An Information-Theoretic Foundation for the Weighted Updating Model

Jesse Aaron Zinn*
Clayton State University
August 17, 2018

Abstract

Weighted Updating generalizes Bayesian updating, allowing for biased beliefs by weighting the likelihood function and prior distribution with positive real exponents. I provide a rigorous foundation for the model by showing that transforming a distribution by exponential weighting (and normalizing) systematically affects the information entropy of the resulting distribution. For weights greater than one the resulting distribution has less information entropy than the original distribution, and vice versa. As the entropy of a distribution measures how informative a decision maker is treating the underlying observation(s), this result suggests a useful interpretation of the weights. For example, a weight greater than one on a likelihood function models an individual who is treating the associated observation(s) as being more informative than a perfect Bayesian would.

JEL Codes: C02, D03, D83
Keywords: Bayesian Updating, Cognitive Biases, Irrational Expectations, Learning, Uncertainty

*JesseZinn@Clayton.edu, Phone: (678) 466-4536, College of Business, 2000 Clayton State Boulevard, Morrow, GA 30260. I appreciate Ted Bergstrom, Gary Charness, Itzhak Gilboa, Zack Grossman, Botond Kőszegi, Greg Leo, Jason Lepore, Dick Startz and seminar participants at UC Santa Barbara, Cal Poly San Luis Obispo, the 2017 North American Summer Meeting of the Econometric Society, and the 2017 Behavioral Models of Politics conference at Duke University for valuable comments and suggestions.
1 Introduction

The weighted updating model generalizes Bayes’ rule to allow for systematically bi-
ased learning. Despite the fact that this model has seen increased use in economics
and other disciplines, there has not yet been a rigorous justification for it. Those who
apply the model have, heretofore, justified it by appealing to intuition. This paper
eliminates this shortcoming with a result that provides a rationale for using the model
as it has been used.

I show that transforming a distribution increases or decreases the information en-
tropy of the new distribution relative to the original if the transformation is strictly
concave or convex, respectively. In particular, the central result of this paper estab-
ishes that whether the information entropy of a distribution transformed by expo-
ential weighting is greater or less than the original depends on the magnitude of the
exponent. If the weight is greater than one then the information entropy of the new
distribution is less than that of the original distribution, and vice versa.

Each of the distributions constituting Bayes’ rule determines how either empirical
observations or prior information affects beliefs. The information entropies of these
distributions measure how informative the individual regards the associated pieces
of information. Therefore, this result provides the interpretation that weighting is a
parametric method with which to model the treatment of data as either more or less
informative than with Bayesian updating. With this result one can say that weighted
updating embodies a theory of biased judgment, wherein these biases are a result of
incorrectly interpreting the information content of data.

Literature that uses the weighting updated model in one form or another includes
Grether (1980, 1992), who estimates the exponential weights on the likelihood func-
tion and the prior distribution to find empirical evidence for the representativeness
heuristic. Ibrahim and Chen (2000) introduces power priors, which allows the researcher to consider data from previous studies by putting a weight in (0, 1) on the likelihood function for that data and using a weight of 1 with current data. Van Benthem et al. (2009) define a weighted product updating rule and go on to prove that Bayes’ rule and the Jeffrey updating rule are both special cases. Palfrey and Wang (2012) use weighted updating to model investor under- and overreaction to public information about financial assets in a model with speculative pricing. Benjamin et al. (2015) use the weighted updating model to study non-belief in the law of large numbers. March (2016) uses weighted updating to model adaptive social learning, where economic agents make inferences based on the previous actions of others. In March and Ziegelmeier (2018) “intuitive” agents overweight public information in a model with excessive herding behavior.

2 The Weighted Updating Model

A decision maker will consider an observation (or sequence of observations) $x$ as an outcome from a stochastic process with probability density function $f(x|\theta)$, where $\theta$ is an unknown parameter that the decision maker considers to be from parameter space $\Theta$. Bayesian beliefs regarding the value of $\theta$ after observing $x$ are completely described by the posterior distribution $\pi(\theta|x)$. If we let $(\Theta, \mathcal{A}, m)$ be a measure space and denote the likelihood function with $f(x|\theta)$ and the prior distribution with $\pi(\theta)$, then Bayes’ rule states

$$
\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta) dm(\theta)}.
$$

Weighted updating augments Bayes’ rule with parameters $\alpha$ and $\beta$ as exponents
respectively on the prior probability distribution and likelihood function. Denote the posterior distribution under weighted updating after observing \( x \) by \( \tilde{\pi}(\theta|x) \). Then the weighted updating model is given by

\[
\tilde{\pi}(\theta|x) = \frac{f(x|\theta)\beta \pi(\theta)\alpha}{\int_{\Theta} f(x|\theta)\beta \pi(\theta)\alpha \, dm(\theta)}.
\] (1)

Both Bayes’ rule and the weighted updating model can be stated without mention of the marginal distribution, which is not a function of \( \theta \) and serves only as a normalizing factor, ensuring that the posterior distribution aggregates to one over its support.\(^1\) Thus, the weighted updating model can be displayed as

\[
\tilde{\pi}(\theta|x) \propto f(x|\theta)\beta \pi(\theta)\alpha.
\] (1')

Stating the model as in expression (1') emphasizes how the nature of the posterior distribution depends entirely on the interaction between the prior distribution and the likelihood function, and how the weights \( \alpha \) and \( \beta \) affect this interaction.

Zinn (2015) expands the weighting updating model to allow each observation \( x_i \) to have its own weight \( \beta_i \), which allows for biases where different observations are treated differently.

---

\(^1\)Throughout the paper I use the word “support” as shorthand for “the set of inputs where a distribution takes positive values.” This is somewhat non-standard, though close in spirit to the usual definition of the word.

I also assume all functions are measurable and integrable so that integrals are finite and well-defined. This assumption includes the functions generated by exponential weighting. In many cases, this assumption is innocuous because weighting a distribution with an exponent and rescaling results in a distribution from the original family, so integrability follows.
In this section, I investigate how exponentially weighting a distribution affects the resulting distribution relative to the original. Note that I change the notation from what is used in the previous section in an effort to make it clear that the following results apply to the likelihood function(s) and the prior distribution individually. Likewise, results involving the weight $\gamma$ can be interpreted as true for $\alpha$, $\beta$, and each $\beta_i$.

Let $(\Omega, S, M)$ be a measure space and consider the transformation

$$g(\omega) \mapsto \frac{g(\omega)^\gamma}{\int_{\Omega} g(\omega)^\gamma \, dM(\omega)}.$$  \hspace{1cm} (2)

If $\gamma > 0$ then (2) is a strictly increasing transformation in the sense that

$$g(\omega_1) > g(\omega_2) \Rightarrow \frac{g(\omega_1)^\gamma}{\int_{\Omega} g(\omega)^\gamma \, dM(\omega)} > \frac{g(\omega_2)^\gamma}{\int_{\Omega} g(\omega)^\gamma \, dM(\omega)}.$$

Therefore, the value(s) that maximize (or minimize) $g$ and the distribution proportional to $g^\gamma$ are identical when $\gamma > 0$. In other words, transformation (2) is necessarily mode-preserving. Note, however, that such a transformation is not necessarily mean-preserving, as is likely to be the case when $g$ is asymmetric.

Most relevant for this work is that the exponent $\gamma$ affects how concentrated or dispersed the resulting distribution is. The following definition precisely describes what I mean by “concentrated” and “dispersed.”

**Definition 1** (Monotone Dispersion, Monotone Concentration). For two non-uniform probability distributions $\Gamma$ and $g$ on the same support $\Omega$, $\Gamma$ is a *monotone dispersion*

---

2Note that “monotone dispersion” is distinct from “monotone spread” (See Quiggin (2012)). One difference between a monotone dispersion and a monotone spread is that the former is necessarily mode-preserving while the latter is necessarily mean-preserving, but not vice-versa.
of $g$ if for all pairs $(\omega_1, \omega_2) \in \Omega^2$ the following two conditions hold:

$$g(\omega_1) > g(\omega_2) \iff \Gamma(\omega_1) > \Gamma(\omega_2), \text{ and}$$

$$g(\omega_1) > g(\omega_2) \implies \frac{g(\omega_1)}{g(\omega_2)} > \frac{\Gamma(\omega_1)}{\Gamma(\omega_2)}. \quad (4)$$

If $\Gamma$ is a monotone dispersion of $g$ then $g$ is a \textit{monotone concentration of} $\Gamma$.\(^3\)

See Figure 1 for an example of two distributions that are related to each other through monotone dispersion and concentration. Note that condition (3) requires that the transformations $g \mapsto \Gamma$ and $\Gamma \mapsto g$ are strictly increasing functions. Another consequence of (3) is that\(^4\)

$$g(\omega_1) = g(\omega_2) \iff \Gamma(\omega_1) = \Gamma(\omega_2).$$

This condition and the monotonicity of such transformations make it clear that two agents with beliefs that are related by monotone dispersion and concentration will agree on a rank ordering of events according to their likelihoods as given by their respective beliefs, even if they do not agree on the specific values of the probabilities. This ensures that any two distributions related by monotone concentration and dispersion will have similar ordinal properties.

Expression (4) describes how the cardinal properties of a monotone dispersion and concentration differ, with a monotone dispersion being closer to a uniform distribution and a monotone concentration having more extreme values for both highs and lows.

\(^3\)Uniform distributions are excluded from Definition 1 because if either $g$ or $\Gamma$ were uniform then the other would necessarily be uniform by condition (3), so they would be the same distribution. If this is the case then condition (4) is only \textit{vacuously} true, which is not useful for our purposes because condition (4) provides an asymmetry that allows one to compare different distributions. Excluding uniform distributions ensures that there is no case in which the relations “is a monotone dispersion of” and “is a monotone concentration of” are symmetric or reflexive.

\(^4\)My thanks to an anonymous referee for pointing this out.
Figure 1: $\Gamma$ is a monotone dispersion of $g$. Equivalently, $g$ is a monotone concentration of $\Gamma$.

\[ g(\omega), \Gamma(\omega) \]

The following theorem shows how the concepts of monotone dispersion and concentration are intimately related to concave and convex transformations.\(^5\)

**Theorem 1.** Let $g : \Omega \rightarrow \mathbb{R}^+$ be any non-uniform probability distribution. Let $T : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a strictly increasing, differentiable function such that $T(0) = 0$ and $T(g)$ is a distribution. If $T$ is strictly concave then $T(g)$ is a monotone dispersion of $g$ and if $T$ is strictly convex then $T(g)$ is a monotone concentration of $g$.

Because exponentiating (whether normalizing or not) is a strictly concave transformation when the exponent is less than one and strictly convex when the exponent is greater than one, Theorem 1 has the following corollary.

**Corollary 1.** Let $g : \Omega \rightarrow \mathbb{R}^+$ be any non-uniform probability distribution. If $\gamma \in (0, 1)$ then

\[
g(\omega)^\gamma \int_{\Omega} g(\omega)^\gamma dM(\omega),
\]

is a monotone dispersion of $g$. If $\gamma > 1$ then (5) is a monotone concentration of $g$.

\(^5\)Proofs can be found in the appendix.
Corollary 1 demonstrates that when a distribution is weighted with a positive power other than one and normalized, the resulting distribution is either a monotone dispersion or concentration of the original distribution depending on whether the weight is less than or greater than one.

4 Information Entropy as a Measure of Dispersion

In this section, I study how the notions of dispersion and concentration from Definition 1 can be measured by a distribution’s information entropy. What immediately follows is a brief discussion of information entropy. I will then illustrate its importance for the weighted updating model.

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

\[
H(g) \equiv -\int_{\Omega} g(\omega) \log g(\omega) \, dM(\omega).
\]

The logarithm ensures that information entropy is additive in the densities of independent random variables. This is because for any two independent random variables \( Y \) and \( Z \) respectively distributed \( g_Y \) and \( g_Z \) and particular pair of events \((y, z)\),

\[- \log g_Y(y) g_Z(z) = - \log g_Y(y) - \log g_Z(z).\]

Tribus (1961) dubbed \(- \log g(\omega)\) the surprisal of \( \omega \) for any distribution \( g \) and particular \( \omega \in \Omega \).\(^6\) Because \(- \log g(\omega)\) is decreasing in \( g(\omega) \), surprisal is greater for \( \omega \) which (according to \( g \)) are less likely and, therefore, more surprising outcomes. The information entropy of a distribution is equivalent to the expected surprisal,

\(^6\)– \( \log g(\omega) \) is also known as the self-information of \( \omega \).
as information entropy is the expected value of the surprisal of a distribution. 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 greater information entropy will, on average, generate observations that have less information content, and vice versa.\footnote{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.}

The following theorem verifies the claim that the information entropy of a monotone dispersion is greater than the information entropy of a monotone concentration.

**Theorem 2.** If $\Gamma$ is a monotone dispersion of $g$ then $H(\Gamma) \geq H(g)$. If, in addition,

$$M(\{\omega : \Gamma(\omega) \neq r\}) > 0$$

for some $r \in [b, B]$, where

$$b \equiv \sup \Gamma(\{\omega : g(\omega) \leq \Gamma(\omega)\}) \in \mathbb{R}$$

and

$$B \equiv \inf \Gamma(\{\omega : g(\omega) \geq \Gamma(\omega)\}) \in \mathbb{R},$$

then $H(\Gamma) > H(g)$.

Together, Theorems 1 and 2 imply that if $g$ and $T(g)$ are two non-uniform distributions and $T$ is a strictly concave then $H(T(g)) \geq H(g)$, and $H(T(g)) \leq H(g)$ if $T$ is strictly convex.
The following corollary is the main result of this paper. It is essentially the conclusion of a syllogism with Corollary 1 and Theorem 2 as premises.

**Corollary 2.** If $g$ and $\frac{g(\omega)^\gamma}{\int_\Omega g(\omega)^\gamma \, dM(\omega)}$ are distributions for some $\gamma > 0$ then

\[
H(g) > H \left( \frac{g(\omega)^\gamma}{\int_\Omega g(\omega)^\gamma \, dM(\omega)} \right) \iff \gamma > 1, \text{ and }
\]
\[
H(g) < H \left( \frac{g(\omega)^\gamma}{\int_\Omega g(\omega)^\gamma \, dM(\omega)} \right) \iff \gamma < 1.
\]

Corollary 2 provides a rationale and rigorous interpretation of the weighted updating model. Consider expression (1). Corollary 2 implies that if $\alpha > 1$ then the prior information is being overemphasized and treated as thought it contains more information content than it actually has, and the opposite holds if $\alpha < 1$. The same interpretation holds for $\beta$ with respect to the how informative $x$ is treated. So if, for example, $\alpha > 1$ and $\beta < 1$ then the beliefs represented by that model are made in a manner in which the prior information is treated with more information content than it actually has and $x$ is treated as though it has too little information content.

## 5 Concluding Remarks

In this paper, I provide an interpretation of weighted updating as a method of modelling individuals who treat information as either more or less informative than under Bayes’ rule. In particular, I show that weighting the functions primitive to Bayes’ rule transforms the functions by monotone dispersion or monotone concentration, and that these transformations systematically affect the information entropies of the resulting likelihood function(s) and prior probability distribution.

This interpretation of weighting a distribution suggests that, on its own, weighted updating may be appropriate to model only those biases in which individuals cor-
rectly 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\(^8\) or the cognitive limitations of utilizing correctly interpreted data. However, 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, it is entirely possible that there are multiple biases simultaneously affecting belief formation. One could, for example, model an individual who misinterprets evidence using the framework of Rabin and Schrag (1999) and then processes the misinterpreted information irrationally using weighted updating.

**Appendix: Proofs**

*Proof of Theorem 1.* Condition (3) is satisfied immediately because \(T\) is strictly increasing. The fact that \(T\) is strictly increasing and \(g\) is non-uniform implies that \(T(g)\) is non-uniform as well, and the only thing left to show is that (4) is satisfied. Let \(T\) be strictly concave. Because \(g\) is non-uniform there exists \(\omega_1\) and \(\omega_2\) such that \(g(\omega_1) > g(\omega_2) > 0\). To economize on notation let \(g_1 = g(\omega_1)\) and \(g_2 = g(\omega_2)\). Strict concavity and differentiability of \(T\) imply

\[
g_1 > g_2 + \frac{1}{T'(g_2)}[T(g_1) - T(g_2)],
\]

where \(T'\) denotes the derivative of \(T\). Because \(g_2 > 0\), dividing expression (5) by \(g_2\) does not change the direction of the inequality, so

\[
\frac{g_1}{g_2} > 1 + \frac{1}{g_2T'(g_2)}[T(g_1) - T(g_2)]. \tag{6}
\]

\(^8\)Self-deception typically involves individuals who downplay or overemphasize the importance of certain pieces of evidence in a systematic way (Hirshleifer, 2001).
The fundamental theorem of calculus and \( T(0) = 0 \) imply

\[
T(g_2) = \int_0^{g_2} T'(x) \, dx. \tag{7}
\]

Because \( T \) is strictly concave \( T'(x) > T'(g_2) \) for all \( x < g_2 \), so

\[
\int_0^{g_2} T'(x) \, dx > \int_0^{g_2} T'(g_2) \, dx = g_2 T'(g_2). \tag{8}
\]

Expressions (7) and (8) imply \( T(g_2) > g_2 T'(g_2) \). Consequently,

\[
1 + \frac{1}{g_2 T'(g_2)} [T(g_1) - T(g_2)] > 1 + \frac{1}{T(g_2)} [T(g_1) - T(g_2)]. \tag{9}
\]

Combining expressions (6) and (9) yields

\[
\frac{g_1}{g_2} > 1 + \frac{1}{T(g_2)} [T(g_1) - T(g_2)] = 1 + \frac{T(g_1)}{T(g_2)} - \frac{T(g_2)}{T(g_2)} = \frac{T(g_1)}{T(g_2)}, \tag{10}
\]

satisfying condition (4) so \( T(g) \) is a monotone dispersion of \( g \). The result for a strictly convex \( T \) follows from the fact that if \( T \) is strictly convex then \( T^{-1} \) is strictly concave,\(^9\) so the logic in the proof for a strictly concave transformation implies \( g = T^{-1}(T(g)) \) is a monotone dispersion of \( T(g) \), which is equivalent to \( T(g) \) being a monotone concentration of \( g \).

Q.E.D.

The proof of Theorem 2 utilizes the following four lemmas.

\(^9\)Note that \( T^{-1} \) exists and is strictly increasing because \( T \) is strictly increasing.
Lemma 1. If \( g \) and \( \Gamma \) are distributions with the same support \( \Omega \) then both of the sets
\[
\{ \omega : g(\omega) \leq \Gamma(\omega) \} \quad \text{and} \quad \{ \omega : g(\omega) \geq \Gamma(\omega) \}
\]
are nonempty.\(^{10}\)

Proof. Suppose \( \{ \omega : g(\omega) \leq \Gamma(\omega) \} = \emptyset \). Then \( g(\omega) > \Gamma(\omega) \) for all \( \omega \in \Omega \). This implies
\[
\int_{\Omega} g(\omega) \, dM(\omega) > \int_{\Omega} \Gamma(\omega) \, dM(\omega),
\]
which, in turn, implies that \( 1 > 1 \). This contradiction means
\[
\{ \omega : g(\omega) \leq \Gamma(\omega) \} \neq \emptyset.
\]
Showing that \( \{ \omega : g(\omega) \geq \Gamma(\omega) \} \neq \emptyset \) follows by simply transposing \( g \) and \( \Gamma \). \( \text{Q.E.D.} \)

Recall from Theorem 2 the following definitions which are used in the next several results:
\[
b \equiv \sup \Gamma(\{ \omega : g(\omega) \leq \Gamma(\omega) \}) \in \mathbb{R}.
\]
\[
B \equiv \inf \Gamma(\{ \omega : g(\omega) \geq \Gamma(\omega) \}) \in \mathbb{R}.
\]

Lemma 2. If \( g \) is a monotone concentration of \( \Gamma \) then \( b, B \in \mathbb{R} \).

Proof. Due to Lemma 1, there exist \( \omega_1 \in \{ \omega : g(\omega) \leq \Gamma(\omega) \} \) and \( \omega_2 \in \{ \omega : g(\omega) \geq \Gamma(\omega) \} \). Suppose \( \Gamma(\{ \omega : g(\omega) \leq \Gamma(\omega) \}) \) is not bounded above. Then, we can choose \( \omega_1 \) such that \( \Gamma(\omega_1) > g(\omega_2) \). To summarize, we have
\[
\Gamma(\omega_1) > g(\omega_2) \geq \Gamma(\omega_2) > 0
\]

\(^{10}\)I doubt that this result is original to this work. I include it, however, for sake of completeness.
and
\[ \Gamma(\omega_1) \geq g(\omega_1). \]

As \( g \) is a monotone concentration of \( \Gamma \), condition (3) stipulates that
\[ \Gamma(\omega_1) > \Gamma(\omega_2) \iff g(\omega_1) > g(\omega_2). \]

Therefore,
\[ \Gamma(\omega_1) \geq g(\omega_1) > g(\omega_2) \geq \Gamma(\omega_2) > 0. \]

This implies
\[ \frac{\Gamma(\omega_1)}{\Gamma(\omega_2)} > \frac{g(\omega_1)}{g(\omega_2)} \quad \text{and} \quad g(\omega_1) > g(\omega_2), \]
contradicting condition (4), which holds since \( g \) is a monotone concentration of \( \Gamma \).

Therefore, \( \Gamma(\omega_1) \leq g(\omega_2) \) and, since \( \omega_1 \) was chosen arbitrarily, this holds for all \( \omega_1 \in \{ \omega : g(\omega) \leq \Gamma(\omega) \} \). Hence, \( g(\omega_2) \) is an upper bound of \( \Gamma(\{ \omega : g(\omega) \leq \Gamma(\omega) \}) \subset \mathbb{R} \).

The completeness of \( \mathbb{R} \), thereby, implies that \( b \in \mathbb{R} \). That \( B \in \mathbb{R} \) can be shown similarly, though it can be a bit simpler because \( \Gamma(\{ \omega : g(\omega) \geq \Gamma(\omega) \}) > 0 \). \textbf{Q.E.D.}

**Lemma 3.** If \( g \) is a monotone concentration of \( \Gamma \) then \( b \leq B \).

**Proof.** Suppose that \( B < b \). Then, use the definition of the supremum of a set to establish that there exists \( \omega_1 \in \Omega \) for which
\[ g(\omega_1) < \Gamma(\omega_1) \quad \text{and} \quad \frac{b + B}{2} < \Gamma(\omega_1) \leq b. \quad (11) \]

Similarly, the definition of the infimum of a set guarantees that there exists \( \omega_2 \in \Omega \) such that
\[ \Gamma(\omega_2) < g(\omega_2) \quad \text{and} \quad B \leq \Gamma(\omega_2) < \frac{b + B}{2}. \quad (12) \]
Expressions (11) and (12) imply that $\Gamma(\omega_2) < \Gamma(\omega_1)$, which is true if and only if $g(\omega_2) < g(\omega_1)$ since $g$ is a monotone concentration of $\Gamma$. Therefore,

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

which, as shown in the proof to Lemma 2, is inconsistent with condition (4). Therefore $b \leq B$. \textbf{Q.E.D.}

**Lemma 4.** If $g$ is a monotone concentration of $\Gamma$ then there exists $r \in \mathbb{R}$ such that

$$\Gamma(\omega) > r \implies g(\omega) > \Gamma(\omega) \quad (13)$$

and

$$\Gamma(\omega) < r \implies g(\omega) < \Gamma(\omega). \quad (14)$$

**Proof.** I prove (13), as the proof of (14) is essentially identical. Lemma 3 guarantees that $[b, B] \neq \emptyset$, so we may consider any $r \in [b, B]$. If $\Gamma(\omega_0) > r$ for some $\omega_0 \in \Omega$ then

$$b \leq r < \Gamma(\omega_0).$$

$\Gamma(\omega_0) > b$ implies that $\omega_0 \notin \{\omega : g(\omega) \leq \Gamma(\omega)\}$, so $g(\omega_0) > \Gamma(\omega_0)$. \textbf{Q.E.D.}

**Proof of Theorem 2.** Gibb’s inequality implies\textsuperscript{11}

$$\int_\Omega g(\omega) \log \Gamma(\omega) \, dM(\omega) \leq \int_\Omega g(\omega) \log g(\omega) \, dM(\omega),$$

\textsuperscript{11}Gibbs’ inequality is a well known fact from statistical physics that is a consequence of Jensen’s inequality, which applies because log functions are concave.
which can also be expressed

$$\int_{\Omega} g(\omega) \log \Gamma(\omega) \, dM(\omega) \leq -H(g).$$  \hfill (15)

Add $H(\Gamma)$ to both sides of expression (15) to find that

$$H(\Gamma) + \int_{\Omega} g(\omega) \log \Gamma(\omega) \, dM(\omega) \leq H(\Gamma) - H(g),$$

implying that

$$\int_{\Omega} [g(\omega) - \Gamma(\omega)] \log \Gamma(\omega) \, dM(\omega) \leq H(\Gamma) - H(g),$$  \hfill (16)

Lemma 3 asserts that $[b, B]$ is non-empty, so we can choose any $r \in [b, B]$ and write:

$$-\log r \int_{\Omega} g(\omega) - \Gamma(\omega) \, dM(\omega) = 0,$$  \hfill (17)

which is true because $g$ and $\Gamma$ are both distributions. Adding expressions (16) and (17) and some rearranging yields

$$\int_{\Omega} [g(\omega) - \Gamma(\omega)](\log \Gamma(\omega) - \log r) \, dM(\omega) \leq H(\Gamma) - H(g).$$  \hfill (18)

By Lemma 4, $r \in [b, B]$ implies that $\log \Gamma(\omega) - \log r$ and $g(\omega) - \Gamma(\omega)$ have the same sign, so the left-hand side of expression (18) is non-negative. Therefore, $H(g) \leq H(\Gamma)$.

If, additionally, $M(\{\omega \in \Omega : \Gamma(\omega) \neq r\}) > 0$ then the integral constituting the left-hand side of expression (18) can be decomposed into

$$\int_{\{\omega: \Gamma(\omega) \neq r\}} [g(\omega) - \Gamma(\omega)](\log \Gamma(\omega) - \log r) \, dM(\omega)$$
plus
\[ \int_{\{ \omega: \Gamma(\omega) = r \}} [g(\omega) - \Gamma(\omega)](\log \Gamma(\omega) - \log r) \, dM(\omega). \]

The former, by Lemma 4, is strictly positive and the latter equals zero, so \( H(g) < H(\Gamma) \). 

\[ \text{Q.E.D.} \]

References

**Benjamin, D. J., M. Rabin, and C. Raymond** (2015): “A Model of Non-Belief in the Law of Large Numbers,” *Journal of the European Economic Association*, 14, 515–544.

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

——— (1992): “Testing Bayes’ Rule and the Representativeness Heuristic: Some Experimental Evidence,” *Journal of Economic Behavior & Organization*, 17, 31–57.

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

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

**March, C.** (2016): “Adaptive Social Learning,” *CESifo Working Paper No. 5783*.

**March, C. and A. Ziegelmeier** (2018): “Excessive Herding in the Laboratory: The Role of Intuitive Judgments,” *CESifo Working Paper No. 6855*. 
Palfrey, T. R. and S. W. Wang (2012): “Speculative Overpricing in Asset Markets with Information Flows,” *Econometrica*, 80, 1937–1976.

Quiggin, J. (2012): *Generalized Expected Utility Theory: The Rank-Dependent Model*, Springer Science & Business Media.

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

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

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

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

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

Zinn, J. A. (2015): “Expanding the Weighted Updating Model,” *Economics Bulletin*, 35, 182–186.