# Modelling Biased Judgement with Weighted Updating

Zinn, Jesse

Clayton State University

30 September 2013

# Modelling Biased Judgement with Weighted Updating

Jesse Aaron Zinn*

Clayton State University

January 16, 2015

**Abstract**

The weighted updating model is a generalization of Bayesian updating that allows for biased beliefs by weighting the constituent functions of Bayes' rule with real exponents. In this paper I show that transforming a distribution by exponential weighting and normalization systematically affects the information entropy of the resulting distribution. Specifically, if the weight is greater then one then the resulting distribution has less information entropy than the original distribution (and vice versa). This result provides a useful interpretation of the model, since, for example a likelihood function with greater entropy translates to the associated data being treated with less information content. The result also justifies using the model as it has been used in the literature, i.e. to model biases in which individuals treat observations as being either more or less informative than they should.

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

*Contact: JesseZinn@clayton.edu. I appreciate Ted Bergstrom, Gary Charness, Zack Grossman, Jason Lepore, and Dick Startz for valuable comments and suggestions.

# 1 Introduction

The weighted updating model generalizes Bayes' Rule to allow for biased learning. Despite the fact that this model has seen some use in economics and other disciplines, there has not yet been a rigorous interpretation of the model or justification for using it. Those who use the model justify its use by appealing to intuition. This paper eliminates this shortcoming, by showing that transforming a distribution by weighting it and normalizing systematically affects the information entropy of the resulting distribution relative to that of the original distribution.

By weighting and normalizing a single distribution, the entropy of the resulting distribution decreases or increases relative to the original depending on whether the weight is greater or less than one. This 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. 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.

Work that utilizes the weighted updating model includes Grether (1980) and Grether (1992), which estimate the exponential weights on the likelihood function and the prior distribution to find empirical evidence for the representativeness heuristic. Ibrahim and Chen (2000) introduced *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 putting a weight of 1 for current data. Van Benthem, Gerbrandy, and Kooi (2009) define a "weighted product updating rule" and go on to prove that Bayes' rule and the Jeffrey updating rule are both special cases of their rule. Palfrey and Wang (2012) use weight updating to model investor under- and overreaction to public information about financial assets in a

model with speculative pricing. Benjamin, Rabin, and Raymond (2013) use the weighted updating model to study "non-belief in the law of large numbers". Zinn (2014) expands upon the basic model to allow the weights to change over time and for individuals to discriminate between observations.

## 2 Interpreting the Weights

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 $\theta$ is an unknown parameter that the decision maker considers to be from parameter space $\Theta$. Bayesian beliefs regarding the value of $\theta$ after observing $h_t$ 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

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

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

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

Both Bayes' rule and the weighted updating model can be stated without mention of the marginal distribution, which is not a function of $\theta$ and serves only as a normalization, ensuring that the posterior distribution aggregates to one over

its support.[1] Thus, the weighted updating model can be displayed as

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

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

## 2.1 Monotone Concentration and Monotone Dispersion

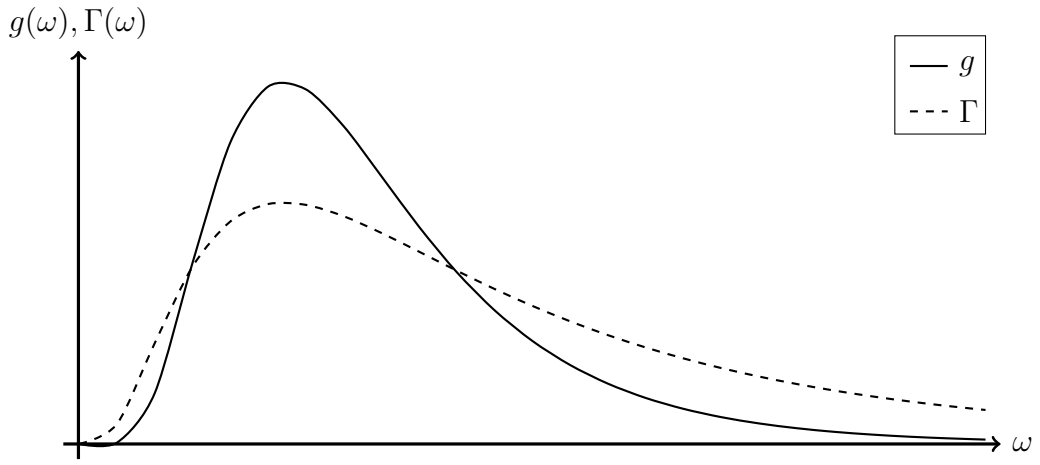
Consider how an exponent $\gamma$ transforms a single probability distribution $g(y)$ to another proportional to $g(y)^{\gamma}$. As long as $\gamma > 0$ taking $g(y)$ to the power $\gamma$ for all $y$ is a monotone transformation, as is dividing by the resulting marginal distribution, which is always positive. As such, the values of $y$ that maximize (or minimize) $g$ and $g^{\gamma}$ are identical. What the exponent $\gamma$ affects is the concentration of the resulting distribution. The following definition describes this notion precisely.[2]

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

[1]Note that throughout the paper it is assumed that $\int_{\Theta} f(h_t|\theta)^{\beta}\pi(\theta)^{\alpha}\, d\theta$ 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.

[2]In a previous version of this paper, the concept of a monotone dispersion was given the name "monotone spread", a related concept due to Quiggin (1988). Note that a monotone dispersion differs from a monotone spread in that the latter is necessarily mean-preserving.

Figure 1: $\Gamma$ is a monotone dispersion of $g$, $g$ is a monotone concentration of $\Gamma$.




*dispersion of* $g$ if for all pairs $(\omega_1, \omega_2) \in \Omega^2$ it is true that

$$g(\omega_1) = g(\omega_2) \quad \Leftrightarrow \quad \Gamma(\omega_1) = \Gamma(\omega_2), \tag{2}$$

$$g(\omega_1) > g(\omega_2) \quad \Leftrightarrow \quad \Gamma(\omega_1) > \Gamma(\omega_2), \text{ and} \tag{3}$$

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

If $\Gamma$ is a monotone dispersion of $g$ then $g$ is a *monotone concentration of* $\Gamma$.[3]

See Figure 1 for an example of two distributions that are a monotone dispersion and concentration of one another.

Interest in monotone dispersions and concentrations is due to the fact that when a distribution is weighted with a positive power and normalized, the result-

[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 (2), so they would be the same distribution. If this is the case then conditions (3) and (4) are only *vacuously* true, which is not useful for our purposes because condition (4) provides an asymmetry that allows one to compare different distributions. Another way of saying this is that such a restriction ensures that there are no case in which the relations "is a monotone dispersion of" and "is a monotone concentration of" are symmetric.

ing distribution is either a monotone dispersion or concentration of the original distribution depending on whether the weight is less than or greater than one, as stated in the following theorem.[4]

**Theorem 1.** 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 $\Gamma$ is a monotone concentration of $g$.

Theorem 1 guarantees that a positively weighted distribution results in either a monotone dispersion or concentration of the original, but what does it mean for two distributions to be related in this way? Looking at Definition 1, expression (3) implies that $\Gamma(\omega_1) > \Gamma(\omega_2)$ whenever $g(\omega_1) > g(\omega_2)$, so if $\Gamma$ is a monotone dispersion of $g$ then the transformation $g \mapsto \Gamma$ is a monotone transformation. This monotonicity ensures that the ordinal properties are identical within pairs of distributions that are dispersions and concentrations of each other. In other words, two agents with beliefs that are related by monotone dispersion or concentration will agree on a rank ordering of events according to their likelihoods as given by their respective beliefs.

Expression (4) describes how the cardinal properties of a monotone dispersion or concentration differ from the original function, with a monotone dispersion being closer to a uniform distribution. Equivalently, a concentration is an exaggeration of the original distribution, with "higher highs" and "lower lows". The following theorem states these notions rigorously.

[4]Proofs for all results are in the appendix.

**Theorem 2.** Let $\Gamma$ 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[5] to Theorem 2 is used in the proof of Theorem 3, which is stated below.

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

## 2.2 Measuring 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 $\mu$ and variance $\sigma^2$. It is straightforward to find that taking this distribution to the power $\gamma > 0$ results in a function that is proportional to the normal distribution with mean $\mu$ and variance $\sigma^2/\gamma$. After doing this manipulation one can utilize Theorem 1 to note that $\gamma < 1$ generates a monotone dispersion of the original distribution with greater variance, while a monotone concentration with less variance is the outcome if $\gamma > 1$.

Despite being true for the normal distribution, it is not the case for all distributions that a monotone dispersion implies greater variance and that a monotone

[5]The appendix contains a proof of this Corollay that is independent of Theorem 2.

concentration has less variance. 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, it is strictly convex with peaks at the extremes of the support $x = 0$ and 1. 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 the beta distribution $B(3/4, 3/4)$ which has a variance of $1/10$. Applying the monotone concentration of raising this distribution to the power $\gamma = 2$ and normalizing yields $B(1/2, 1/2)$, which has a variance of $1/8$. Thus, $B(1/2, 1/2)$ is a monotone concentration of $B(3/4, 3/4)$, yet it has a greater variance, providing a counter-example against the general statement that a monotone concentration has less variance (and vice-versa).

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 relation with 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[6]

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

For any distribution $g$ and particular $\omega \in \Omega$, Tribus (1961) dubbed $-\log g(\omega)$ the *surprisal* of $\omega$. 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 logarithm ensures that surprisal is additive in the densities of independent random variables, because 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 $\Omega$. Distributions with higher entropy then can be interpreted as having higher expected surprisal. If outcomes from one distribution are, on av-

[6]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 $-1/2 \log_2(1/2) - 1/2 \log_2(1/2) = 1$ 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. 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.

erage, 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.[7]

The following theorem verifies the claim that transforming a distribution by monotone dispersion results in an increase in entropy and that monotone concentration decreases entropy.

**Theorem 3.** Let $\Gamma$ be a monotone dispersion of $g$. Then the entropy of $\Gamma$ 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 Conclusion

This paper provides 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.

[7]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.

This 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[8] 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.

# Appendix

*Proof of Theorem 1.* Let $\gamma \in (0,1)$. Conditions (2) and (3) 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

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

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

Dividing both the numerator and denominator of the center term by the normalizing 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 $\Gamma$ is a monotone dispersion of $g$. The case for $\gamma > 1$ yielding a monotone concentration is proved analogously. **Q.E.D.**

*Proof of Theorem 2.* Let

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

As $\Gamma$ 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 $\Gamma$ is a monotone dispersion of $g$ implies that both $\Gamma$ 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 expression (3) from Definition 1 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 $\omega_0$, integrating over $\Omega$ yields

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

As both $g$ and $\Gamma$ 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 3 requires the following two lemmas and a fact (Gibb's Inequality) from statistical physics.

**Lemma 1.** Let $\Gamma$ 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 $\Gamma$ is positive on its support $\Omega$ 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 $\Gamma$ is a monotone dispersion of $g$, expression (4) in particular. Therefore it must be the case that $b \leq B$. **Q.E.D.**

**Lemma 2.** Let $\Gamma$ be a monotone dispersion of $g$ and let there exist some $\omega_j$ 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 $\omega_j$ 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 3.* By Gibbs' Inequality

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

which implies

$$\int_\Omega g(\omega) \log g(\omega) - \Gamma(\omega) \log \Gamma(\omega)\, d\omega \geq \int_\Omega (g(\omega) - \Gamma(\omega)) \log \Gamma(\omega)\, d\omega. \tag{5}$$

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

$$0 = -\log r \int_\Omega g(\omega) - \Gamma(\omega)\, d\omega. \tag{6}$$

Adding expressions (5) and (6) gives

$$\int_\Omega g(\omega) \log g(\omega) - \Gamma(\omega) \log \Gamma(\omega)\, d\omega \geq \int_\Omega [g(\omega) - \Gamma(\omega)](\log \Gamma(\omega) - \log r)\, d\omega. \tag{7}$$

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 (7) is non-negative. And so,

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

If, additionally, $\{\omega : g(\omega) > \Gamma(\omega)\}$ or $\{\omega : g(\omega) < \Gamma(\omega)\}$ have positive measure then the right-hand side of expression (7) is strictly positive, so inequality (8) is strict. **Q.E.D.**

# References


Benjamin, D. J., M. Rabin, and C. Raymond (2013): "A Model of Non-Belief in the Law of Large Numbers," *Oxford University Department of Economics Discussion Paper No. 672*.

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.

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

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., 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.

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

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

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

Zinn, J. A. (2014): "Expanding the Weighted Updating Model," *Available at SSRN*.