.g. many distributions that are multi-peaked.[6]

The following definition encapsulates Axioms 1, 2, and 3.

Definition 1 (Monotone Dispersion, Monotone Concentration)**.** For two non-uniform probability distributions Γ and g on the same support Ω, Γ is a *monotone dispersion of* g if for all pairs $(\omega_1, \omega_2) \in \Omega^2$ Axioms 1, 2, and 3 are satisfied. If Γ is a monotone dispersion of g then g is a *monotone concentration of* Γ.[7]

Essentially, if a distribution g is a monotone concentration of Γ then one could say that g is stronger than Γ, so throughout the rest of the paper we will use these more technical terms in place of colloquial terms such as "stronger" and "weaker".

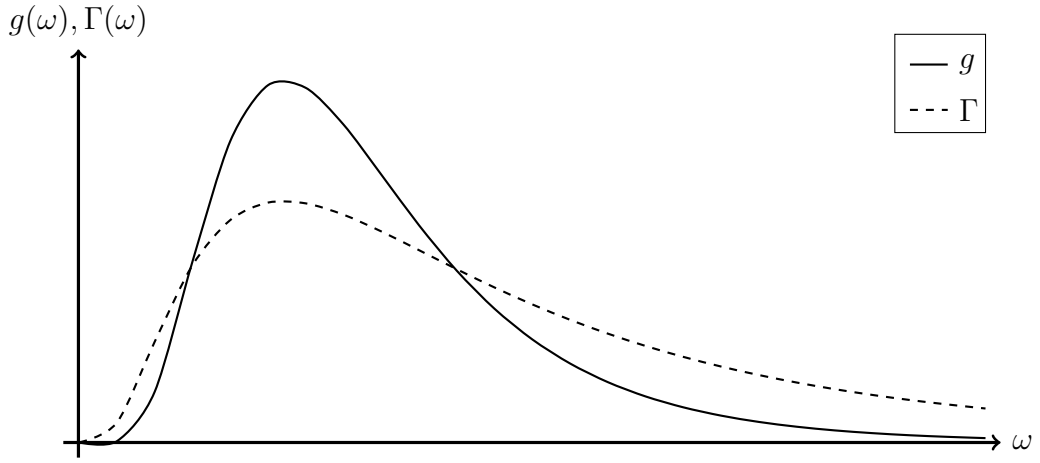
See Figure 1 for an example of two distributions to which these terms apply. Note that a monotone dispersion differs from a "monotone spread", a related concept due to Quiggin (1988) that is necessarily mean-preserving.

The following theorem shows that a monotone concentration has "higher highs"

[6]This is discussed further in the next subsection, as it is shown why variance does not necessarily represent orderings described by Axioms 1, 2, and 3.

[7]Uniform distributions are excluded from Definition 1 because if either g or Γ were uniform then the other would necessarily be uniform by Axiom 1, so they would be the same distribution. If this is the case then Axioms 2 and 3 are only *vacuously* true, which is not useful for our purposes because Axiom 3 provides an asymmetry that allows one to compare different distributions. Another way of saying this is that such a restriction ensures that the relations "is a monotone dispersion of" and "is a monotone concentration of" are not symmetric.

Figure 1: Γ is a monotone dispersion of g. Equivalently, g is a monotone concentration of Γ.



and "lower lows" than any of its monotone dispersions, characteristics that were mentioned above as indicative of a distribution being relatively stronger than another.

Theorem 1. Let Γ be a monotone dispersion of g. For any $\omega_1, \omega_2 \in \Omega$,

$$g(\omega_1) > g(\omega_2) \geq \Gamma(\omega_2) \quad \Rightarrow \quad g(\omega_1) > \Gamma(\omega_1).$$

Also,

$$g(\omega_1) < g(\omega_2) \leq \Gamma(\omega_2) \quad \Rightarrow \quad g(\omega_1) < \Gamma(\omega_1).$$

The following Corollary[8] to Theorem 1 is used in the proof of Theorem 2 below.

Corollary 1. Let Γ be a monotone dispersion of g and let ω^* be a maximizer of g and Γ. Then $g(\omega^*) > \Gamma(\omega^*)$.

Now that the notions of monotone concentration and monotone dispersion

[8]The appendix contains a proof of this Corollay that is independent of Theorem 1.

have been characterized, with a concentration having higher highs and lower lows than a dispersion, the following subsection discusses how these notions could be represented by a quantitative measure.

2.1 Measuring Concentration and Dispersion

As variance is a widely used measure of dispersion, one may suspect that a monotone dispersion results in a distribution with greater variance and a monotone concentration less variance than the original distribution. For many distributions this is indeed the case. Consider the normal distribution with mean μ and variance σ^2. It is straightforward to find that increasing (decreasing) the variance leads to a monotone dispersion (concentration).

Despite being true for normal distributions, it is not the case for all families of distributions that a monotone dispersion implies greater variance and that a monotone concentration has less variance, which may be unsurprising to those readers who followed the discussion of Axiom 3 above.

Consider the beta distribution $B(a, b)$ which is proportional to $x^{\alpha-1}(1-x)^{b-1}$ for parameters $a, b > 0$. Cases in which $a, b \in (0, 1)$ result in a u-shaped distribution, as any such distribution would be strictly convex with peaks at the extremes of the support, $x = 0$ and 1. Conversely, applying a monotone dispersion results in a flatter distribution with less variance and applying a monotone concentration shifts mass toward the end-points of $[0, 1]$ resulting in greater variance. In particular, consider $B(1/2, 1/2)$ which has a variance of $1/8$ while $B(3/4, 3/4)$ has a variance of $1/10$, despite the fact that $B(1/2, 1/2)$ is a monotone concentration of $B(3/4, 3/4)$.

The reason variance does not have a consistent relationship with monotone dispersions and concentrations is because it is a measure of dispersion *from the*

mean of the distribution. For a consistent representation of monotone dispersion and concentration it is necessary to have a measure of dispersion that is independent of reference points. As will be shown before the end of the current section, a distribution's information entropy, as defined in Shannon (1948), is a measure of dispersion or uncertainty that invariably increases for monotone dispersions and decreases for monotone concentrations.

Definition 2 (Information Entropy, (Shannon, 1948))**.** For any distribution $g : \Omega \to \mathbb{R}_{++}$, the *information entropy of* g is given by

$$H(g) \equiv -\int_{\Omega} g(\omega) \log g(\omega)\, d\omega.$$

Entropy is usually introduced using a discrete distribution g, for which the entropy is defined analogously as $H(g) \equiv -\sum_{\Omega} g(\omega) \log_c g(\omega)$, where the base c determines unit of measure (e.g. *bits* for $c = 2$). The concept defined in Definition 2 is usually known as *differential entropy* or *continuous entropy* and is typically denoted with h rather than H. The continuous version is studied because, for our purposes, its analysis is not as straightforward and the results for discrete distributions follow by analogy.

One reason that information theorists typically present entropy using discrete densities is because the entropy of a discrete distribution can be interpreted as the average length of code necessary for the efficient transmission of information regarding outcomes from that distribution. For a coin flip the length of the average code should be $-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$ bit per signal because it would be efficient to let, say, 1 encode *heads* and 0 encode *tails*. However, for some continuous distributions this interpretation of entropy is nonsensical because the entropy could be negative or not an integer. For example, the uniform distribution

over $[0, 1/2]$ has entropy $-\int_0^{1/2} 2 \log_2 2 \, dx = -1$ bits per signal. This paper is interested in comparing the entropies of distributions rather than interpreting entropy as the efficient average length of a message, so the paper does not focus on discrete densities.

For any distribution g and particular $\omega \in \Omega$, Tribus (1961) dubbed $-\log g(\omega)$ the *surprisal* of ω. Because $-\log g(\omega)$ is decreasing in $g(\omega)$, surprisal is greater for ω which (according to g) are less likely and, therefore, more *surprising* outcomes. The logarithm ensures that surprisal is additive in the densities of independent random variables, as for any two independent random variables X and Y respectively distributed g_X and g_Y, the surprisal for any particular pair of events (x, y) is

$$-\log g_X(x) g_Y(y) = -\log g_X(x) - \log g_Y(y).$$

Defining $-\log g(\omega)$ as the surprisal suggests that the information entropy of a distribution is equivalent to the *expected surprisal*, as entropy is equivalent to weighting the surprisal for each $\omega \in \Omega$ by the associated density $g(\omega)$ and aggregating over Ω. Distributions with higher entropy then can be interpreted as having higher expected surprisal. If outcomes from one distribution are, on average, more surprising than outcomes from another distribution, then the first distribution can be thought of as containing less information than the second. Thus, distributions with higher entropy typically generate observations that have less information content.[9]

The following theorem verifies the claim that transforming a distribution by

[9]The interpretation of information entropy as a measure of the uninformativeness of a distribution is consistent with the idea that physical entropy, which is proportional to information entropy by Boltzmann's constant, is a measure of one's *ignorance* of a system. See, for example, the discussion in Sethna (2006, §5.3) for this interpretation of physical entropy along with a discussion of its relationship with information entropy.

monotone dispersion results in an increase in entropy and that monotone concentration decreases entropy.[10]

Theorem 2. Let Γ be a monotone dispersion of g. Then the entropy of Γ is at least as great as the entropy of g. That is

$$-\int_{\Omega} \Gamma(\omega) \log \Gamma(\omega)\, d\omega \geq -\int_{\Omega} g(\omega) \log g(\omega)\, d\omega.$$

If, in addition, either of the sets $\{\omega : g(\omega) > \Gamma(\omega)\}$ or $\{\omega : g(\omega) < \Gamma(\omega)\}$ have positive measure, then the inequality is strict.

3 Exponentially Weighting a Distribution

This section shows that weighting a distribution by a positive weight results in a monotone dispersion (if the weight is less than one) or concentration (if the weight is greater than one) of that distribution. Then it is shown that replacing Axiom 3 with a more restrictive axiom necessitates a weighting transformation. Afterwards, the section provides a discussion of how, out of all possible transformations that entail either monotone concentration or dispersion, weighting a distribution is a particularly desirable transformation, as it is parsimonious and will typically maintain tractability.

[10]The reader may be interested to know that I considered using the entropy order in Axiom 3, but I decided to go with what I felt would be the likelihood concentration order because I felt it would be more intuitive for readers unfamiliar with information theory and because the math within my proof of the following result involves defining the likelihood concentration order anyway.

3.1 Weighting, Concentration, and Dispersion

Consider how an exponent γ transforms one probability distribution g to another proportional to g^γ. As long as $\gamma > 0$ taking g to the power γ is a monotone, increasing transformation, as is dividing by the resulting marginal distribution, which is always positive. Thus, weighting a distribution results in a new distribution and this pair of distributions satisfies Axioms 1 and 2. Less obviously, weighting also satisfies Axiom 3, so that it entails either a monotone concentration or dispersion, as the following theorem shows.

Theorem 3. Let $g : \Omega \to \mathbb{R}$ be any non-uniform probability distribution. If $\gamma \in (0, 1)$ then the distribution $\Gamma : \Omega \to \mathbb{R}$, defined as

$$\Gamma(\omega) \equiv \frac{g(\omega)^\gamma}{\int_\Omega g(\omega)^\gamma \, d\omega},$$

is a monotone dispersion of g. If it is the case that $\gamma > 1$ then Γ is a monotone concentration of g.

Thus, weighting a distribution is a particular method with which to generate a monotone concentration or dispersion, allowing all of the interpretations and results from the previous section to be applied.

3.2 Monotonicity & Proportional Elasticity

This subsection shows that exponential weighting follows from the following pair of axioms on belief formation. Comparing these axioms to Axioms 1, 2, and 3 is useful for understanding the differences between, on one hand, monotone concentrations and dispersions generally as compared to the particular transformations given by weighting distributions, on the other.

Note that the second of these axioms requires that g and $T \circ g$ are continuous distributions.

Axiom 4 (Monotonically Increasing Function)**.** The transformation T is a monotonically increasing function.

Axiom 5 (Proportional Elasticity)**.** For each $\omega \in \Omega$, the elasticity of the change in $T \circ g$ given a change in ω is proportional to the elasticity of a change in g. That is, if $T \circ g$ and g are differentiable at ω,

$$\frac{d \log T(g(\omega))}{d\omega} = k \frac{d \log g(\omega)}{d\omega},$$

for some $k \in \mathbb{R}$.

Axiom 4 ensures that ordinal rankings regarding the densities on Ω are identical between g and $T \circ g$. In other words if ω_1 is more likely than ω_2 according to g, then ω_1 is also more likely than ω_2 according to $T \circ g$. Note that a strict version of Axiom 4 is equivalent to the conjunction of Axioms 1 and 2. Thus, any pair of distributions related through a transformation that violates Axiom 4 could not be a monotonic dispersion or concentration of one another.

While Axiom 4 imposes an ordinal restriction, Axiom 5 restricts the cardinal properties of two distributions related through transformation, which makes sense at it replaces Axiom 3, which plays the same role but in a less restrictive manner. That is, Axiom 5 dictates how much densities can vary in a transformed distribution relative to the variation in the original distribution. Specifically, Axiom 5 entails that degrees of belief vary proportionately across the two distributions g and $T \circ g$, so a marginal change in ω induces a relative change in g that is proportional to a relative change in $T \circ g$, and the factor of proportionality is constant

for any given T.

The following theorem proves that only transformations involving exponentiating with a non-negative constant and normalizing satisfy Axioms 4 and 5.

Theorem 4. Let T and g satisfy Axioms 4 and 5. Then there exists some $\gamma \geq 0$ such that for each $\omega \in \Omega$

$$T(g(\omega)) \equiv \frac{g(\omega)^{\gamma}}{\int_{\Omega} g(\omega)^{\gamma}\, d\omega}.$$

3.3 Parsimony & Tractability

Putting an exponential weight on a distribution and normalizing is not the only transformation that yields a monotone dispersion or concentration, allowing the interpretation, via Theorem 2, that such a transformation alters the perceived information content of observations on average. Surely there are other families of transformations that depict certain biases in a more realistic fashion than weighted updating. Why focus on exponential weighting? Unfortunately, we do not currently have any idea of which transformations realistically model biased belief formation, so at this nascent stage in our understanding we should strive for characteristics other than realism – characteristics such as parsimony, tractability, and others that make a model valuable in scientific investigation.[11] This subsection provides an argument that weighted updating is typically both parsimonious and tractable.

That exponential weighting of a distribution is a parsimonious method of transformation is obvious – it involves doing one fairly basic operation with a single

[11] Gabaix and Laibson (2008) list seven key properties for economic models: parsimony, tractability, conceptual insightfulness, generalizability, falsifiability, empirical consistency, and predictive precision. Obviously, the latter two coincide with a model being realistic.

parameter. What about other basic operations that utilize a single parameter? It is possible to obtain transformations by adding or subtracting some positive constant and then normalizing (and being careful to avoid negative values when subtracting), but doing so typically results in a distribution that is not in the same family as the original distribution and, as such, would in all likelihood be extremely difficult to analyze. Thus, adding or subtracting a real number from each value of a distribution is parsimonious but would often result in a loss of tractability. Multiplying and dividing by some constant does not even result in a transformation, since the operation will be undone by normalizing. So, multiplication and division is parsimonious but entirely ineffectual.

In contrast, exponential weighting is typically both parsimonious and tractable, since the resulting distribution is often of the same family of distributions as the original. This is particularly important for our purposes because tractability can be easily lost when multiplying two (or more) distributions as is done in Bayes' rule, which is why Bayesian statisticians tend to study models wherein the prior distribution is conjugate to the likelihood function.[12]

An illustration of tractability being maintained by exponential weighting is provided by Theorem 5 in Section 6 utilizing Fact 2, that log-concavity is preserved after exponential weighting.

4 The Weighted Updating Model

In this section, we substitute the weighted distributions suggested in Section 3 into Bayes' rule to generate the weighted updating model. After introducing the model, I explain why weights on *both* the likelihood function and the prior

[12]Using a conjugate prior distribution guarantees (by definition) that the resulting posterior distribution is from the same family as that of the likelihood function.

probability distribution are generally necessary.

Recall Bayes' rule:

$$\pi(\theta|h_t) = \frac{f(h_t|\theta)\pi(\theta)}{\int_\Theta f(h_t|\theta)\pi(\theta)\,d\theta}.$$

Weighted updating augments Bayes' rule with real-valued parameters α and β as exponents respectively on the likelihood function and prior probability distribution. Denoting the posterior distribution under weighted updating after observing history h_t by $\tilde{\pi}(\theta|h_t)$, this form of weighted updating is given by[13]

$$\tilde{\pi}(\theta|h_t) = \frac{f(h_t|\theta)^\beta \pi(\theta)^\alpha}{\int_\Theta f(h_t|\theta)^\beta \pi(\theta)^\alpha\,d\theta}. \tag{1}$$

Both Bayes' rule and the weighted updating model can be stated without mention of the marginal distribution, which is not a function of θ and serves only as a normalization, ensuring that the posterior distribution aggregates to one over its support. Thus, Bayes' rule is often stated as

$$\pi(\theta|h_t) \propto f(h_t|\theta)\pi(\theta)$$

and, analogously, the weighted updating model can be displayed as

$$\tilde{\pi}(\theta|h_t) \propto f(h_t|\theta)^\beta \pi(\theta)^\alpha. \tag{1'}$$

[13]Note that throughout the paper it is assumed that the denominator $\int_\Theta f(h_t|\theta)^\beta \pi(\theta)^\alpha\,d\theta$ is finite so that $\tilde{\pi}(\theta|h_t)$ is well-defined. For many cases this assumption is innocuous because weighting a distribution with an exponent and rescaling results in a distribution from the original family. However, this assumption is not always satisfied. For example, the function $(1-p)x^{-p}$ represents a distribution over $x \geq 1$ if and only if $p > 1$. Taking such a distribution to a power $\alpha < 1/p$ and doing the usual normalization does not result in another distribution, as the integral over $[1, \infty)$ of the resulting function diverges.

4.1 Are Two Weights Necessary?

Putting more weight on the prior is qualitatively dual to putting less on the likelihood function, and vice-versa. This duality suggests that perhaps one can represent any given weighted posterior distribution with just one parameter, effectively restricting the other to a value of one. For example, one could model a bias that involves over-weighting prior information relative to non-prior information in at least two possible ways with the weighted updating model. For example, one could put a weight of one on the likelihood function and a weight greater than one on the prior distribution, or put a weight of one on the prior distribution and a weight less than one on the likelihood function. Could these approaches be equivalent, in that they are capable of resulting in identical weighted posterior distributions?

It is straightforward to see that this would not generally be the case. If it were, then, for example, given any h_t there would exist some $c > 0$ such that

$$f(h_t|\theta)^{\beta}\pi(\theta)^{\alpha} = cf(h_t|\theta)^{\gamma}\pi(\theta) \tag{2}$$

for all $\theta \in \Theta$. But if α, β, h_t, and c are fixed then expression (2) represents γ as an implicit function of θ. In other words γ would not necessarily be a constant.

Despite the fact that both parameters are necessary to study an entire weighted posterior distribution $\tilde{\pi}$, it may still be useful to transform the distribution so that there is effectively one parameter. For example, if one were to study point estimates by maximum likelihood (the subject of Section 6), then maximizing either $\tilde{\pi}^{\frac{1}{\alpha}}$ or $\tilde{\pi}^{\frac{1}{\beta}}$ would yield the same estimate of θ as maximizing $\tilde{\pi}$, since taking anything to these powers is a monotonic transformation.[14]

[14]By Theorem 3 both $\tilde{\pi}^{\frac{1}{\alpha}}$ and $\tilde{\pi}^{\frac{1}{\beta}}$ are either monotone dispersions and concentrations both of

5 Expanding the Framework

Expression (1′) introduced the weighted updating model

$$\tilde{\pi}(\theta|h_t) \propto f(h_t|\theta)^{\beta}\pi(\theta)^{\alpha}.$$

This introductory model involves two restrictions that can be relaxed to allow for a greater variety of biases that weighted updating can model. These restrictions are that (i) the agent treats each datum x_j as being exactly as informative as any other datum in h_t and (ii) the weights are constant over time. More general frameworks involve discarding either (or both) of these restriction, allowing for different weights on likelihood functions associated with different pieces of data or allowing the weights to vary over time.

5.1 Discrimination Between Data

Relaxing the restriction that each x_j is weighted by the same weight β involves utilizing the definition of conditional distribution functions, which says that for any $t \in \mathbb{N}$ and likelihood function $f(h_t|\theta)$,

$$f(h_t|\theta) = f(x_t|h_{t-1}, \theta)f(h_{t-1}|\theta).$$

Repeated iteration yields

$$f(h_t|\theta) = \prod_{j=1}^{t} f(x_j|h_{j-1}, \theta),$$

each other and of $\tilde{\pi}$.

which motivates setting up the weighted updating model as

$$\tilde{\pi}(\theta|h_t) \propto \pi(\theta)^{\alpha} \prod_{j=1}^{t} f(x_j|h_{j-1},\theta)^{\beta_j}, \tag{3}$$

where α remains the weight on the prior distribution and β_j, for each $j \in \{1,\ldots,t\}$, is the weight associated with the jth datum x_j. This is a generalization of the introductory framework because $(1')$ is a special case of (3), the special case being $\beta_j = \beta$ for each $j \in \{1,\ldots,t\}$.

In light of Theorems 2 and 3, we can say that if an individual's beliefs evolve according to the weighted updating model in (3), then, compared to a perfect Bayesian, the individual is subjectively treating the component distributions $\pi(\theta)^{\alpha}$ and $f(x_j|h_{j-1},\theta)^{\beta_j}$ for $j = 1,\ldots,t$ each as containing either more or less information depending on the levels of $\alpha, \beta_1, \ldots, \beta_t$, depending on how each of these weights compare to one. As the prior $\pi(\theta)$ summarizes prior information and each likelihood function $f(x_j|h_{j-1},\theta)$ represents the influence of an individual datum x_j, the weighted updating model in expression (3) essentially allows the individual to treat the prior information and each datum x_j at individualized levels of information content.

Additional biases that the generalized weighted updating model (3) is capable of modelling include anchoring; the availability heuristic; order effects, such as primacy and recency; and self-attribution bias. The remainder of this subsection discusses how to model these biases with weighted updating. Table 1 summarizes this discussion.

The availability heuristic generates biases due to certain observations being more available in memory (Tversky and Kahneman, 1973). This can be modelled using weighted updating simply by assuming that an economic agent puts greater

Table 1: Biases Involving Discrimination between Non-Prior Data

Cognitive Bias	Weights
Availability	β_j high for x_j that are more salient
Primacy Effect	β_j decreasing in j
Recency Effect	β_j increasing in j
Self-Attribution	β_j low if x_j is undesirable

weights on β_j's that correspond to x_j's that are relatively memorable.

Order effects occur when the relative position of observations seems to affect beliefs formed from those observations. Experimental subjects typically exhibit either the primacy effect, where earlier observations are more salient than later observations, or the recency effect, where the opposite occurs (Hogarth and Einhorn, 1992). To model the primacy effect with the weighted updating model would require that β_j decreases as j rises, while modelling the recency effect involves assuming that β_j is increasing in j.

Self-attribution bias occurs when individuals credit their own ability for desirable outcomes but blame undesirable outcomes on external factors, such as luck. This suggests that agents put greater weights on x_j that are desirable and lower weights on x_j that are undesirable. Section 7 contains a more in-depth discussion of self-attribution bias, before it is shown that a decision maker with such a bias will also exhibit optimism bias.

5.2 Dynamically Inconsistent Weights

This paper has, up to this point, presented the weighted updating model as one in which the weights are fixed. This subsection discusses relaxing this restriction

so that weights can change over time.

To allow the weights to change over time simply involves allowing them to be functions of time, which can be defined exogenously or endogenously depending on the nature of the application. Denote these functions $\alpha(t)$ and $\beta(t)$, so that after observing h_t the base weighted updating model in expression $(1')$ becomes

$$\tilde{\pi}(\theta|h_t) \propto f(h_t|\theta)^{\beta(t)}\pi(\theta)^{\alpha(t)}.$$

A bias that can be modelled with weights that change over time is base-rate neglect. As its name suggests, base-rate neglect involves ignoring prior information. However, subjects who exhibit base-rate neglect typically do not ignore prior information until after they have observed some non-prior information. A classic experiment on base-rate neglect is described in Kahneman and Tversky (1973). In this experiment, base rates differed between subjects: one group was told that the descriptions they observed were drawn from a population of 70 lawyers and 30 engineers, while the other group was told that they were drawn from a population with the frequencies reversed, 30 lawyers and 70 engineers. When experimental subjects observed a purposefully uninformative description of a man and were asked to guess whether he is an engineer or a lawyer, the average guess at the probability that the man was an engineer was approximately 50% *in both groups*. This base-rate neglect occurred even though the likelihoods participants gave were consistent with base rates before observing the irrelevant information, suggesting that participants utilized base rates then ignored them after observing the uninformative description. Such a phenomenon can be modelled by defining $\alpha(t)$ such that $\alpha(0) > 0$ (so that agents utilize prior information) and $\alpha(t) = 0$ for $t > 0$ (so

that they ignore the prior information after observing any history h_t).[15]

6 Maximizing the Weighted Updating Model

Although the weighted posterior distribution $\tilde{\pi}(\theta|h_t)$ fully describes an agent's beliefs regarding θ, it is often useful to work with point estimates. This section considers properties of the likelihood function and prior probability distribution that lend themselves to obtaining point estimates through maximization of the weighted posterior distribution. Relatively few distributions that see widespread use are concave, so it is desirable to consider weaker properties that may be useful for maximizing a weighted posterior distribution $\tilde{\pi}(\theta|h_t)$. A result below provides that, under typical conditions, log-concavity is the weakest assumption one can make on all of the primitive distributions in Bayes' rule and still be ensured of obtaining unique results from the analysis of first-order conditions.

A function g is (strictly) log-concave if $\log g$ is (strictly) concave. Equivalently, if g is (strictly) log-concave then for any $\lambda \in (0,1)$

$$g(\lambda\omega_1 + (1-\lambda)\omega_2) \quad (>) \geq \quad g(\omega_1)^{\lambda} g(\omega_2)^{(1-\lambda)}. \tag{4}$$

It turns out that many densities commonly used in economics are log-concave. Perhaps the most notable distribution that is log-concave despite not being concave is the normal distribution.[16]

The following Theorem provides a set of sufficient conditions, an element of

[15] Rabin (1996) points out that weighted updating with constant weights cannot account for base-rate neglect in light of irrelevant information. (See footnote 60 of Rabin (1996). Note that the version of this paper published in 1998 in *Journal of Economic Literature* does not include this discussion.)

[16] Bagnoli and Bergstrom (2005) contains an extensive classification of distributions that are and are not log-concave. See also Boyd and Vandenberghe (2004, Chapter 3)

which is log-concavity, for obtaining a unique maximizer of a weighted posterior distribution. This result is a testament to how transforming distributions with exponential weights maintains much tractability, as discussed in Section 3.

Theorem 5. Let $\alpha, \beta_1, \ldots, \beta_t > 0$, let Θ be a convex subset of $\mathbb{R}$, and let the prior distribution $\pi(\theta)$ and the likelihood functions $f(x_j|h_{j-1}, \theta)$, for all $t \in \mathbb{N}$ and $j \in \{1, \ldots, t\}$, each be positive-valued, twice continuously differentiable, and log-concave, with at least one of these $t + 1$ functions strictly log-concave. Then

$$\tilde{\theta}(t) \equiv \arg\max_{\theta \in \Theta} \ \tilde{\pi}(\theta|h_t)$$

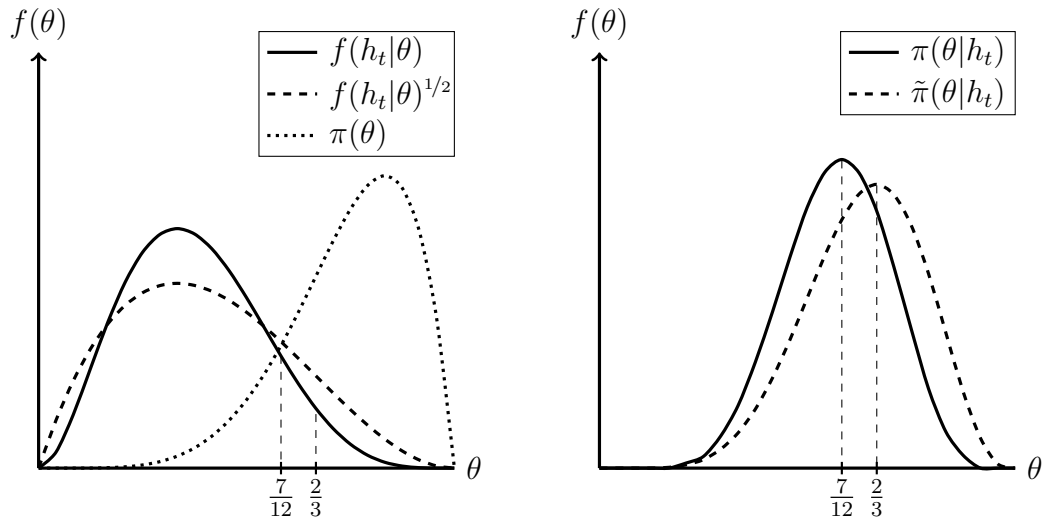
is a continuously differentiable function. Moreover, the sign of the partial derivative of $\tilde{\theta}(t)$ with respect to α is the same as the sign of $\pi'(\tilde{\theta}(t))$ and the sign of the partial derivative with respect to β_j is the same as the sign of $f_\theta(x_j|h_{j-1}, \tilde{\theta}(t))$ for each $j \in \{1, \ldots, t\}$.

With the other hypotheses given,[17] if any of the likelihood functions or if the prior distribution were not log-concave then the conclusions of Theorem 5 would not necessarily be true without imposing additional structure. As such, as long as the other hypotheses are maintained, it would be fruitless to consider properties weaker than log-concavity (e.g. quasiconcavity) to impose on *all* of the distributions primitive to Bayes' rule and expect to find that they are sufficient for the conclusions obtained in Theorem 5.

The comparative static results from Theorem 5 are that $\tilde{\theta}_\alpha(t)$ has the same sign as $\pi'(\tilde{\theta}(t))$ and each $\tilde{\theta}_{\beta_j}(t)$ has the same sign as $f_\theta(x_j|h_{j-1}, \tilde{\theta}(t))$. *Ceteris paribus*, if more weight is put on the prior $\pi(\theta)$ by increasing α then $\tilde{\theta}(t)$ will shift towards the

[17] The other hypotheses are $\alpha, \beta_1, \ldots, \beta_t > 0$; Θ a convex subset of $\mathbb{R}$; and $\pi(\theta)$ and $f(x_j|h_{j-1}, \theta)$, for all $t \in \mathbb{N}$ and $j \in \{1, \ldots, t\}$, each being positive-valued and twice continuously differentiable.

Figure 2: An illustration of comparative statics for log-concave weighted updating.



mode of $\pi(\theta)$. Similarly, if more weight is put on the jth datum x_j by increasing β_j then $\tilde{\theta}(t)$ will shift towards the mode of $f_\theta(x_j|h_{j-1},\theta)$. Thus, in either case a distribution with greater weight will pull the maximum *a posteriori* estimate towards its maximum. Figure 2 illustrates the comparative statics results using the introductory weighted updating model (1). It compares a perfect Bayesian (solid curves) with a weighted updater (dashed curves) who is putting a weight of $1/2$ on the likelihood function associated with h_t. Notice that the maximum for the Bayesian's posterior distribution $\pi(\theta|h_t)$ is at $\theta = 7/12$ and that at this point the slope of the likelihood function is negative: $f(h_t|7/12) < 0$, suggesting that a decrease in α will induce an increase in the value of θ that maximizes the posterior distribution. This is indeed the case, as the maximum of the weighted updating model $\tilde{\pi}(\theta|h_t)$ is at $\theta = 2/3 > 7/12$.

7 Optimism Bias via Self-Attribution

This section utilizes some of the findings from previous sections of the paper in an application of the weighted updating model. Specifically, this section illustrates how the weighted updating model can be used to show that optimism bias can be caused by self-attribution bias.

One of the hallmarks of optimism bias is the statistical impossibility of individuals who, on average, expect to do better than average in some realm of their lives. A classic example is provided by Weinstein (1980), who finds that students expect to live longer and be healthier than their peers, while believing themselves less likely to experience negative outcomes such as divorce and heart attacks. In regards to *how* optimism bias occurs, Sharot, Riccardi, Raio, and Phelps (2007) find differences in areas of the brain activated depending on whether imagined future events were desirable or not, and the differences suggest a heightened role for areas of the brain associated with monitoring emotional salience when desirable outcomes are imagined.

Evidence suggests that optimism bias has measurable effects on peoples' beliefs in many facets of life, including those that are economic in nature. Hoch (1985) finds that business school students overestimate their job prospects. Weinstein and Klein (1996) find that smokers tend to believe other smokers are more likely to suffer from lung cancer than themselves. In a cross-country analysis, Koellinger, Minniti, and Schade (2007) find an association between optimism and business start-ups and a negative association between optimism and survival of new firms. Chapin and Coleman (2009) find evidence suggesting people tend to believe they are less likely to be the victim of a crime than others. Bain (2009) finds optimism bias amongst private sector financiers regarding forecasts of toll road demand.

As Hirshleifer (2001) points out, rational learning would gradually erode optimism bias. Thus, for optimism bias to maintain, there must be some bias in the learning process. One explanation of the kind of irrational learning that leads to optimism bias is that such individuals tend to attribute undesirable outcomes to "luck" while attributing desirable outcomes to some quality of themselves, such as ability. If this is true, then weighted updating should be able to model such a learning process, as attributing some outcome to luck entails treating the outcome as relatively uninformative, in which case a low weight would be put on the likelihood distribution associated with that outcome. This section shows that weighted updating can capture this phenomenon and generate beliefs that exhibit optimism bias.

Assume that higher levels of x_t are in some sense desirable, so that somebody who exhibits optimism bias expects higher levels of x_t than a perfect Bayesian who has witnessed the same history and has the same prior.[18] Modelling this requires that the parameter θ defines some sort of stochastic ordering so that either higher or lower levels of θ are associated with higher levels of x_t in each period.[19] Without loss of generality, greater θ are associated with larger x_t.

The stochastic ordering utilized is the strict single crossing property defined by Milgrom and Shannon (1994). This property is chosen over other stochastic orders because of its generality and because it is maintained under monotone dispersion and concentration, making it particularly useful for analyses involving the weighted updating model. If $g(y, \omega)$ satisfies the strict single crossing property

[18] It may be helpful to think of x_t as income.

[19] If one thinks of x_t as income, then values of θ might represent beliefs about one's earning power ("ability", in a loose sense).

in $(\omega; y)$ then for all $y' > y$ and $\omega' > \omega$ it is true that

$$g(y, \omega') \geq g(y, \omega) \quad \Rightarrow \quad g(y', \omega') > g(y', \omega).$$

We will assume that likelihood functions f satisfy the strict single crossing property in $(\theta; x)$. The essence of this assumption is that in the case that after witnessing some level of x_t the decision maker views θ' as more likely than θ, where $\theta' > \theta$, it follows that had a higher level of x_t been witnessed then θ' would still have been deemed more likely than θ.

We will model optimism bias by assuming that an individual will put more weight on likelihood functions associated with high levels of x_t and less weight on lower levels of x_t than a perfect Bayesian. Determination of which are higher and lower levels of x_t will depend upon a reference level, which is defined as the minimal level of x_{t+1} that would be sufficient for $\tilde{\theta}(t+1) \geq \tilde{\theta}(t)$.[20] That is, the reference level immediately after observing h_t is

$$\tilde{x}(t) \equiv \inf \left\{ x : \arg\max_{\theta \in \Theta} f(x|\theta) \geq \tilde{\theta}(t) \right\}.$$

The weights the individual who exhibits optimism bias will utilize are $\beta_t = \beta(x_t, t) > 0$, where each function $\beta(x, t) - 1$ has a single crossing at $x = \tilde{x}(t-1)$. That is, for all $t \in \mathbb{N}$,

$$\beta(x_t, t) \quad \begin{cases} > 1 & \text{if } x_t > \tilde{x}(t-1) \\ = 1 & \text{if } x_t = \tilde{x}(t-1) \\ \in (0,1) & \text{if } x_t < \tilde{x}(t-1). \end{cases} \tag{5}$$

[20] Recall that $\tilde{\theta}(t) \equiv \arg\max_\theta \tilde{\pi}(\theta|h_t)$.

Now it can be shown that under such a setup optimism bias will occur, in that regardless of the history h_t an individual who forms beliefs according to these rules will believe θ is greater than a Bayesian who has observed the same history and has the same prior distribution $\pi(\theta)$.[21]

Proposition 1. Let Θ be a convex subset of $\mathbb{R}$, and let the prior distribution $\pi(\theta)$ and the likelihood functions $f(x_j|h_{j-1},\theta)$, for all $t \in \mathbb{N}$ and $j \in \{1,\ldots,t\}$, each be positive-valued, twice continuously differentiable, and strictly log-concave. Also, let each likelihood function f satisfy the strict single crossing property in $(\theta; x)$ and let the functions $\beta(x_t, t)$ for each t be consistent with (5). Then for all $t \in \mathbb{N}$,

$$\arg\max_{\theta\in\Theta} \quad \pi(\theta)\prod_{j=1}^{t} f(x_j|h_{j-1},\theta)^{\beta(x_j,j)} \quad \geq \quad \arg\max_{\theta\in\Theta} \quad \pi(\theta)\prod_{j=1}^{t} f(x_t|h_{j-1},\theta).$$

A model of optimism bias such as this one can be used in any number of applications. A few that come to mind are (i) a job-search model where the job hunter is overly-optimistic about future job offers and turns down offers that a perfect Bayesian or agent with rational expectations would accept, (ii) a price-setting firm that is overly-optimistic regarding stochastic demand and thereby sets prices higher than would be profit-maximizing,[22] and (iii) a firm that over-estimates the expected return on an investment project from which a perfect Bayesian would abstain.

[21] It may be of interest to note that the necessity that the reference level $\tilde{x}(t)$ changes over time stems from the fact that relatively weak assumptions were used regarding $\beta(x,t)$. If β is monotonically increasing in x then a constant reference level would be sufficient for optimism bias to manifest.

[22] For example, a monopolist who optimistically under-estimates the price elasticity of demand.

8 Concluding Remarks

This paper presents weighted updating as a generalization of Bayes' rule that is capable of systematically producing biased judgement in economic agents. I provide an interpretation of weighted updating as a method by which individuals treat information as either more or less informative than under Bayes' rule. In particular, it is shown that weighting the functions primitive to Bayes' rule transforms the functions by monotone dispersion or monotone concentration, and that these transformations affect the information entropy of the resulting primitives.

These results provide the interpretation that weighted updating is a parametric method with which to model the treatment of data as either more or less informative than with Bayesian updating. As such, weighted updating embodies a theory of biased judgement, wherein these biases are a result of the treatment of data as containing inaccurate levels of information content. I should emphasize that the interpretation of weighting a distribution suggests that, on its own, weighted updating may be appropriate to model only those biases in which individuals correctly interpret information, but for some reason do not use the information in a rational way. Thus, for example, weighted updating may be utilized to model biases based on self-deception[23] or the cognitive limitations of utilizing correctly interpreted data, but it may not be appropriate for modelling the type of confirmation bias studied by Rabin and Schrag (1999), which involves decision makers who misinterpret information. Still, there is no reason why there should be only one type of bias affecting belief formation; one could, for example, model individuals who misinterpret evidence using the framework of Rabin and Schrag (1999) and then process the misinterpreted information irrationally using weighted updating.

[23]Self-deception typically involves individuals who downplay or overemphasize the importance of certain pieces of evidence in a systematic way (Hirshleifer, 2001).

Appendix

Proof of Theorem 1. Let

$$g(\omega_1) > g(\omega_2) \geq \Gamma(\omega_2).$$

As Γ is a monotone dispersion of g, $g(\omega_1) > g(\omega_2)$ implies

$$\frac{g(\omega_1)}{g(\omega_2)} > \frac{\Gamma(\omega_1)}{\Gamma(\omega_2)},$$

which can be rearranged to obtain

$$\frac{\Gamma(\omega_2)}{g(\omega_2)} > \frac{\Gamma(\omega_1)}{g(\omega_1)}.$$

Now utilize $g(\omega_2) \geq \Gamma(\omega_2)$ to augment the above inequality to obtain

$$1 \geq \frac{\Gamma(\omega_2)}{g(\omega_2)} > \frac{\Gamma(\omega_1)}{g(\omega_1)}.$$

And so, $g(\omega_1) > \Gamma(\omega_1)$. The other case implying the opposite conclusion is symmetric. **Q.E.D.**

Proof of Corollary 1. As

$$\omega^* \in \arg\max_{\omega \in \Omega} g(\omega),$$

we have $g(\omega^*) \geq g(\omega)$ for each $\omega \in \Omega$. The hypothesis that Γ is a monotone dispersion of g implies that both Γ and g are non-uniform, so there exists some $\omega_0 \in \Omega$ such that $g(\omega^*) > g(\omega_0)$. Thus,

$$\frac{g(\omega^*)}{g(\omega)} \geq \frac{\Gamma(\omega^*)}{\Gamma(\omega)} \quad \text{for all } \omega \in \Omega.$$

Note that Axiom 2 guarantees that this inequality is strict at $\omega = \omega_0$. These conditions imply

$$\frac{g(\omega)}{g(\omega^*)} \leq \frac{\Gamma(\omega)}{\Gamma(\omega^*)} \quad \text{for all } \omega \in \Omega,$$

with strict in inequality for $\omega = \omega_0$. As these conditions hold for all $\omega \in \Omega$ with strict inequality at ω_0, integrating over Ω yields

$$\frac{\int_\Omega g(\omega)\, d\omega}{g(\omega^*)} < \frac{\int_\Omega \Gamma(\omega)\, d\omega}{\Gamma(\omega^*)}.$$

As both g and Γ are probability distributions, they integrate to unity over their support, so this condition is equivalent to

$$\frac{1}{g(\omega^*)} < \frac{1}{\Gamma(\omega^*)},$$

which is true only if $g(\omega^*) > \Gamma(\omega^*)$. **Q.E.D.**

The proof of Theorem 2 requires the following two lemmas and a fact (Gibb's Inequality) from statistical physics.

Lemma 1. Let Γ be a monotone dispersion of g. Then

$$\sup \Gamma(\{\omega : g(\omega) < \Gamma(\omega)\}) \leq \inf \Gamma(\{\omega : g(\omega) > \Gamma(\omega)\})$$

Proof. Let $b = \sup \Gamma(\{\omega : g(\omega) < \Gamma(\omega)\})$ and $B = \inf \Gamma(\{\omega : g(\omega) > \Gamma(\omega)\})$. Suppose for purposes of contradiction that $b > B$. Then completeness of the interval $(B, b) \subset \mathbb{R}_{++}$ implies that there exist $\omega_1, \omega_2 \in \Omega$ such that $\Gamma(\omega_1) > \Gamma(\omega_2)$,

$$\Gamma(\omega_1) \in \Gamma(\{\omega : g(\omega) < \Gamma(\omega)\}),$$

and

$$\Gamma(\omega_2) \in \Gamma(\{\omega : g(\omega) > \Gamma(\omega)\}).$$

By the definition of monotone dispersion and monotone concentration, $\Gamma(\omega_1) > \Gamma(\omega_2)$ if an only if $g(\omega_1) > g(\omega_2)$. This, the above two conditions, and the fact that Γ is positive on its support Ω imply

$$\Gamma(\omega_1) > g(\omega_1) > g(\omega_2) > \Gamma(\omega_2) > 0,$$

from which it follows that

$$\frac{\Gamma(\omega_1)}{\Gamma(\omega_2)} > \frac{g(\omega_1)}{g(\omega_2)} > 1,$$

contradicting the fact that Γ is a monotone dispersion of g, Axiom 3 in particular. Therefore it must be the case that $b \leq B$. **Q.E.D.**

Note that for continuous distributions g and Γ, it will necessarily be the case that b = B so that Lemma 1 is automatic. Also, for a discrete distribution min and max can be respectively substituted for inf and sup, making the proof somewhat less technical.

Lemma 2. Let Γ be a monotone dispersion of g and let there exist some ω_j such that $\Gamma(\omega_j) > g(\omega_j)$. Then there exists $r \in \mathbb{R}$ such that

$$\Gamma(\omega) > r \quad \Rightarrow \quad g(\omega) > \Gamma(\omega)$$

and

$$\Gamma(\omega) < r \quad \Rightarrow \quad g(\omega) < \Gamma(\omega).$$

Proof. Corollary 1 guarantees the existence of some $\omega \in \Omega$ such that $g(\omega) > \Gamma(\omega)$.

Define B as in the proof of Lemma 1, and it follows that $\Gamma(\omega) \geq B$. If it is the case that $\Gamma(\omega) > B$ then by definition $g(\omega) > \Gamma(\omega)$. In summary, $\Gamma(\omega) > B$ implies that $g(\omega) > \Gamma(\omega)$.

The hypothesis that there exists some ω_j such that $\Gamma(\omega_j) > g(\omega_j)$ establishes the existence of $\omega \in \Omega$ such that $\Gamma(\omega) \leq b$, where b is defined in the proof of Lemma 1. A symmetric argument to the above guarantees that $\Gamma(\omega) > g(\omega)$ whenever $\Gamma(\omega) < b$.

Thus, for any $r \in [b, B]$, which is non-empty by Lemma 1, it follows that

$$\Gamma(\omega) > r \quad \Rightarrow \quad g(\omega) > \Gamma(\omega)$$

and

$$\Gamma(\omega) < r \quad \Rightarrow \quad g(\omega) < \Gamma(\omega). \qquad \textbf{Q.E.D.}$$

We will make use of the following fact from the field of statistical physics.

Fact 1 (Gibbs' Inequality)**.** For any two probability distributions $p, q : X \to \mathbb{R}_{++}$

$$\int_X p(x) \log p(x)\, dx \geq \int_X p(x) \log q(x)\, dx.$$

Proof of Theorem 2. By Gibbs' Inequality

$$\int_\Omega g(\omega) \log g(\omega)\, d\omega \geq \int_\Omega g(\omega) \log \Gamma(\omega)\, d\omega,$$

which implies

$$\int_\Omega g(\omega) \log g(\omega) - \Gamma(\omega) \log \Gamma(\omega)\, d\omega \geq \int_\Omega (g(\omega) - \Gamma(\omega)) \log \Gamma(\omega)\, d\omega. \qquad (6)$$

Lemma 1 asserts that $[b, B]$ is non-empty. Consider any $r \in [b, B]$. As, g and Γ are both distributions,

$$0 = -\log r \int_{\Omega} g(\omega) - \Gamma(\omega)\, d\omega. \tag{7}$$

Adding expressions (6) and (7) gives

$$\int_{\Omega} g(\omega) \log g(\omega) - \Gamma(\omega) \log \Gamma(\omega)\, d\omega \geq \int_{\Omega} [g(\omega) - \Gamma(\omega)](\log \Gamma(\omega) - \log r)\, d\omega. \tag{8}$$

By Lemma 2, $r \in [b, B]$ implies that $\log \Gamma(\omega) - \log r$ has the same sign as $g(\omega) - \Gamma(\omega)$, so the right-hand side of expression (8) is non-negative. And so,

$$-\int_{\Omega} \Gamma(\omega) \log \Gamma(\omega)\, d\omega \geq -\int_{\Omega} g(\omega) \log g(\omega)\, d\omega. \tag{9}$$

If, additionally, $\{\omega : g(\omega) > \Gamma(\omega)\}$ or $\{\omega : g(\omega) < \Gamma(\omega)\}$ have positive measure then the right-hand side of expression (8) is strictly positive, so inequality (9) is strict. **Q.E.D.**

Proof of Theorem 3. Let $\gamma \in (0, 1)$. Axioms 1 and 2 are satisfied immediately. As g is non-uniform there exists a pair $(\omega_1, \omega_2) \in \Omega^2$ for which $g(\omega_1) > g(\omega_2)$. For any such pair, multiplying each term of the relations $0 < \gamma < 1$ by $\log(g(\omega_1)/g(\omega_2))$ yields

$$0 < \gamma \log \frac{g(\omega_1)}{g(\omega_2)} < \log \frac{g(\omega_1)}{g(\omega_2)},$$

which implies that

$$1 < \frac{g(\omega_1)^{\gamma}}{g(\omega_2)^{\gamma}} < \frac{g(\omega_1)}{g(\omega_2)}.$$

Dividing both the numerator and denominator of the center term by the normal-

izing factor $\int_\Omega g(\omega)^\gamma \, d\omega > 0$ yields

$$1 < \frac{g(\omega_1)^\gamma / \int_\Omega g(\omega)^\gamma \, d\omega}{g(\omega_2)^\gamma / \int_\Omega g(\omega)^\gamma \, d\omega} < \frac{g(\omega_1)}{g(\omega_2)},$$

which is another way of stating that

$$1 < \frac{\Gamma(\omega_1)}{\Gamma(\omega_2)} < \frac{g(\omega_1)}{g(\omega_2)}.$$

This proves that Γ is a monotone dispersion of g. The case for $\gamma > 1$ yielding a monotone concentration is proved analogously. **Q.E.D.**

Proof of Theorem 4. We have from Axiom 5

$$\frac{d \log T(g(\omega))}{d\omega} = k \frac{d \log g(\omega)}{d\omega}.$$

Solving this differential equation implies that

$$T(g(\omega)) = c g(\omega)^k$$

for some $c > 0$ and $k \in \mathbb{R}$. Let $\gamma = k$. The fact that $\gamma \geq 0$ follows from Axiom 4. The value of c is determined by the fact that $T \circ g$ is a distribution which necessitates $c = 1/\int_\Omega g(\omega)^\gamma \, d\omega$. **Q.E.D.**

The following results provide properties due to log-concavity that are useful for maximizing the weighted updating model. Note that only the strict cases are shown in all proofs, as the non-strict cases are nearly identical.[24]

[24]Fore sake of completeness the proofs are presented even though they are almost certainly not new.

Fact 2. For any γ $(>) \geq$ 0, if a function $g : \Omega \to \mathbb{R}$ is (strictly) log-concave then $g^\gamma : \Omega \to \mathbb{R}$ is (strictly) log-concave.

Proof. For any $(\omega_1, \omega_2) \in \Omega^2$ and any $\lambda \in (0, 1)$ it is necessary that

$$g(\lambda\omega_1 + (1-\lambda)\omega_2) > g(\omega_1)^\lambda g(\omega_2)^{(1-\lambda)}.$$

Taking logs, multiplying each side by $\gamma > 0$, and some rearranging yield

$$\log g(\lambda\omega_1 + (1-\lambda)\omega_2)^\gamma > \lambda \log g(\omega_1)^\gamma + (1-\lambda)\log g(\omega_2)^\gamma. \qquad \textbf{Q.E.D.}$$

Fact 3. For any $\gamma > 0$, if a function $g : \Omega \to \mathbb{R}$ is (strictly) log-concave then $\gamma g : \Omega \to \mathbb{R}$ is (strictly) log-concave.

Proof. Multiplying each side of the strict case of expression (4) by γ and distributing on the right-hand side yield

$$\gamma g(\lambda\omega_1 + (1-\lambda)\omega_2) > [\gamma g(\omega_1)]^\lambda [\gamma g(\omega_2)]^{(1-\lambda)}. \qquad \textbf{Q.E.D.}$$

Fact 4. If $f, g \subset \Omega \times \mathbb{R}_+$ are both log-concave functions then their pointwise product, denoted fg, is log-concave. If, in addition, either f or g is strictly log-concave and both are positive then fg is strictly log-concave.

Proof. Without loss of generality, let f be strictly log-concave and g log-concave. For any $\lambda \in (0, 1)$ it follows that

$$f(\lambda\omega_1 + (1-\lambda)\omega_2) > f(\omega_1)^\lambda f(\omega_2)^{(1-\lambda)}$$

and

$$g(\lambda\omega_1 + (1-\lambda)\omega_2) \geq g(\omega_1)^{\lambda} g(\omega_2)^{(1-\lambda)}.$$

All values in these expressions are positive, so multiplying left-hand sides and right-hand sides and utilizing the fg notation yield

$$fg(\lambda\omega_1 + (1-\lambda)\omega_2) > fg(\omega_1)^{\lambda} fg(\omega_2)^{(1-\lambda)} \qquad \textbf{Q.E.D.}$$

Facts 2, 3, and 4 imply the following result.

Corollary 2. For $\alpha, \beta_1, \ldots, \beta_t \geq 0$, if the likelihood functions $f(x_j|h_{j-1}, \theta)$ and the prior distribution $\pi(\theta)$ are log-concave then the weighted posterior distribution $\tilde{\pi}(\theta|h_t)$ is log-concave on Θ. If any one of the likelihood functions or the prior distribution is strictly log-concave and if the weight on that function is strictly positive then $\tilde{\pi}(\theta|h_t)$ is strictly log-concave

Proof. Let $\alpha, \beta_1, \ldots, \beta_t \geq 0$ and $\pi(\theta), f(x_1|\theta), \ldots, f(x_t|h_{t-1}, \theta)$ be log-concave and at least one of them strictly so. By Fact 2, each of the functions

$$\pi(\theta)^{\alpha}, f(x_1|\theta)^{\beta_1}, \ldots, f(x_t|h_{t-1}, \theta)^{\beta_t}$$

are log concave. Moreover, the weighted version of any strictly log concave function is strictly log concave. Fact 4 ensures that the product

$$\pi(\theta)^{\alpha} \prod_{j=1}^{t} f(x_j|h_{j-1}, \theta)^{\beta_j} \tag{10}$$

is strictly log-concave. Finally, since the marginal distribution

$$\int_{\Theta} \pi(\theta)^{\alpha} \prod_{j=1}^{t} f(x_j|h_{j-1},\theta)^{\beta_j}\, d\theta \tag{11}$$

is a positive constant (i.e. not a function of θ), dividing expression (10) by (11) to obtain $\tilde{\pi}(\theta|h_t)$ yields a strictly log-concave function by Fact 3. **Q.E.D.**

Corollary 2 asserts that if the prior distribution and likelihood functions primitive to Bayes rule are log-concave, then any associated weighted posterior distribution is log concave as long as the weights are all non-negative. Moreover, if any one of the primitive distributions is strictly log-concave and its weight positive, then, as long as the others are log-concave, the weighted posterior distribution is strictly log-concave. Strict log-concavity is very useful in maximization, because it ensures that any maximum is a unique global maximum.

The following fact of log-concavity will prove useful.

Fact 5. If $g : \mathbb{R} \to \mathbb{R}_+$ is log-concave and twice continuously differentiable then

$$g''(x)g(x) - g'(x)^2 \leq 0. \tag{12}$$

This inequality is strict if g is strictly log-concave.[25]

Proof. A twice continuously differentiable function f is strictly concave if $f'' < 0$.

[25]For $g(x) > 0$, rearranging expression (12) provides an illuminating implication of log-concavity when a function is twice continuously differentiable:

$$g''(x) \leq \frac{g'(x)^2}{g(x)}.$$

Thus, the second derivative of a log-concave function can be positive, which contrasts with the fact that a concave and twice continuously differentiable function has non-negative second derivative on its domain. This is one way of illustrating that log-concavity is a weaker condition than concavity.

As a function g is strictly log-concave if and only if $\log g$ is strictly concave, or $(\log g)'' < 0$ which yields expression (12). **Q.E.D.**

Proof of Theorem 5. Ignoring the marginal distribution and taking the natural logarithm is a monotonic transformation, leaving the problem

$$\max_{\theta \in \Theta} \ \alpha \log \pi(\theta) + \sum_{j=1}^{t} \beta_j \log f(x_j|h_{j-1}, \theta).$$

The first-order condition of maximization with respect to θ is

$$\alpha \frac{\pi'(\tilde{\theta}(t))}{\pi(\tilde{\theta}(t))} + \sum_{j=1}^{t} \beta_j \frac{f_\theta(x_j|h_{j-1}, \tilde{\theta}(t))}{f(x_j|h_{j-1}, \tilde{\theta}(t))} = 0, \tag{13}$$

By Corollary 2, $\tilde{\pi}(x_t|h_t, \theta)$ is strictly log-concave so $\tilde{\theta}(t)$ is a unique maximum on Θ. To prove that (13) defines $\tilde{\theta}(t)$ as a continuously differentiable function of $(\alpha, \beta_1, \ldots, \beta_t, h_t)$ one can use the implicit function theorem, a sufficient condition of which is the derivative of the left-hand side of (13) with respect to θ is non-zero at $\tilde{\theta}(t)$. In fact, it will be shown that this derivative is negative, which is also the second-order condition for maximization. This derivative is

$$\begin{aligned} \frac{\partial \text{FOC}}{\partial \theta} \equiv \alpha & \frac{\pi''(\tilde{\theta}(t))\pi(\tilde{\theta}(t)) - \pi'(\tilde{\theta}(t))^2}{\pi(\tilde{\theta}(t))^2} \\ & + \sum_{j=1}^{t} \beta_j \frac{f_{\theta\theta}(x_j|h_{j-1}, \tilde{\theta}(t)) f(x_j|h_{j-1}, \tilde{\theta}(t)) - f_\theta(x_j|h_{j-1}, \tilde{\theta}(t))^2}{f(x_j|h_{j-1}, \tilde{\theta}(t))^2} \end{aligned}$$

By hypothesis, $\alpha, \beta_1, \ldots, \beta_t > 0$. The denominators of each term are also positive. Utilizing Fact 5, log-concavity of all of the functions $\pi(\theta), f(x_1|\theta), \ldots, f(x_t|h_{t-1}, \theta)$ implies that all of the numerators are non-positive and since at least one of these

functions is strictly log-concave at least one of the numerators is negative. Thus,

$$\frac{\partial \text{FOC}}{\partial \theta} < 0,$$

establishing that the fist-order condition (13) defines $\tilde{\theta}(t)$ as a continuously differentiable implicit function of $(\alpha, \beta_1, \ldots, \beta_t, h_t)$.

The comparative static results are found by solving for the appropriate derivatives after finding the expressions for:

$$\frac{d}{d\alpha}\frac{\partial \text{FOC}}{\partial \theta} \qquad \text{and} \qquad \frac{d}{d\beta_j}\frac{\partial \text{FOC}}{\partial \theta},$$

for each $j \in \{1, \ldots, t\}$. Respectively, these derivatives are

$$\tilde{\theta}_\alpha(t) = -\frac{\pi'(\tilde{\theta}(t))}{\pi(\tilde{\theta}(t))} \Big/ \frac{\partial \text{FOC}}{\partial \theta} \qquad \text{and} \qquad \tilde{\theta}_{\beta_j}(t) = -\frac{f_\theta(x_j|h_{j-1}, \tilde{\theta}(t))}{f(x_j|h_{j-1}, \tilde{\theta}(t))} \Big/ \frac{\partial \text{FOC}}{\partial \theta}.$$

To determine the signs of these expressions first note that $\frac{\partial \text{FOC}}{\partial \theta} < 0$, which essentially cancels out the minus sign in each expression. Combining these with the facts $\pi(\tilde{\theta}(t)) > 0$ and $f(x_j|h_{j-1}, \tilde{\theta}(t)) > 0$, for each j, implies that the signs of these expressions are respectively the same as the signs of $\pi'(\tilde{\theta}(t))$ and $f_\theta(x_j|h_{j-1}, \tilde{\theta}(t))$. **Q.E.D.**

Proof of Proposition 1. As,

$$\tilde{\pi}(\theta|h_t) \propto \tilde{\pi}(\theta|h_{t-1}) f(x_t|h_{t-1}, \theta)^{\beta(x_t,t)}$$

we can write the first-order condition of maximization for the weighted updater

as

$$\frac{\tilde{\pi}_\theta(\tilde{\theta}(t)|h_{t-1})}{\tilde{\pi}(\tilde{\theta}(t)|h_{t-1})} + \beta(x_t, t)\frac{f_\theta(x_t|h_{t-1}, \tilde{\theta}(t))}{f(x_t|h_{t-1}, \tilde{\theta}(t))} = 0.$$

This first-order condition implies that the signs of $\tilde{\pi}_\theta$ and f_θ are either opposite or both zero at $\tilde{\theta}(t)$. There are two cases to consider.

Case 1: $x_t \geq \tilde{x}(t-1)$. In this case, the single crossing property on f guarantees

$$\arg\max_{\theta \in \Theta} f(x_t|\theta) \geq \tilde{\theta}(t-1).$$

As both $f(x_t|\theta)$ and $\tilde{\pi}(\theta|h_{t-1})$ are log-concave on Θ, this implies that $f_\theta \geq \tilde{\pi}_\theta$ at $\tilde{\theta}(t)$. So $f_\theta \geq 0$ since it must have sign opposite that of $\tilde{\pi}_\theta$. The comparative static results from Theorem 5 implies that $\tilde{\theta}_{\beta_j}(t) \geq 0$. Since in this case $\beta_j \geq 1$ and $\tilde{\theta}_{\beta_j}(t) \geq 0$, it is necessary that the weighted updater will increase their maximizing value of θ at least as much as the perfect Bayesian.

Case 2: $x_t < \tilde{x}(t-1)$. Now the single crossing property on f implies

$$\arg\max_{\theta \in \Theta} f(x_t|\theta) < \tilde{\theta}(t-1).$$

Using arguments analogous to those in *Case 1* entails $\tilde{\theta}_{\beta_j}(t) < 0$. Also, as $\beta_j < 1$ in this case and $\tilde{\theta}_{\beta_j}(t) < 0$, the weighted updater will decrease the maximizing value of θ strictly less than the perfect Bayesian will.

We have determined that it is either the case that the weighted updater will increase the maximized value of θ at least as much or will decrease the maximized value of θ less than the perfect Bayesian will. As it is assumed that they have identical prior distributions $\pi(\theta)$, which are strictly log-concave and therefore have unique and identical maximizers, the weighted updater's maximizing value of θ can never be less than that of the perfect Bayesian. **Q.E.D.**

References

Bagnoli, M., and T. Bergstrom (2005): "Log-Concave Probability and its Applications," *Economic Theory*, 26(2), 445–469.

Bain, R. (2009): "Error and Optimism Bias in Toll Road Traffic Forecasts," *Transportation*, 36(5), 469–482.

Benjamin, D. J., M. Rabin, and C. Raymond (2011): "A Model of Non-Belief in the Law of Large Numbers," *Working Paper available on the Social Science Research Network*.

Boyd, S. P., and L. Vandenberghe (2004): *Convex Optimization*. Cambridge University Press.

Chapin, J., and G. Coleman (2009): "Optimistic Bias: What You Think, What You Know, or Whom You Know," *North American Journal of Psychology*, 11(1), 121–132.

DellaVigna, S. (2009): "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, pp. 315–372.

Gabaix, X., and D. I. Laibson (2008): "The Seven Properties of Good Models," in *The Foundations of Positive and Normative Economics*, pp. 292–299. Oxford University Press.

Grether, D. M. (1980): "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, 95(3), 537–557.

——— (1992): "Testing Bayes' Rule and the Representativeness Heuristic: Some

Experimental Evidence," *Journal of Economic Behavior & Organization*, 17(1), 31–57.

Hirshleifer, D. (2001): "Investor Psychology and Asset Pricing," *Journal of Finance*, 56(4), 1533–1597.

Hoch, S. J. (1985): "Counterfactual Reasoning and Accuracy in Predicting Personal Events.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4), 719–731.

Hogarth, R. M., and H. J. Einhorn (1992): "Order Effects in Belief Updating: The Belief-Adjustment Model," *Cognitive Psychology*, 24(1), 1–55.

Ibrahim, J. G., and M.-H. Chen (2000): "Power Prior Distributions for Regression Models," *Statistical Science*, 15(1), 46–60.

Kahneman, D., and A. Tversky (1973): "On the Psychology of Prediction," *Psychological Review*, 80(4), 237–251.

Koellinger, P., M. Minniti, and C. Schade (2007): "'I Think I Can, I Think I Can': Overconfidence and Entrepreneurial Behavior," *Journal of Economic Psychology*, 28(4), 502–527.

Milgrom, P., and C. Shannon (1994): "Monotone Comparative Statics," *Econometrica*, 62(1), 157–180.

Palfrey, T. R., and S. W. Wang (2012): "Speculative Overpricing in Asset Markets with Information Flows," *Econometrica*, 80(5), 1937–1976.

Quiggin, J. (1988): "Increasing Risk: Another Definition," *paper presented at 4th Conference on Foundations of Utility Research, Budapest.*

Rabin, M. (1996): "Psychology and Economics," *UC Berkeley Working Paper*.

Rabin, M., and J. L. Schrag (1999): "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics*, 114(1), 37–82.

Sethna, J. P. (2006): *Statistical Mechanics: Entropy, Order Parameters, and Complexity*. Oxford University Press.

Shaked, M., and J. G. Shanthikumar (2007): *Stochastic Orders*. Springer.

Shannon, C. E. (1948): "A Mathematical Theory of Communication," *The Bell System Technical Journal*, 27(3), 379–423 and 623–656.

Sharot, T., A. M. Riccardi, C. M. Raio, and E. A. Phelps (2007): "Neural Mechanisms Mediating Optimism Bias," *Nature*, 450(7166), 102–105.

Tribus, M. (1961): *Thermostatics and Thermodynamics: An Introduction to Energy, Information and States of Matter, With Engineering Applications*. Van Nostrand, Princeton, NJ.

Tversky, A., and D. Kahneman (1973): "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, 5(2), 207–232.

Van Benthem, J., J. Gerbrandy, and B. Kooi (2009): "Dynamic Update with Probabilities," *Studia Logica*, 93(1), 67–96.

Weinstein, N. D. (1980): "Unrealistic Optimism About Future Life Events.," *Journal of Personality and Social Psychology*, 39(5), 806–820.

Weinstein, N. D., and W. M. Klein (1996): "Unrealistic Optimism: Present and Future," *Journal of Social and Clinical Psychology*, 15(1), 1–8.