An Axiomatic Approach to the Law of Small Numbers*

Jawwad Noor and Fernando Payró

June 24, 2023

Abstract

Studies show that people's beliefs about randomness are systematically misspecified – for instance they do not expect streaks to persist. In a canonical coin-tossing environment, this paper models such "streak aversion" in terms of a belief in Mean Reversion along the sequence of outcomes. Beliefs that exhibit Mean Reversion can be represented as if the bias of the coin is path-dependent and self-correcting. Consistent with other findings, such beliefs may fail the Law of Large Numbers. In the setting of Bayesian inference, Mean Reversion ensures that the agent never rules out the true parameter. In an evolutionary setting, Mean Reversion agents are never pushed out of the evolutionary race by standard agents who correctly understand randomness. This paper suggests several directions for both theoretical and empirical investigation of beliefs about randomness.

JEL Classification: D01, D9.

Keywords: Law of Small Numbers, Belief Biases, Heuristics, Gambler's Fallacy, Learning, Misspecified beliefs, Evolution.

1 Introduction

A large number of empirical studies show that people do not understand the nature of randomness, in that their static beliefs about the outcomes of an i.i.d. random process are systematically misspecified. Two well-known findings are:

Gambler's Fallacy: Subjects believe that the probability of tails is higher following a streak of heads: for instance, $P(HHHT) > P(HHHH)$. While there is a large experimental literature establishing the Gambler's Fallacy (see Benjamin (2019) for a comprehensive review), there is also substantial evidence from the field. Studying betting on lottery numbers, Terrell (1994) demonstrates that people are significantly less likely to bet on a lottery number if it was recently a winning number (see Suetens et al (2016) for a more recent study). Using data from a field experiment,

*Noor is at the Department of Economics, Boston University, 270 Bay State Road, Boston MA 02215. Email: jnoor@bu.edu. Payró is at the Department of Economics and Economic History, Universitat Autonoma de Barcelona, Barcelona School of Economics and Center for the Study of Organizations and Decisions in Economics, Campus de la UAB, Edifici B, Bellaterra, 08193. Email: fernando.payro@uab.cat. The authors would like to thank seminar audiences at Boston University, UPenn, U of Edinburgh, Manchester, Durham, LSE, U Autonoma, U Pompeu Fabra and attendees at the SAe 2022 conference, the East Coast Behavioral and Experimental 2023 workshop, Bounded Rationality in Choice conference 2023, CESC 2023, MOVE-ISER 2023 workshop, BSE Summer Forum and Risk Uncertainty and Decisions conference 2023 for helpful feedback. Payró gratefully acknowledges financial support from the Ministerio de Economía y Competitividad and Feder (PGC2018-094348-B-I00 and PID2020-116771GB-I00) and financial support from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S).The usual disclaimer applies.

Chen, Moskowitz and Shue (2016) document that loan officers are more likely to reject (accept) a loan application after they have accepted (rejected) the previous application, and estimate that in some treatments up to 9% of decisions are erroneous due to such a sequencing effect. They find a similar bias in decisions by US judges in refugee asylum cases. Jin and Peng (2022) show that various puzzles in finance (such as the disposition effect, where traders hold on to recently losing stock and sell recently winning stock) can be explained by the Gambler's Fallacy.

Excessive Alternation: Subjects believe that random processes feature a lot of switching and few streaks on the path of any sequence of outcomes. For instance, $P(HHTHTH) > P(HHHHTT)$. Rapoport and Budescu (1997), and Bar Hillel and Wagenaar (1991) propose that subjects believe in an alternation rate of approximately 60%. In 2014, Spotify changed its perfectly random shuffling algorithm to a non-random algorithm, due to complaints by users that the original shuffle was not random because one artist's songs would sometimes play in succession.[1]

The celebrated Heuristics and Biases paradigm due to Kahneman and Tversky proposes a theory of beliefs based on heuristics. In order to explain the above findings, they hypothesize that people believe that "even small samples are highly representative of the populations from which they are drawn" (Tversky and Kahneman (1974, pg 1125-1126), a property dubbed *the Law of Small Numbers*. In the context of a fair coin, this asserts the belief that small samples will have approximately an equal proportion of heads and tails. The term "small samples" extends further to segments within any sample: "each segment of the [sequence] is highly representative of the "fairness" of the coin" Tversky and Kahneman (1971, pg 106). [2] The Law of Small Numbers generates a *disbelief in streaks*, which explains both Excessive Alternation and the Gambler's Fallacy.

The Law of Small Numbers is an informal theory.[3] The formalization of the Law of Small Numbers in economics is due to the seminal work by Rabin (2002). Given a coin of perceived bias $\theta^* \in [0,1]$, the belief assigned to a sequence $x^n = (x_1, ..., x_n)$ of outcomes of n tosses (for odd n) in the model is:

$$P(x^n) = P(x_1x_2) \times P(x_3x_4) \times .... \times P(x_{n-1}x_n),$$

where each $P(x_{i-1}x_i)$ is the distribution generated by an urn containing N balls with an integer $\theta^* N$ number of balls labelled "heads", from which draws are made *without* replacement.[4] Sampling without replacement from urns is the tool used to generate the Gambler's Fallacy. The "i.i.d. by pairs" feature of the model makes it tractable for analysis and enables an exploration of the economic implications of the Gambler's Fallacy.

This paper is interested in foundational questions pertaining to the Law of Small Numbers. Specifically:

Q1. What is the general structure required for a model to capture LSN?

Taking the axiomatic approach, we first identify a property of beliefs that captures the essence of the Law of Small Numbers, and then we prove a representation theorem that delivers a class of models. We propose two nested defining properties. Our first proposal is that the Law of Small Numbers is defined by a *Mean Reversion* axiom: in a canonical coin-tossing context, the agent believes that the sample mean will tend to stay close to the bias of the coin *along* the entire sequence. We prove a representation theorem that states that beliefs satisfy Mean Reversion, a weak

[1] https://engineering.atspotify.com/2014/02/how-to-shuffle-songs/

[2] The term "Law of Small Numbers" is coined in Tversky and Kahneman (1971) but, in their 1974 paper, the authors appear to prefer to use the term "Local Representativeness" instead. Local Representativeness describes the belief that "the essential characteristics of the process will be represented, not only globally in the entire sequence, but also locally in each of its parts" (Tversky and Kahneman (1974, pg 1125)). In this paper we treat the two terms as interchangeable.

[3] Indeed, the lack of formal specification of heuristics has been criticized by Gigerenzer (1996) on the grounds that they offer too much flexibility and risk being unfalsifiable.

[4] Therefore the belief that the outcome of a single flip of the coin is heads is $P(H) = \theta^*$ while the belief that two flips generate a heads followed by a tails $P(HT) = \theta^* \frac{(1-\theta^*)N}{N-1}$. Since the urn is "renewed" after every two periods, the belief in HTH is $P(HTH) = P(HT)P(H) = \theta^* \frac{(1-\theta^*)N}{N-1} \times \theta^*$.

Independence property and a standard Marginal Consistency property if and only if it is as if the agent believes the bias $\theta_{i,\overline{x}^{i-1}}$ of the coin is path-dependent and self-correcting:

$$P(x^n) = \prod_{i=1}^{n} (\theta_{i,\overline{x}^{i-1}})^{x_i} (1 - \theta_{i,\overline{x}^{i-1}})^{1-x_i},$$

where $\overline{x}^{i-1}$ is the sample mean for the sequence $x^{i-1} = (x_1, ..., x_{i-1})$.

Our second, more general, proposal is that LSN is defined by *Local Mean Reversion*, an analog of Mean Reversion where the sample mean is computed "locally" at each point of the sequence. The corresponding representation requires the bias of the coin to be *locally* self-correcting: the bias in toss i depends not on the sample mean of $(x_1, ..., x_{i-1})$ at toss $i-1$ but rather that of the segment $(x_{i-k}, ..., x_{i-1})$ defined by the "last k tosses". We argue that Local Mean Reversion is closer in spirit to Tversky and Kahneman (1971, 1974), and show that it nests Rabin (2002).

We find that the "segments" that determine what "local" means are not generally identified in the model. Thus, the generality of LSN as a hypothesis comes with an identification problem. Mean Reversion is a specification of the model that yields complete uniqueness.

Q2. How are we to formalize the statement that LSN is a theory of "streak aversion"?

To our knowledge, a "streak" has never been formally defined. It is easy to agree that 10 consecutive heads is constitutes a streak for a fair coin, but what is the smallest number of heads that constitute a streak for a coin with general bias $\theta^* \in (0, 1)$? We define a (generalized) streak of heads not in terms of a number of consecutive heads, but rather in terms of the sample mean induced locally by the sequence. Thus, Streak Aversion is the property that if the local sample mean exceeds the bias (that is, if there are "too many heads in the last few tosses"), a tails is considered more likely on the next toss. We show that Local Mean Reversion implies Streak Aversion. The result invites further experimental exploration of the Gambler's Fallacy: is a higher belief in tails on the next toss driven by the contiguity of heads preceding it, or on the local concentration of heads? When being elicited the probability of tails on the next toss, would a subject pay to know the outcomes that occurred prior to a contiguous streak of heads?

Q3. Does a belief in LSN imply a Law of Large Numbers?

We show that in a large class of Mean Reversion models, the Law of Large Numbers generically fails. Specifically, the sampling distribution does not collapse in the limit – this property is dubbed *Nonbelief in the Law of Large Numbers* by Benjamin, Rabin and Raymond (2016), and evidence is provided in Benjamin, Moore and Rabin (2018). We therefore show that LSN may serve as a unifying principle for more than just the Gambler's Fallacy and Excessive Alternation.

Future research might also study if our models can accommodate the stronger finding of Sample Size Neglect, where sampling distributions are believed to be insensitive to sample size more generally (Kahneman and Tversky (1972)). The literature has treated LSN and Sample Size Neglect as nonintersecting phenomena (Tversky and Kahneman (1974), Rabin (2002), Benjamin, Rabin and Raymond (2016)), but our analysis suggests that they may be related.

Q4. What do LSN agents learn?

We study Bayesian inference under Mean Reversion. We establish that, in the limit, the Mean Reversion agent almost surely puts strictly positive probability on the true parameter θ^*. Intuitively, this is because the Mean Reversion agent believes that the sample mean tends to the (unknown) true parameter, while the Law of Large Numbers ensures that it does. This is in contrast with Local Mean Reversion (and therefore Rabin (2002)) where the agent may become fully confident in the wrong parameter in the limit. This establishes that learning depends on how one formalizes LSN. Moreover, the formulation in terms of Mean Reversion has some appealing implications for learning.

Q5. Does LSN offer an evolutionary advantage?

We consider an evolutionary setting, where a population containing Mean Reversion agents and IID agents (that is, agents who understand i.i.d. randomness) must decide whether to hunt in a "safe small stakes" hunting ground or a "risky high stakes" hunting ground. Their decision is based

on a public signal about the risky hunting ground, which they use to determine their beliefs about a parameter that captures the desirability of the risky ground. We show that, almost surely, the Mean Reversion agents will eventually become more confident about the true parameter than the IID agents, and consequently, relative to the IID population, a larger proportion of the Mean Reversion population will hunt in the "correct" hunting ground. This will guarantee that the Mean Reversion population will survive asymptotically.[5] We also establish a lower bound on the probability that the IID population does not survive asymptotically. This lower bound gets arbitrarily close to 1, as the desirability of the risky ground gets close to 1.

The paper is organized as follows. Section 2 axiomatizes the "correctly specified" model. Sections 3 presents the Mean Reversion model, while Section 4 generalizes it to allow for Local Mean Reversion. Section 5 derives implications of the models, and discusses how they connect with the evidence on beliefs about randomness. Section 6 studies Bayesian inference while Section 7 contains an application to evolution. Section 8 relates this paper to the literature. All proofs are relegated to appendices.

2 Rational Benchmark

The evidence on beliefs about randomness is both static (ex-ante, is it more likely to get HHHH or HHHT?) and dynamic (given HHH, is it more likely that the next toss is H or T?). The static evidence is, by definition, revealing properties of the agent's ex-ante beliefs about a sequence of coin tosses. In particular the evidence necessitates a model of misspecified beliefs about the data generating process. An important observation is that the dynamic evidence, which reveals properties of posterior beliefs, is entirely consistent with Bayesian updating of an incorrect prior. Parsimony demands that we visualize the evidence in terms of incorrect ex-ante beliefs alone rather than incorrect updating as well. Consequently, the theory we present is static, with the understanding that dynamic applications will assume Bayesian updating.

2.1 Primitives

Consider a canonical coin-tossing environment: the possible realizations of a coin toss in any period i are $\Omega_i = \Omega = \{0, 1\}$, and the space of all realizations of any $n \leq \infty$ tosses is $\Omega^n = \prod_{i=1}^{n} \Omega_i$. Throughout, we use $x = (x_1, x_2, ...) \in \Omega^\infty$ to denote an infinite sequence and $x^n = (x_1, ..., x_n) \in \Omega^n$ to denote a finite sequence of length n. The concatenation of two sequences $x^n \in \Omega^n$ and $y^m \in \Omega^m$ is denoted $x^n y^m \in \Omega^{n+m}$. Our results do not hinge on the binariness of Ω, which we maintain for simplicity of exposition, and can be readily extended at least to any finite set Ω.

Our primitive consists of a family of beliefs,

$$\{P^n\}_{n=1}^{\infty},$$

where each P^n is a probability measure (henceforth, *belief*) on the measurable space (Ω^n, Σ^n) defined by the sample space Ω^n and the σ-algebra $\Sigma^n = 2^{\Omega^n}$ of all subsets $A^n \subset \Omega^n$. As our axioms are ordinal in nature, they can be derived from the betting behavior of Subjective Expected Utility agents. Since the translation to betting behavior is obvious, we take beliefs $\{P^n\}$ directly as our primitive and interpret them as behavioral objects.

[5] It is natural to consider whether survival can be ensured in a market setting as well. Sandroni (2000) shows that agents who eventually make accurate forecasts will push out agents who do not. This applies to all agents with misspecified priors, which include Mean Reversion agents. However, his result hinges on the market for assets for being complete, as well as other technical assumptions that ensure that two Bayesian agents will eventually agree in the limit. Acemoglu et al (2016) point out that such Bayesian asymptotic agreement results are fragile.

We presume throughout that the coin tosses are objectively i.i.d. with bias $\theta^* \in [0,1]$, so that the objective probability for any sequence x^n is given by:

$$Q(x^n) = \prod_{i=1}^{n} (\theta^*)^{x_i}(1-\theta^*)^{1-x_i}.$$

In this paper we will model the agent's misspecified understanding of what an i.i.d. data-generating process looks like.

Throughout the paper, our examples will presume a fair coin unless stated otherwise.

2.2 Correctly Specified Beliefs

As a benchmark we characterize the model with correct understanding of the data generating process. There are several ways of doing so but we present one that provides perspective on subsequent results.

In economics it is typical to posit the existence of a belief P^∞ over the infinite horizon sample space $(\Omega^\infty, \Sigma^\infty)$, and to consider its marginals.[6] If we begin with a family of beliefs $\{P^n\}_{n=1}^\infty$, Kolmogorov's extension theorem tells us that a necessary and sufficient condition on $\{P^n\}_{n=1}^\infty$ for the existence of P^∞ is:

Axiom 1 *(Marginal Consistency) For any n and any event $A^n \subset \Omega^n$,*

$$P^n(A^n) = P^{n+1}(A^n).$$

Marginal Consistency embodies horizon-independence of beliefs, in the sense that the agent does not think differently about a given event A^n if the horizon is extended by a period. Such horizon-independence is satisfied by standard models. An interesting question is whether the evidence on beliefs about randomness is at odds with Marginal Consistency. We will see that it is not.

The next axiom expresses that the agent knows that the objective bias of the coin is θ^* and that it is constant across tosses. It states that the agent's marginal belief on obtaining a heads in toss n is precisely θ^*.

Axiom 2 *(Knowledge of Bias)For any n,*

$$P^n(\Omega^{n-1}1_n) = \theta^*.$$

The next property involves a recognition that the tosses are independent:

Axiom 3 *(Independence)For any n, any $x^n, y^n \in \Omega^n$ and any $x_{n+1} \in \Omega$ s.t. $P^n(y^n) > 0$ and $P^{n+1}(y^n x_{n+1}) > 0$,*

$$\frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} = \frac{P^{n+1}(y^n x_{n+1})}{P^n(y^n)}.$$

It will be useful to interpret Independence in the following way. Define the conditional probability of x_{n+1} given $x_1, ..., x_n$ by

$$P^{n+1}(x_{n+1}|x_1, .., x_n) := \frac{P^{n+1}(x_1, ..., x_n, x_{n+1})}{P^n(x_1, ..., x_n)}.$$

This is not the usual definition of conditional probability (given by $\frac{P^{n+1}(x_1,....,x_n,x_{n+1})}{P^{n+1}(x_1,....,x_n)}$), but it is equivalent to the usual definition when Marginal Consistency is satisfied. Nevertheless, it indicates

[6] As is standard, we identify any n-event $A^n \in \Sigma^n$ with the event $A^n\Omega^\infty = \{(x^n z) \in \Omega^\infty : x^n \in A^n \text{ and } z \in \Omega^\infty\}$ in the infinite horizon space Ω^∞ known as the *n-cylinder*. Let $\Sigma^\infty = \sigma(\cup_{n=1}^\infty \Sigma^n)$ denote the σ-algebra generated by all the n-cylinders, $n = 1, 2, ...$ Then P^∞ is a probability measure of a well-defined space $(\Omega^\infty, \Sigma^\infty)$. The marginal belief on (Ω^n, Σ^n) is defined by $P^n(A^n) = P^\infty(A^n\Omega^\infty)$ for each n-event $A^n \in \Sigma^n$.

the agent's belief about the probability of outcome x_{n+1} given the history $x_1, .., x_n$. Independence states that the conditional probability of x_{n+1} is independent of the history: for any n and any $x_{n+1} \in \Omega$ and $x^n, y^n \in \Omega^n$,

$$P^{n+1}(x_{n+1}|x_1, .., x_n) = P^{n+1}(x_{n+1}|y_1, .., y_n).$$

A similar property is that the relative probability of sequence x^n and y^n does not change if both are shifted into the future by one step and a common first period outcome x_1 is appended to them. This "stationarity" type property is interpreted in terms of time-invariance of the bias of the coin. This is clearly also contradicted by the Gambler's Fallacy.

Axiom 4 *(Time-Invariant Bias)For any n, and $x^n, y^n \in \Omega^n$ and any $x_1 \in \Omega$ s.t. $P^n(y^n) > 0$ and $P^{n+1}(x_1 y^n) > 0$,*

$$\frac{P^{n+1}(x_1 x^n)}{P^{n+1}(x_1 y^n)} = \frac{P^n(x^n)}{P^n(y^n)}.$$

We show that these properties characterize an agent who correctly understands the coin. See Appendix B for more general versions of the results in this section that drop Marginal Consistency and Knowledge of Bias.

Theorem 1 *A family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ satisfies Knowledge of Bias, Marginal Consistency, Independence and Time-Invariant Bias if and only if it is θ^*-i.i.d: for all n and $x^n \in \Omega^n$,*

$$P^n(x^n) = \prod_{i=1}^{n} (\theta^*)^{x_i} (1 - \theta^*)^{1 - x_i}.$$

The evidence is incompatible with Independence and Time-Invariant Bias. For instance, the Gambler's Fallacy requires that the presence of a streak of heads prior to toss $n + 1$ is believed to impact the probability of heads on $n + 1$, directly contradicting Independence. Similarly, in contradiction to the Gambler's Fallacy, Time-Invariant Bias requires that $\frac{P(HH)}{P(HT)} = \frac{P(H)}{P(T)} = 1$ for a fair coin. In the sequel, we will drop Time-Invariant Bias and relax Independence. The proof of Theorem 1 shows that dropping Time-Invariant Bias leads to a similar representation where the bias θ_i can vary exogenously with each toss i. By relaxing Independence our results will generalize the model even further, so that the bias varies endogenously with the outcomes of previous tosses.

3 Mean Reversion

As noted in the Introduction, the Law of Small Numbers is the term given to the belief that "even small samples are highly representative of the populations from which they are drawn" (Tversky and Kahneman (1974, pg 1125-1126). We seek to articulate this as an axiom on beliefs. We interpret the Law of Small Numbers as a belief that the sample mean in any finite sequence will be close to the bias, and moreover, that this belief extends also to subsegments in the sequence (Tversky and Kahneman (1971, 1974)). Since the Law of Small Numbers is an informal hypothesis, it does not specify which subsegments it extends to. Indeed, the property has almost no content if it is meant to apply to all subsets of the sequence. In this section, we formulate the parsimonious case where it applies to segments that start from the first toss. In Section 4, we relax this to obtain a more general model.

3.1 Axioms

Denote the *sample mean* number of heads at any point n of a sequence $x = (x_1, x_2, ...)$ by

$$\overline{x}^n := \frac{\sum_{i \leq n} x_i}{n},$$

and denote the distance between the sample mean and the true bias by

$$d(x^n) = |\overline{x}^n - \theta^*|.$$

Adopt the convention that $d(x^{i-1}) = \overline{x}^{i-1} = 0$ when $i = 1$.

The following axiom captures the Law of Small Numbers through the idea that beliefs are driven by a consideration of the distance between sample mean and true bias, and that "dominance on path" implies higher beliefs.

Axiom 5 *(Mean Reversion) For any n and $x, y \in \Omega^n$ s.t. $d(x^{n-1}) = d(y^{n-1})$,*

$$d(x^i) \leq d(y^i) \text{ for all } i \leq n \implies P^n(x) \geq P^n(y).$$

The axiom states that if the sample means of the sequences x and y are equally distant from θ^* at throw $n-1$, and if x dominates y on path for all n throws, then x^n is deemed more likely than y^n. It is immediate that Mean Reversion implies a degree of disbelief in streaks: for instance it implies the Gambler's Fallacy, $P(HHHT) > P(HHHH)$, due to the obvious dominance on path. It also gives rise to Excessive Alternation – taking $P(HTHTTH) > P(HHHTTT)$ as an illustrative example, Table 1 computes the sample mean of both sequences at each toss and verifies that there is dominance on path. However, Mean Reversion does not embody unconditional disbelief in streaks. For instance, it implies that $HHTTTT$ is more likely than $TTTHHT$, because the former better maintains the sample mean on path, despite the longer streaks.

	$HTHTTH$	$HHHTTT$
$n=1$	1	1
$n=2$	$\frac{1}{2}$	1
$n=3$	$\frac{2}{3}$	1
$n=4$	$\frac{1}{2}$	$\frac{3}{4}$
$n=5$	$\frac{2}{5}$	$\frac{3}{5}$
$n=6$	$\frac{1}{2}$	$\frac{1}{2}$

Table 1. The entries are $\overline{x}^n = \dfrac{\sum_{i=1}^n x_i}{n}$ for $x = HTHTTH, HHHTTT$ and $n = 1, ..., 6$.

By definition, Mean Reversion embodies a degree of awareness that the true bias is θ^*, since beliefs respond to deviations from θ^*. While such an expression of awareness is suggested by the Gambler's Fallacy and Excessive Alternation, future research might investigate whether the additional expression of knowledge captured in the Knowledge of Bias axiom holds empirically.

The requirement "$d(x^{n-1}) = d(y^{n-1})$" is often satisfied in the evidence. For instance it is satisfied in the Gambler's Fallacy where $HHHT$ is deemed more likely than $HHHH$, and it is also satisfied in the Excessive Alternation example where $HTHTTH$ is deemed more likely than $HHHTTT$. Nevertheless, we now write a weaker and a stronger version of Mean Reversion that will appear in our applications. The Weak Mean Reversion axiom weakens Mean Reversion by restricting attention to x, y that have the same sample mean in throw $n-1$, rather than the same distance to the mean:

Axiom 6 *(Weak Mean Reversion) For any n and $x, y \in \Omega^n$ s.t. $\overline{x}^{n-1} = \overline{y}^{n-1}$,*

$$d(x^i) \leq d(y^i) \text{ for all } i \leq n \implies P^n(x) \geq P^n(y).$$

In contrast, the Strong Mean Reversion axiom requires that dominance on path is respected for all pairs of sequences.

Axiom 7 *(Strong Mean Reversion) For any n and $x, y \in \Omega^n$,*

$$d(x^i) \leq d(y^i) \text{ for all } i \leq n \implies P^n(x) \geq P^n(y).$$

The Mean Reversion axioms only place structure on beliefs that can be ranked by dominance on path. Such beliefs admit a very general representation – see Appendix B. But economic applications will require additional structure to generate sharper predictions. For such additional structure it is worth retaining some version of standard axioms, so that our model serves as a generalization of the standard model presented in Section 2.2. In our main result we will maintain Marginal Consistency and drop Time-Invariant Bias. Independence is too strong since it requires $\frac{P(HH)}{P(TH)} = \frac{P(H)}{P(T)} = 1$, in contradiction to Excessive Alternation. However, consider the following weakening of Independence:

Axiom 8 *(MR Independence)For any n, any $x^n, y^n \in \Omega^n$ and any $x_{n+1} \in \Omega$ s.t. $P^n(y^n) > 0$ and $P^{n+1}(y^n x_{n+1}) > 0$,*

$$\overline{x}^n = \overline{y}^n \implies \frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} = \frac{P^{n+1}(y^n x_{n+1})}{P^n(y^n)}.$$

MR-Independence states that the conditional probability of x_{n+1} can depend on the history x^n only through its sample mean $\overline{x}^n$. This is consistent with Mean Reversion, but not implied by it, since Mean Reversion does not make any statement about conditional probabilities. MR-Independence rules out some degree of disbelief in streaks that might cause the agent to feel that the conditional probability of tails is lower when the history is $HHTTT$ rather than $TTTHH$, that is, $\frac{P^6(HHTTT,T)}{P^5(HHTTT)} < \frac{P^6(TTTHH,T)}{P^5(TTTHH)}$.

3.2 Representation Result

The well-specified model (Theorem 1) satisfies the standard Marginal Consistency condition, allowing beliefs to be fully described by a belief P on the infinite sample space Ω^∞. Can findings such as the Gambler's Fallacy be modeled with a belief P on Ω^∞? Our main theorem assures us that Mean Reversion is not fundamentally incompatible with Marginal Consistency.

Recall that for any sequence $x \in \Omega^\infty$, the sequence truncated at i is denoted $x^i \in \Omega^i$. Adding an outcome of heads (respectively, tails) in the $i+1^{st}$ toss yields $x^i 1 \in \Omega^{i+1}$ (respectively $x^i 0 \in \Omega^{i+1}$).

Theorem 2 *A family of full support beliefs $\{P^n\}_{n=1}^\infty$ satisfies Weak Mean Reversion, MR Independence and Marginal Consistency iff for each $i \geq 1$ there exists a continuous function $g^i : [0,1]^2 \to (0,1]$ that is weakly decreasing in its first argument, and for all n and $x^n \in \Omega^n$,*

$$P^n(x^n) = \prod_{i=1}^{n} (\theta_{i,\overline{x}^{i-1}})^{x_i} (1 - \theta_{i,\overline{x}^{i-1}})^{1-x_i},$$

where

$$\theta_{i,\overline{x}^{i-1}} = g^i(d(x^{i-1}1), \overline{x}^{i-1}) \text{ and } 1 - \theta_{i,\overline{x}^{i-1}} = g^i(d(x^{i-1}0), \overline{x}^{i-1}).$$

In the representation, $\{P^n\}$ satisfies Mean Reversion (resp. Strong Mean Reversion) iff g^i has $d(x^{i-1})$ as its second argument (resp. g^i is constant in its second argument).

The representation is reminiscent of a "belief that a random process is self-correcting" (Kahneman and Tversky (1974)). The result tells us that the axioms are characterized by a path-dependent bias that is self-correcting in that it varies so as to keep the sample mean near the bias θ^*. Specifically, $\theta_{i,\overline{x}^{i-1}} \in (0,1)$ is the propensity for heads in the i^{th} flip, conditional on a sample mean $\overline{x}^{i-1}$ up to that point. The dependence of the bias $\theta_{i,\overline{x}^{i-1}}$ on i and $\overline{x}^{i-1}$ is described by the g^i function. This function is weakly decreasing in the distance to θ^*, hence the bias is only weakly self-correcting. When g^i is constant in its arguments then $\theta_{i,\overline{x}^{i-1}} = \frac{1}{2}$ and the agent perceives a fair coin at $i, \overline{x}^{i-1}$.

Observe that g^i depends on the toss i, so that the model permits the sensitivity of beliefs on the sample mean to change with the toss – for instance, sensitivity may increase as we go deeper into the sequence. This reflects the generality of the Mean Reversion axioms, which are silent on how

the sample means at different points i in the sequence impact the probability of heads on the n^{th} toss.

The representation is "essentially" unique. The bias $\theta_{n,\overline{x}^{n-1}}$ is identified uniquely by conditional beliefs:

$$\theta_{n,\overline{x}^{n-1}} = \frac{P^n(x^{n-1}1)}{P^{n-1}(x^{n-1})} = P^n(x^{n-1}1|x^{n-1}).$$

However, the model is defined using $g^n(d(x^n), \overline{x}^{n-1})$ that is a function on $[0,1]^2$, whereas in toss n there are only finitely many values for $(d(x^n), \overline{x}^{n-1})$ generated by all sequences $x^n \in \Omega^n$. Thus g^n is uniquely defined by beliefs only on a finite set of points, and can be defined arbitrarily on other points.

The next theorem collects several results of interest. It shows that Weak Mean Reversion and MR Independence imply that the representation can be written as a product of the g^i functions. Some of our applications will use this formulation of the model for its generality.[7]

Theorem 3 *Beliefs $\{P^n\}$ satisfy Weak Mean Reversion and MR Independence if and only if for each $i \geq 1$ there exists a continuous function $g^i : [0,1]^2 \to (0,1]$ that is weakly decreasing in its first argument such that for any n and $x^n \in \Omega^n$,*

$$P^n(x^n) = \prod_{i=1}^{n} g^i(d(x^i), \overline{x}^{i-1}).$$

In the representation, $\{P^n\}$ satisfies Mean Reversion (resp. Strong Mean Reversion) iff g^i has $d(x^{i-1})$ as its second argument (resp. g^i is constant in its second argument).

Furthermore:

(i) $\{P^n\}$ satisfies Marginal Consistency iff $g^i(d(x^{i-1}1), \overline{x}^{i-1}) + g^i(d(x^{i-1}0), \overline{x}^{i-1}) = 1$ for all x, i.

(ii) $\{P^n\}_{n=1}^{\infty}$ satisfies Knowledge of Bias if and only if $\{g^i : [0,1]^2 \to (0,1]\}_{i=1}^{\infty}$ inductively satisfies

$$g^1(1, \cdot) = \theta^*,$$

and for each $n+1$,

$$\sum_{x^n} g^{n+1}(d(x^n 1), \overline{x}^n) P^n(x^n) = \theta^*.$$

(iii) If $\{P^n\}$ satisfies Strong Mean Reversion and Marginal Consistency then it can be represented by taking $g^n = \frac{1}{2}$ for all $n > 3$.

Part (i) is a restatement of Theorem 2. It shows that Marginal Consistency is characterized by the requirement that the bias for heads and the bias for tails necessarily sum to 1 for any $i, \overline{x}_{i-1}$, leading to the readily interpretable self-correcting bias representation in Theorem 2.

Mean Reversion already embeds awareness of the true bias θ^*, through the reliance on the distance d between the sample mean and θ^*, but the marginal probability she assigns to a heads in toss n is allowed to differ from θ^*. A stronger statement of awareness is to assume Knowledge of Bias. Part (ii) provides a characterization. At each step n, the functions $g^1, .., g^n$ define P^n through the representation. This in turn defines the restriction on the average of $g^{n+1}(d(x^n 1), \overline{x}^n)$ wrt P^n stated in the result. Observe that this a restriction on g^{n+1} only for sequences $x^n 1$ that end in a heads.

Finally, part (iii) demonstrates that Strong Mean Reversion, while yielding an attractively simple model, is quite strong. When imposed together with Marginal Consistency, the model reduces to one where the agent believes the coin to be fair after the 4th toss, regardless of θ^*.

[7] Speaking of generality, one may wonder what an extension of our model may look like when the outcome of the random process lies in $\mathbb{R}$. Suppose the objective distribution over sequences of outcomes in $\mathbb{R}^n$ is a product of normals $N(\mu, \sigma^2)$ with some mean μ and variance σ^2. Let the density of any normal be denoted by $\varphi(\mu, \sigma^2)$. In this case our model could be written in terms of a density f^n that maps sequences of outcomes in $\mathbb{R}^n$ to $\mathbb{R}_+$. It would be such that $f^n(x^n) = \prod_{i=1}^{n} \varphi(\mu_{\overline{x}^{i-1}}, \sigma^2)(x_i)$ where the mean $\mu_{\overline{x}^{i-1}}$ is self-correcting as in our model. The variance could also depend on $\overline{x}^{i-1}$, reflecting the strength of the self-correction as a function of the deviation from the true mean.

3.3 Illustration

The *Friedman Urn* (Friedman (1949)) is a special case of our model with $\theta^* = \frac{1}{2}$. Imagine that an urn initially contains one ball labelled "heads" and one ball labelled "tails". Inductively, at any toss n, if a heads (resp. tails) is drawn, then one ball labelled "tails" (resp. "heads") is added to the urn. This generates conditional probabilities for any $n \geq 1$ given by

$$P_F(H_n|x^{n-1}) = \frac{(n-1)(1-\overline{x}^{n-1})+1}{n+1} \text{ and } P_F(T_n|x^{n-1}) = \frac{(n-1)\overline{x}^{n-1}+1}{n+1}$$

with the convention that $\overline{x}^0 = 1$, so that $P_F(H_1) = P_F(T_1) = \frac{1}{2}$. The Friedman Urn is a probability distribution P_F on Ω^∞ that satisfies:

$$P_F(x^n) = \prod_{i=1}^{n} P_F(x_i|x^{i-1}).$$

The Friedman Urn corresponds to our model where $g^n(d(x^n), \overline{x}^{n-1}) = P_F(x_n|x^{n-1})$.[8] The Friedman Urn satisfies Weak Mean Reversion and MR Independence. It also satisfies Marginal Consistency (since $g^{n+1}(\overline{x^n 1}, \overline{x}^n) + g^{n+1}(\overline{x^n 0}, \overline{x}^n) = 1$) and Knowledge of Bias (since the symmetry implies that for every sequence with mean $\overline{x}^n$ there exists a sequence – with tails substituted for heads – with the same probability but mean $1 - \overline{x}^n$). Thus it satisfies many axioms of interest for the case $\theta^* = \frac{1}{2}$.

Consider a more flexible specification of the model that might be useful for empirical work: take a one-parameter specification of the Strong Mean Reversion model

$$g^i(d) = \frac{1}{Z^i}\left(\frac{1}{1+d}\right)^{\lambda_i}$$

where, of course, the normalizing constant Z^i satisfies $Z^i = \sum_{x^i \in \Omega^i}\left(\frac{1}{1+d(x^i)}\right)^{\lambda_i}$. Then according to the model

$$\frac{P(x^{i-1}1)}{P(x^{i-1}0)} = \left(\frac{1+d(x^{i-1}0)}{1+d(x^{i-1}1)}\right)^{\lambda_i}.$$

Observe that λ_i is completely identified by this equation, and indeed can be estimated empirically using data on beliefs. A comparative static wrt λ_i is useful to interpret this parameter. Taking a derivative yields

$$\frac{d\frac{P(x^{i-1}1)}{P(x^{i-1}0)}}{d\lambda_i} = \left(\frac{1+d(x^{i-1}0)}{1+d(x^{i-1}1)}\right)^{\lambda_i} \ln\left(\frac{1+d(x^{i-1}0)}{1+d(x^{i-1}1)}\right).$$

Thus, if heads on toss i brings the mean closer to the bias than does a tails ($d(x^{i-1}1) < d(x^{i-1}0)$) then the derivative is strictly positive and so increasing λ increases the relative likelihood $\frac{P(x^{i-1}1)}{P(x^{i-1}0)}$. That is, a higher λ captures a stronger belief in Mean Reversion.

An example of an empirical question is whether the strength of mean reversion (parametrized by λ_i) changes with the toss i. Do subjects believe that mean reversion becomes stronger further down the sequence, or do they become more tolerant towards deviations from θ^* ?

4 Local Mean Reversion

While Mean Reversion is global in the sense of computing the sample mean at toss i using the entire sequence of outcomes $x_1....x_i$, we now formulate a notion where the agent may be concerned with

[8] To see that g^n is well-defined, first observe that $P_F(\cdot|x^{n-1})$ depends only on n and the sample mean of the history. Next observe that if $d(x^{n-1}1) = d(x^{n-1}0)$ then it must in fact be that $\overline{x^n} = \frac{1}{2}$. In that case, $P_F(H_n|x^{n-1}) = P_F(T_n|x^{n-1})$, since there must be an equal number of heads and tails in the urn.

the sample mean computed using only the outcomes of the "last few" tosses. This is compatible with Tversky and Kahneman (1974)'s Local Representativeness and Rabin (2002)'s seminal model of the Gambler's Fallacy (see Section 8). It is also plausible when n is large, since cognitive constraints may lead the agent to focus on parts of the sequence – although alternatively they could form a coarse perception of the mean number of heads in the entire sequence.

4.1 Axioms

Recall that the conditional probability of x_n given $x_1, ..., x_{n-1}$ is defined by $P^n(x_n|x_1, .., x_{n-1}) := \frac{P^n(x_1,....,x_{n-1},x_n)}{P^{n-1}(x_1,....,x_{n-1})}$. Mean Reversion and MR-Independence permit the entire history $x_1, .., x_{n-1}$ to matter (through the sample mean $\overline{x}^{n-1}$). Consider now that the agent may only be concerned with a subset of the history, which we refer to as the "segment at n". An immediate question that arises is: how do we identify the segment the agent is looking at? We posit that at the very least:

Definition 1 *(Segments) For any n, a segment at n is a set of contiguous indices $W_n = \{k_n, .., n\} \subseteq \{1, .., n\}$ containing n and satisfying*

$$P^n(x_n|x_1...x_{k_n-1}x_{k_n}...x_{n-1}) = P^n(x_n|y_1...y_{k_n-1}x_{k_n}...x_{n-1}) \text{ for all } x, y \in \Omega^\infty.$$

Thus, tosses $\{k_n, ..., n\}$ are referred to as a segment at n if the outcomes *outside* those tosses never impact the conditional probability of x_n. The definition assumes that a segment necessarily consists of contiguous tosses and depends only on n (rather than the outcome x_n). This is to keep notation clean and the model simple, and can be easily relaxed. We expect that, empirically, for very long sequences attentional constraints may prompt the agent to focus on recent history, though this does not preclude the possibility that early history may still attract attention if it has very skewed outcomes. An empirical evaluation of segments is an interesting direction for future research.

Segments are always nonempty (since they contain the last toss n). Segments also always exist. For instance, $\{1, .., n\}$ is trivially always a segment. In the (strict version of the) Mean Reversion model it is the unique segment at n. But in general, there are potentially many segments at n that satisfy the definition. For now, we suppose that there is some family of segments $\{W_n\}_{n\geq 1}$ selected by the analyst. We assume that the family has some structure that is possessed by Mean Reversion and the literature as well (Section 8). Specifically we assume that a segment does not extend back farther than the previous segment: if $W_n = \{k_n, .., n\}$ and $W_n = \{k_{n+1}, .., n+1\}$ then $k_n \leq k_{n+1}$. An equivalent way to state this is:

Axiom 9 *(Segment Regularity) For all n,*

$$W_{n+1} \backslash \{n+1\} \subseteq W_n.$$

Define the segment mean and distance by the following natural generalizations of sample mean and distance

$$\overline{x}^n(W_n) = \frac{\sum_{i\in W_n} x_i}{|W_n|} \text{ and } d_{W_n}(x^n) = |\overline{x}^n(W_n) - \theta^*|.$$

The Mean Reversion axiom can be generalized as follows:

Axiom 10 *(Local Mean Reversion) For any n and $x, y \in \Omega^n$ s.t. $\overline{x}^{n-1}(W_{n-1}) = \overline{y}^{n-1}(W_{n-1})$,*

$$d_{W_i}(x^i) \leq d_{W_i}(y^i) \text{ for all } i \leq n \implies P^n(x) \geq P^n(y).$$

To illustrate, consider the Excessive Alternation example where $P(HTHTTH) > P(HHHTTT)$ for a fair coin, and suppose the segment length is fixed at 2, so that the relevant segment are $\{1, 2\}, \{2, 3\}, .., \{5, 6\}$. The sample mean in each segment is closer to $\frac{1}{2}$ in sequence $HTHTTH$ than

in sequence $HHHTTT$ — see Table 2. The axiom implies $P(HTHTTH) > P(HHHTTT)$. For a fair coin, if the segment is $\{n\}$

	$HTHTTH$	$HHHTTT$
$\{1,2\}$	$\frac{1}{2}$	1
$\{2,3\}$	$\frac{1}{2}$	1
$\{3,4\}$	$\frac{1}{2}$	1
$\{4,5\}$	0	0
$\{5,6\}$	$\frac{1}{2}$	0

Table 2. The entries are $\overline{x}^n(\{n-1,n\}) = \dfrac{x_{n-1}+x_n}{2}$ for $x = HTHTTH, HHHTTT$ and $n = 2, ..., 6$.

Local Mean Reversion is strictly more general than Mean Reversion. For instance if $\theta^* = \frac{3}{4}$ and $W_1 = \{1\}$, $W_2 = \{1,2\}$ and $W_3 = \{3\}$, then HTH dominates HHT according to Local Mean Reversion, and so $P(HTH) \geq P(HHT)$. Mean Reversion on the other hand requires both sequences to be equally likely.

Finally, we weaken MR Independence in the natural way:

Axiom 11 *(Local MR Independence) For all n,*

$$\overline{x}^n(W_n) = \overline{y}^n(W_n) \implies \frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} = \frac{P^{n+1}(y^n x_{n+1})}{P^n(y^n)}.$$

Local Mean Reversion and Local MR Independence rely on the family of segments $\{W_n\}_{n\geq 1}$ selected by the analyst. The agent will not satisfy Local Mean Reversion for every family of segments, and therefore the axiom itself becomes a means of practically determining the family $\{W_n\}_{n\geq 1}$. There may in principle still be more than one family that is consistent with Local Mean Reversion and Local MR Independence (in the extreme case of uniform beliefs, all segments are consistent with the axiom). This reveals an identification problem in "local" versions of the Mean Reversion model, in the absence of further structure. Further structure can be exploited in an experimental setting. For instance, subjects can be asked their assessed probability of heads following a concealed sequence x^n, and they can then be asked how much they would pay to see different lengths of the concealed history.

Our Mean Reversion model can alternatively be generalized to *Weighted Mean Reversion* where, instead of tracking the sample mean $\overline{x}^n = \frac{\sum_{i=1}^n x_i}{n}$, the agent tracks a weighted sample mean such as

$$\frac{\sum_{i=1}^n \delta^{n-i} x_i}{\sum_{i=1}^n \delta^{n-i}}.$$

Mean Reversion obtains as the special case where $\delta = 1$. Such a model is different in spirit than Tversky and Kahneman (1974)'s Local Representativeness and it excludes Rabin (2002), but is reminiscent of Rabin and Vayanos (2010)'s history dependence of signals in a learning problem. We expect an identification problem to arise in this model as well. We leave an analysis of this notion of Mean Reversion to future research.

Terminology notwithstanding, Local Mean Reversion contains Mean Reversion as a special case. When $W_n \subsetneq \{1, ..., n\}$ for some n, the model can be referred to as one of nontrivial Local Mean Reversion.

4.2 Representation Result

Theorem 4 *A family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ satisfies Marginal Consistency, Local Mean Reversion, LMR Independence and Segment Regularity iff there exists a regular family of segments*

$\{W_n\}_{n=1}^{\infty}$ and for each $n \geq 1$ there exists a continuous function $g^n : [0,1]^2 \to (0,1]$ that is weakly decreasing in its first argument such that for any n and $x^n \in \Omega^n$,

$$P^n(x^n) = \prod_{i=1}^{n} (\theta_{i,\overline{x}^{i-1}(W_{i-1})})^{x_i} (1 - \theta_{i,\overline{x}^{i-1}(W_{i-1})})^{1-x_i},$$

where

$$\theta_{i,\overline{x}^{i-1}(W_{i-1})} = g^i(d_{W_i}(x^{i-1}1), \overline{x}^{i-1}(W_{i-1})) \text{ and } 1 - \theta_{i,\overline{x}^{i-1}(W_{i-1})} = g^i(d_{W_i}(x^{i-1}0), \overline{x}^{i-1}(W_{i-1})).$$

The representation is similar to that for Mean Reversion except that the bias at toss i is determined by $d_{W_i}(x^{i-1}1)$, rather than $d(x^{i-1}1)$. The bias is greater than $\frac{1}{2}$ in toss i if and only if $d_{W_i}(x^{i-1}1) \leq d_{W_i}(x^{i-1}0)$, that is, if and only if a heads leads to mean reversion locally within segment ending in toss i. There is an obvious counterpart of Theorem 3 that we omit.

In general, the segments proposed by the representation are not unique, that is, two different families of segments may be consistent with the same beliefs (albeit with different g functions).[9] Under a stronger Local MR Independence condition we can show that the segments must be unique.

Proposition 1 *Suppose that, given a regular family of segments $\{W_n\}_{n=1}^{\infty}$, the family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ satisfies Local Mean Reversion and for all n,*

$$d_{W_n}(x^n) = d_{W_n}(y^n) \iff \frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} = \frac{P^{n+1}(y^n x_{n+1})}{P^n(y^n)}.$$

Then in any representation $\{(g^n, \hat{W}_n)\}_{n=1}^{\infty}$ of $\{P^n\}_{n=1}^{\infty}$, it must be that $W_n = \hat{W}_n$ for all n.

A parsimonious specification of the Local Mean Reversion model is one where g^n is independent of n, and the segments have fixed length of $k \geq 1$ for all $n \geq k$ and length n for $n < k$. Such a model can be viewed in terms of a Markov Chain. Define a Markov Chain $\{s_t\}$ that map S into itself as follows. The state space is the space of all configurations of a segment, $S = \Omega^k$. The transition probability $p(s'|s)$ for going to state s' from s equals zero except when the states are overlapping in the sense that $s = (x_i, .., x_{i+k})$ and $s' = (x_{i+1}, .., x_{i+k+1})$, in which case the transition probability is defined by

$$p(x_{i+1}, .., x_{i+k+1} | x_i, .., x_{i+k}) = g(d_W(x_i, .., x_{i+k}), \overline{x_{i+1}, .., x_{i+k+1}}).$$

Initial probabilities of the states are defined by $\pi(x_1, .., x_k) = \prod_{i=1}^{k-1} g(d(x^i), \overline{x}^{i-1})$. The probability of a sequence of states $(s, s', s'', ..)$ is given by $\mu(s, s', s'', ..) = \pi(s)p(s'|s)p(s''|s')....$ Then for $n \geq k$ we see that

$$P^n(x^n) = \prod_{i=1}^{k-1} g(d(x^i), \overline{x}^{i-1}) \times \prod_{i=k}^{n} g(d_{W_i}(x^{i-1}1), \overline{x}^{i-1}(W_{i-1})) = \mu((x_1, .., x_k), (x_2, .., x_{k+1}), ...).$$

4.3 Illustration: Rabin (2002)

Rabin (2002)'s model is a special case of our Local Mean Reversion model. For any θ^*, it corresponds to the model where (i) for even n the relevant segment is $\{n-1, n\}$ and for odd n it is $\{n\}$,

[9] Take $\theta^* = \frac{1}{2}$ and a Local Mean Reversion model defined as follows: $W_n = \{n\}$ for any odd n and $W_n = \{n-1, n\}$ for all even n, $g^n(0) = 1$ for all even n and $g^n(\cdot) = \frac{1}{2}$ for all odd n. The corresponding beliefs are described as follows. If $\overline{x}^{n-1} = 0.5$, the conditional probability of a heads is $\frac{1}{2}$ in period n. If $\overline{x}^{n-1} > 0.5$ (resp. $\overline{x}^{n-1} < 0.5$) then the conditional probability of a tails (resp. heads) is 1 in period n. These beliefs satisfy the Strong Mean Reversion and MR-Independence axioms. They can be represented using segments $\hat{W}_n = \{1, ..., n\}$ and the same g^n as before.

(ii) $g^i(d_W(H)) = \theta^*$ for all odd i and (iii) $g^i(d_{W_i}(HT), 1) = \frac{(1-\theta^*)N}{N-1}$, $g^i(d_{W_i}(TH), 0) = \frac{\theta^* N}{N-1}$, $g^i(d_{W_i}(HH), 1) = \frac{\theta^* N-1}{N-1}$ and $g^i(d_{W_i}(TT), 0) = \frac{(1-\theta^*)N-1}{N-1}$ for all even i. When $\theta^* = \frac{1}{2}$ then g^i is constant in the second argument for all i, and so the model satisfies a strong version of Local Mean Reversion.

A property of Rabin (2002) is that HT and TH are equally likely. This is due to the exchangeability of sampling without replacement from an urn. Local Mean Reversion permits violations of this exchangeability property, as HT may be viewed as more likely if $\theta^* > \frac{1}{2}$. Thus Local Mean Reversion defines a strictly larger class of models.

As a model of nontrivial Local Mean Reversion, Rabin (2002) can violate Weak Mean Reversion: when beliefs have full support and $\theta^* = \frac{1}{2}$, then the model requires $P(HHHTTH) > P(TTHHHH)$ while Weak Mean Reversion requires $P(HHHTTH) \leq P(TTHHHH)$. Note however that the model violates the condition in Proposition 1, since for n even it implies $\frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} = \frac{P^{n+1}(y^n x_{n+1})}{P^n(y^n)}$ regardless of x^n, y^n. Indeed, the example in footnote 9 shows how a special case of Rabin (2002) can be replicated by Mean Reversion.

5 Out-of-Sample Predictions

5.1 Streak Aversion: Gambler's Fallacy and Excessive Alternation

It is not possible to formally prove that our model exhibits, say, the Gambler's Fallacy without first defining the Gambler's Fallacy formally. However, further research is required to understand the conditions under which a Gambler's Fallacy arises (e.g. do subjects think in terms of local segments or the entire sequence, do they make an assessment of the arithmetic mean, do they use some generalized mean, etc?), and the formulation of a Gambler's Fallacy axiom is best left pending until then. To the extent that Gambler's Fallacy and Excessive Alternation both intuitively express a disbelief in streaks, we formalize a "Streak Aversion" axiom and note that our general model satisfies it.

The term "streak" is used pervasively in the literature, but a moment's thought reveals that it is hardly obvious what a streak is. Two agents may agree that 6 heads constitutes a streak, but they may differ on whether 1 head constitutes a streak. More to the point, what constitutes a streak if the bias is 0.6? A formalization of the notion of a disbelief in streaks, which we refer to as Streak Aversion, demands us to make precise the notion of a streak.

If a streak of heads means "too many heads have occurred recently", we require two components: a "recent" history and the proportion of heads in it. Given our general model, if we are interested in the outcome of toss n then it is natural to take the recent history to be $W_n \backslash \{n\}$, the relevant segment at n excluding n. It is natural to say that "too many heads have occurred" if the sample mean within the recent history $W_n \backslash \{n\}$ has already exceeded θ^*.

Definition 2 *(Streak) For any sequence $x \in \Omega^\infty$, there exists a* streak of heads preceding toss $n-1$ *if*

$$\frac{\sum_{i \in W_n \backslash \{n\}} x_i}{|W_n|} \geq \theta^*.$$

Similarly, there exists a streak of tails preceding toss $n-1$ *if* $\frac{\sum_{i \in W_n \backslash \{n\}} x_i}{|W_n|} \leq \theta^*$.

Our definition of streak of (say) heads does not require contiguous heads. For instance if the agent's relevant segment is of length 5 then HHTH contains a streak of heads, as does THHH. We define Streak Aversion as a belief in mean reversion in toss n conditional on a streak preceding n.

Axiom 12 *(Streak Aversion) For any n, outcome $\omega \in \{0,1\}$ and sequence $x \in \Omega^n$,*

$$x^n \text{ contains a streak of } \omega \text{ at } n-1 \implies P^n(x^{n-1}(1-\omega)) \geq P^n(x^{n-1}\omega).$$

A trivial observation is that:

Proposition 2 *If $\{P_n\}_{n=1}^{\infty}$ satisfies Local Mean Reversion, then it exhibits Streak Aversion.*

Local Mean Reversion readily implies Streak Aversion. It is strictly stronger since Streak Aversion has bite only when the segment mean in the relevant segment at n has already crossed θ^* by toss $n-1$, whereas Local Mean Reversion has bite even if it has not. The proposition tells us that our model explains the Gambler's Fallacy and Excessive Alternation to the extent that these are driven by an aversion to streaks. The proposition also suggests further research on the Gambler's Fallacy: is a higher conditional belief in tails driven by the contiguity of a streak of heads, or a local concentration of heads? It would be interesting to explore if there exists a segment $\{k_n, .., n\}$ such that the Gambler's Fallacy (leading to a tails, say) arises when the sample mean is too high within it:

$$\frac{\sum_{i=k_n}^{n} x_i}{n-k_n+1} > \theta^* \implies P^{n+1}(x^nT) \geq P^{n+1}(x^nH).$$

We close with comments on evidence related to the Gambler's Fallacy.

The *Retrospective Gambler's Fallacy* refers to the belief that outcome of the flip *preceding* a streak of heads is most likely to be a tails (Oppenheimer and Monin (2009)). This is directly implied by Local Mean Reversion. It is consistent with Mean Reversion, albeit with some nuance. For instance, observe that when $\theta^* = \frac{1}{2}$, Weak Mean Reversion implies

$$P^4(HHHH) < P^4(THHH),$$

because $THHH$ dominates $HHHH$ on path. However, unlike a comparison of sequences of the form x^nH vs x^nT used in the Gambler's Fallacy, there does not always exist dominance in sequences of the form $(H, x_2...x_{n+1})$ vs $(T, x_2...x_{n+1})$. For instance, when $\theta^* = \frac{1}{2}$, the sequence $HTHH$ dominates $TTHH$ by the end of the the second toss, but is dominated by it at the end of the fourth toss. Consequently, Mean Reversion is silent on such comparisons.

The *Long-Distance Gambler's Fallacy* in Benjamin, Moore and Rabin (2018). Following a streak of $r = 1, 2, 5$ heads on *consecutive* flips up to the n^{th} flip, their subjects exhibited that a probability of heads on flip $n+1$ was respectively 44%,41% and 39%. But when the streak came from nonconsecutive draws from *random* locations flips, the probability of heads on another randomly chosen flip was 45%,42% and 41% resp. That is, Gambler's Fallacy appeared in randomly chosen subsequences from the original sequence. This is difficult to reconcile in any of the models in the literature, especially if the number n of flips is large. Accommodating it in our model requires reinterpreting the primitive: a stream x^n presented to the agent is not necessarily generated by consecutive tosses.

5.2 Belief in Early Switching

A direct consequence of Mean Reversion is a belief in early switching. For instance, consider

$$P(HTH) > P(HHT).$$

Although both sequences reach the same mean by the last toss, an early switch guarantees dominance on path. This is a general consequence of Mean Reversion.[10]

To see that the above example does not hold in general for Local Mean Reversion, consider the case where segments have length 2 (except that $W_1 = \{1\}$). Then $\frac{P(HTHH)}{P(HHTH)} = \frac{g^1(1)g^2(0.5,1)g^3(0.5,0.5)}{g^1(1)g^2(1,1)g^3(0.5,1)} = \frac{g^2(0.5,1)}{g^2(1,1)}\frac{g^3(0.5,0.5)}{g^3(0.5,1)}$. Although $\frac{g^2(0.5,1)}{g^2(1,1)} \geq 1$, there is no restriction on $\frac{g^3(0.5,0.5)}{g^3(0.5,1)}$ since this involves a comparison across different segment means. If g is constant in the second argument then a belief in early switching is recovered.

[10]For an example that holds constant the total number of switches in the two sequences, take $P(HTHH) > P(HHTH)$.

Although the effect does not exist in general under Local Mean Reversion, it exists in a limited form in earlier throws. Consider the largest n s.t. $W_n = \{1, .., n\}$. By definition, the model satisfies Mean Reversion for sequences up to that n, and thus exhibits a belief in early switching there. For instance, if $W_1 = \{1\}$, $W_2 = \{1, 2\}$, $W_3 = \{1, 2, 3\}$ and all subsequent segments exclude the first toss, then we will have $P^3(HTH) > P^3(HHT)$.

5.3 Non-Belief in the Law of Large Numbers

If, as per the Law of Large Numbers, the sampling distribution generated by a fair coin collapses on θ^* as $n \to \infty$, then should it not also collapse on θ^* if the coin is continually self-correcting towards θ^*? Perhaps surprisingly, the answer is no. Intuitively, Mean Reversion requires a negative correlation between outcomes on consecutive tosses i and $i+1$, but a positive correlation between outcomes on i and $i+2$. The evolution of these correlations with i is not restricted by the axioms, and consequently the Law of Large Numbers is not guaranteed.

In the case of $\theta^* = \frac{1}{2}$ and an extreme version of Weak Mean Reversion, where the agent is sure that the outcomes will alternate perfectly after every toss, it is easy to see that the Law of Large Numbers holds. In the next theorem we prove formally that the Law of Large Numbers is not generally implied by Weak Mean Reversion.

Theorem 5 *If $\{P^n\}$ satisfies Weak Mean Reversion then the Law of Large Numbers is not implied, that is, it may not be the case that for all $\epsilon > 0$,*

$$\lim_{n\to\infty} P^n(|\overline{x}^n - \theta^*| > \epsilon) = 0,$$

if the limit exists.

The proof constructs an example of a Weak Mean Reversion model that satisfies Exchangeability and Marginal Consistency, but generically fails MR-Independence.[11] The sampling distribution generated by the Rabin (2002) model necessarily collapses to a degenerate distribution.[12] Our result clarifies that this feature is not driven by the Law of Small Numbers but rather by the model's simplifying assumption that pairs of throws are i.i.d.

A striking finding in the literature is that of *Sample Size Neglect*: subjects do not recognize that sampling variance decreases with sample size.[13] Sample Size Neglect suggests that, in contrast to the

[11] Exchangeability states that for any n and $k \leq n$, all sequences of length n with k heads are deemed equally likely.

[12] The proof is as follows. For any even $n+1$ consider the segments $\{i, i+1\}$ for $i = 1, 3, 5, ..., n$. There are a total of $\frac{n+1}{2}$ segments and each segment generates a segment mean of $\frac{x_i+x_{i+1}}{2}$ that can only take values $\lambda = \frac{1}{2}, 1, 0$. Let $I(\frac{x_i+x_{i+1}}{2} = \lambda)$ denote the indicator function for whether segment $\{i, i+1\}$ generates a mean λ. By the Strong Law of Large Numbers, the limit of the mean of any sequence $x \in \Omega^\infty$ is then

$$\lim_{n\to\infty} \overline{x}^{n+1} = \lim_{n\to\infty} \sum_{i=1,3,..,n} \frac{2}{n+1} \frac{x_i + x_{i+1}}{2}$$

$$= \lim_{n\to\infty} \frac{\sum_{i=1,3,..,n} I(\frac{x_i+x_{i+1}}{2} = \frac{1}{2})}{\frac{n+1}{2}} \frac{1}{2} + \frac{I(\frac{x_i+x_{i+1}}{2} = 1)}{\frac{n+1}{2}} 1 + \frac{I(\frac{x_i+x_{i+1}}{2} = 0)}{\frac{n+1}{2}} 0$$

$$= U^N_{\theta^* N}(\{HT, TH\}) \frac{1}{2} + U^N_{\theta^* N}(\{HH\}) 1 + U^N_{\theta^* N}(\{TT\}) 0.$$

where $U^N_{\theta^* N}$ is the hypergeometric distribution generated by sampling without replacement from an urn with N balls of which $\theta^* N$ are labelled "heads". This limit equals $\frac{1}{2}$ for $\theta^* = \frac{1}{2}$ but in general deviates from θ^*.

[13] Kahneman and Tversky (1972) report an experiment where subjects are told that 45 babies are born per day in a large hospital and 15 babies are born per day in a small hospital, and each hospital has recorded the daily gender distribution over a full year. Subjects were asked which hospital had more days with over 60% boy births. Subjects had to respond "larger hospital", "smaller hospital" or "about the same", and the vast majority believed that both hospitals had a similar number of such days, not recognizing that the variance of the sampling distribution should be higher in the small hospital.

Law of Large Numbers, people believe that the sampling distribution will not concentrate around the population as the sample size gets large – this is referred to as *Non-Belief in the Law of Large Numbers* (Benjamin, Rabin and Raymond (2016)), and Benjamin, Moore and Rabin (2018) provide evidence in an incentivized experiment. The literature has different explanations and models for the Gambler's Fallacy and Sample Size Neglect/Non-Belief in the Law of Large Numbers ((Tversky and Kahneman (1974), Rabin (2002), Benjamin, Rabin and Raymond (2016)). Our theorem, however, suggests that these findings may arise from a single building block, namely, Mean Reversion. We leave it to future research to explore this message.

5.4 Hot-hand Effect

The *Hot Hand Effect* is the finding that subjects sometimes expect a streak to be more likely to continue, in sharp contrast to the Gambler's Fallacy. This has been demonstrated in a sports context (Gilovich, Tversky and Vallone (1985)). To our knowledge there is no evidence of a Hot Hand Effect in the context of coin tosses. The leading explanation for the Hot Hand Effect is that it arises from uncertainty about the true bias of the coin.

We note that the Hot Hand Effect can be generated by a standard well-specified agent engaging in Bayesian inference: if an agent's prior is that the bias of a coin is either 0.5, 0 or 1, then a streak of (say) heads will naturally push the Bayesian posterior towards the high bias, and thus lead to a belief in the continuation of the streak. The intuition in Gilovich, Tversky and Vallone (1985), formalized in Rabin and Vayanos (2010), is the same except that the prior is misspecified due to the Gambler's Fallacy: for any medium or low bias, after seeing a heads the agent expects a tails, but seeing another heads will push her posterior closer to the higher bias. Thus she expects the streak to continue. The Hot Hand Effect can similarly be generated in our model. We eschew a demonstration.

5.5 Disbelief in Patterns

Studies show that subjects believe that a random sequence will not having any discernible systematic patterns, such as the recursions $HTHTHT$ and $HHTTHHTT$ (Wagenaar (1970), Kahneman and Tversky (1972)). This countervails Excessive Alternation which taken to its logical conclusion would favor the maximally alternating sequences $HTHT..$ and $THTH..$

Both Mean Reversion and Local Mean Reversion (with constant segment length 2) regard the recursions HTHTHT.. and THTHTH.. as being *among* the most likely sequences. But since these sequences are recursions, a Disbelief in Patterns would lead to an expectation of some streaks that break the recursion.

The upshot is that a Disbelief in Patterns is not consistent with Mean Reversion nor Local Mean Reversion. We regard some notion of "Order Aversion" to be a psychological building block besides Local Mean Reversion. The question of what constitutes "lack of order" is conceptually difficult and is addressed in the computer science literature on pseudo-randomness. An interesting direction for future research is to take inspiration from the computer science literature in modeling Order Aversion in the context of beliefs about randomness. Kahneman and Tversky (1972) suggest that random-seeming sequences will have the longest descriptions, in that they will not be compressible into short descriptions like "repeat HT three times". This suggests modeling Order Aversion as Kolmogorov complexity, but there are sequences (such as 314159265359 which correspond to π) that are of low Kolmogorov complexity and yet may not be regarded as orderly.

5.6 The St. Petersburg Paradox

The classic justification for Expected Utility is the St. Petersburg paradox where it is observed that people generally would not pay a lot of money to play the following gamble with an infinite expected

value:a fair coin is tossed repeatedly until it lands heads, in which case it pays $\$2^n$, where n is the toss yielding a heads. A strong belief in Mean Reversion would imply that the subjective expected value is not very large.

6 Bayesian Inference

In this section we study whether an agent with misspecified beliefs about randomness can learn the bias of a coin after observing an infinite sequence of i.i.d. outcomes.

6.1 Model

Let $\Theta = \{\theta_1, ..., \theta_n\}$ and let $P^n_\theta(x^n)$ denote the ex-ante probability she assigns to a sequence x^n conditional on the true bias being θ. Define $d_\theta(x^i) := |\overline{x}^n - \theta|$. Suppose it is given by the following model which satisfies Weak Mean Reversion and MR Independence:

$$P^n_\theta(x) = \prod_{i=1}^{n} g^i(d_\theta(x^i), \overline{x}^{i-1}).$$

where we assume that g^i is strictly positive. (so that P^n_θ has full support) and strictly decreasing. Moreover, we assume g^i is independent of θ.

Suppose she has a prior $\mu \in \Delta(\Theta)$ over the parameter. Then, her ex-ante beliefs over sequences of length n is given by

$$P^n(x^n) = \sum_\theta P^n_\theta(x^n)\mu(\theta).$$

Let $P^n(\theta|x^n)$ denote her Bayesian posterior after observing x^n:

$$P^n(\theta|x^n) = \frac{P^n_\theta(x^n)\mu(\theta)}{\sum_{\theta' \in \Theta} P^n_{\theta'}(x^n)\mu(\theta')}.$$

Since for each given θ, the family of beliefs$\{P^n_\theta\}$ is permitted to violate Marginal Consistency, the Bayesian posteriors may be computed with respect to ex-ante beliefs $P^n \in \Delta(\Theta \times \Omega^n)$ that are not consistent in the sense that $P^n(x^n)$ need not be equal to $P^{n+1}(x^n\Omega)$. However, if each of these families satisfies Marginal Consistency, our model reduces to a Bayesian Model that is misspecified in that the true data generating process (the i.i.d. model) is not in the support of the prior.

6.2 Results

Suppose the data is generated by p^θ on $(\Omega^\infty, \Sigma^\infty)$ that is i.i.d. with bias θ^*. We first establish a general property of the model: the agent always places a non-vanishing probability on the true parameter.

Theorem 6 *Assume that conditional on a bias θ, the agent satisfies Weak Mean Reversion and MR Independence, and the corresponding beliefs have full support. Then for any prior over the bias, $\mu \in \Delta(\Theta)$,*

$$\liminf_n P^n(\theta^*|x^n) > 0 \ \ a.s\text{-}p^{\theta^*}.$$

There is no guarantee that posteriors converge along each sequence x. Nevertheless, the theorem establishes that across all streams outside a set of measure 0, the posterior always places a non-vanishing probability on the true parameter. The reason is that the agent believes the sample mean tends to the true parameter at every point of the path, and the Law of Large Numbers ensures that

the agent does not rule out the true parameter. Her misspecified prior, however, may keep her from ruling out other parameters when she sees unexpected patterns along the path.

The result stands in contrast with Rabin (2002), where the agent's beliefs always converge a.s. to a degenerate posterior, but may well be degenerate on the wrong parameter. The reason is that in Rabin's model, for any given bias different from $\frac{1}{2}$, the agent's beliefs predict a different proportion of heads that the one implied by LLN. Therefore, in a learning context, Rabin's agent places probability zero on the true proportion of heads and her beliefs concentrate on the least implausible bias. As Rabin (2002) is a special case of Local Mean Reversion, we conclude that such mislearning is possible under Local Mean Reversion but impossible under Mean Reversion.

The following result provides sufficient conditions on $\{P^n\}$ for $P^n(\theta^*|x^n)$ to converge, providing both a case where the agent learns the truth and a case where she fails to rule out some wrong parameters.

Theorem 7 *Suppose $\mu \in \Delta(\Theta)$ and each $P^n \in \Delta(\Omega^n)$ have full support, and that g^i is strictly decreasing in its first argument and continuous in its second argument for each i.*

1. If $g^i \to c$ uniformly faster than $\frac{1}{n^2} \to 0$,[14] where $c > 0$ is a constant function, then $p^{\theta^*}(\lim_{n\to\infty} P^n(\theta^*|x^n) \in (0,1)) = 1$, that is,

$$0 < \lim_{n\to\infty} P^n(\theta^*|x^n) \neq 1 \text{ a.s.- } p^{\theta^*}.$$

2. If $g^i = g$ for all $i > 1$, then $p^{\theta^*}(\lim_{n\to\infty} P^n(\theta^*|x^n) \to 1) = 1$, that is,

$$\lim_{n\to\infty} P^n(\theta^*|x^n) \to 1 \text{ a.s.- } p^{\theta^*}.$$

Claim (i) assumes that g^i approaches a constant g "fast enough". As noted after Theorem 3, a constant g corresponds to a belief that the bias is constant and equals $\frac{1}{2}$. In the current context, the agent believes that, for every θ, the coin becomes less self-correcting with n, that is, the belief in Mean Reversion weakens with n. As a result, the progression of the sample mean is viewed as less informative about the true parameter, and the posteriors correspondingly become less responsive to the sample mean as n grows. Indeed, posteriors may be critically shaped by what she sees *early* in any sequence x^n. The result states that the agent's posterior beliefs will not converge to a degenerate distribution almost surely. In line with Theorem 6, in the limit the agent places strictly positive probability on the true parameter θ^*. Indeed, patterns observed early in the sequence can never lead the agent to discard the truth. Claim (ii) assumes that $g^i = g$ does not change with i and so the agent maintains a consistent degree of belief in Mean Reversion. In this case the agent eventually learns the true parameter θ^*.

7 Application: Evolutionary Survival of MR vs IID Agents

We study the survival of Mean Reversion (MR) agents in a specific evolutionary context. Consider two populations of agents. The "IID agent" has an accurate perception of i.i.d. sequences. The "MR" agent is specified by the following simple Strong Mean Reversion model: there exists $\epsilon^* > 0$ and $\alpha > \frac{1}{2}$ such that for all $i > 1$, each g^i is given by

$$g^i_{\epsilon^*}(|\overline{x}^i - \theta^*|) = \begin{cases} \frac{\alpha}{Z^i} & |\overline{x}^i - \theta^*| \leq \epsilon^* \\ \frac{1-\alpha}{Z^i} & otherwise \end{cases} \tag{1}$$

where $\{Z^i\}_{i=1,..,n}$ normalize the representation so that each P^i, $i = 1,..,n$ is a probability. Thus, when facing sequence x, the agent "rewards" (in the sense of boosting the probability of the sequence)

[14] That is, there exists N such that for all $n > N$, $|g^n(a,\theta) - c| < \frac{1}{n^2}$ for all a, θ in the support of g^n.

the outcome of flip n by $\alpha > \frac{1}{2}$ if $\overline{x}^n$ is within ϵ^* of θ^*, and otherwise "punishes" it by $1-\alpha$ when determining her belief $P^n(x^n)$.

Suppose there is a continuum of "safe" hunting grounds where hunting yields a small reward $r = 1$ (a "rabbit"). There is a continuum of "risky" hunting grounds of which a fraction $\theta \in \Theta = \{\hat{\theta}, 1-\hat{\theta}\}$, where $\hat{\theta} > \frac{1}{2}$, contains a large reward, $r = 2$ (a "deer"), and the remaining fraction contain no reward, $r = 0$. The fraction θ is an unknown parameter. In every period, one risky ground is randomly chosen and publicly sampled by both MR and IID agent types. The agents update their beliefs about θ based on whether a deer is sighted in the sampling hunting ground. Let $x_i = 0$ (resp. $x_i = 1$) denote that the deer was not present (resp. present), in which case we can write the reward in period i as $r = 2x_i$. Each type $A = IID, MR$ determines the fraction of its population, $k^A_{x^i} \in [0,1]$, that hunts in the risky grounds in period i conditional on having observed x^i signals, which the remainder fraction $1 - k^A_{x^i}$ hunting in the safe ground.[15] Letting Λ^A_{i-1} denote the population of type A at the start of period i. The total reward per capita received by type A is

$$c^A_{x^i} := \frac{R^A_{x^i}}{\Lambda^A_{i-1}} = k^A_{x^i}(2\theta) + (1 - k^A_{x^i}).$$

Both types of agents have a common prior over Θ:

$$\mu_{MR}(\hat{\theta}) = \mu_{IID}(1-\hat{\theta}) = \frac{1}{2}.$$

Both maximize expected utility using a common strictly increasing strictly concave utility index u to determine the optimal $(k^A_{x^i}, 1 - k^A_{x^i})$ based on history of deer sightings x^i from the sampled hunting ground. We assume that agents do not observe the outcome of other agents' hunting. Consequently, they cannot deduce θ by observing the fraction of agents that found a deer.

The population of type A agents grows by a factor of $\lambda^{c^A_{x^i}}$ in period i, where $\lambda > 1$. Thus, higher per capita consumption leads to faster growth in the population. Assuming that both populations start with the same size, we are interested in determining which grows faster over time, that is, we are interested in the ratio:

$$\prod_{i=1}^{n} \frac{\lambda^{c^{MR}_{x^i}}}{\lambda^{c^{IID}_{x^i}}},$$

and in particular how this ratio grows as $n \to \infty$. Note that we have effectively assumed that even if an entire population A hunts in the risky grounds and no deer appears, then that population is not wiped out, but rather it does not grow in that period.

We show that:

Proposition 3 *Denote the true parameter as $\theta \in \Theta = \{\hat{\theta}, 1-\hat{\theta}\}$. Then the following hold for MR agents with $0 < \epsilon^* < 2\hat{\theta} - 1$.*

(i) The MR agents are eventually more confident about the true parameter, a.s.:

$$P^n_{MR}(\theta|x^n) \geq P^n_{IID}(\theta|x^n) \text{ for all sufficiently large } n \in \mathbb{N} \text{ a.s.-}p^\theta$$

(ii) The population of MR agents never vanishes, a.s.:

$$\liminf_{n\to\infty} \prod_{i=1}^{n} \frac{\lambda^{c^{MR}_{x^i}}}{\lambda^{c^{IID}_{x^i}}} > 0 \quad a.s\text{-}p^\theta.$$

[15] An interpretation is that each agent has to choose between spending their full day in the safe or the risky hunting ground, and they randomize by flipping a coin with bias $k^A_{x^i}$ and they go to the risky hunting ground if there is a heads. If all agents do this, then fraction $k^A_{x^i}$ goes to the risky hunting ground.

The first part of the result states that along any realized path x, almost surely, MR agents will eventually be more confident in the true parameter than the IID agents. The reason is that by LLN the sample mean will eventually be within ϵ^* of the true parameter θ, and the MR agents will take this as a stronger indication that the true parameter is θ than not, relative to the IID agents. This leads a larger proportion of the MR population, relative to the IID population, to hunt in the "correct" hunting ground (the correct hunting ground is the risky one iff the true parameter is $\theta = \hat{\theta}$), and therefore to grow faster than the IID population. Accordingly, the second part of the result states that the MR population is never pushed out of the evolutionary race by the IID population.

Since we have already seen (in Theorem 6) that in general MR agents may not become sure about the truth in the limit, it is possible to construct settings where MR agents may not survive relative to IID agents with positive probability. The above result, however, shows us that MR agents do not possess an inherent evolutionary disadvantage that they carry into all possible settings. As long as they can learn, the fact that their beliefs are misspecified relative to IID agents and cause misinferences does not threaten their survival. In fact it can benefit their survival. For instance:

Proposition 4 *If* $\hat{\theta} \geq \frac{3}{4}$ *and* $\epsilon^* = \hat{\theta} - \frac{1}{2}$, *then* $P_{MR}(\liminf_{n\to\infty} \prod_{i=1}^{n} \frac{\lambda^{c^{LSN}}_{x^i}}{\lambda^{c^{IID}}_{x^i}} = \infty) > \frac{1}{2}$.

That is, when $\hat{\theta}$ is high enough, then with probability strictly greater than $\frac{1}{2}$, IID agents lose the evolutionary race against MR agents.

8 Related Literature

The literature on intuitive likelihood judgements investigates both static beliefs (properties of priors) and dynamic beliefs (properties of updating). The evidence on beliefs about randomness is both static (e.g. it asks about the likelihood of HHT vs HHH) and dynamic (e.g. it asks about the likelihood of tails given that HH has already happened). While a misspecified prior is necessary to explain the static evidence, it is also sufficient to explain the dynamic evidence using Bayesian updating: if an agent believes ex-ante that HHT is more likely than HHH, then Bayesian updating leads them to believe that T is more likely than H conditional on HH. Consequently, this paper lies outside the literature on non-Bayesian updating (see for instance Epstein, Noor and Sandroni (2008, 2010)). It intersects with the literature on learning with misspecified beliefs (spawned by Berk (1966)) in that (i) the prior is degenerate on the wrong model and (ii) we have an application to learning (Section 6).

There has been little theoretical work in economics since the seminal work of Rabin (2002).[16] In order to study further how inference with Gambler's Fallacy can lead to the Hot Hand Fallacy, Rabin and Vayanos (2010) study an alternative model where the agent receives a sequence of noisy real-valued signals $s_n = \theta + \epsilon_n$ about a state θ but mistakenly believes that the errors ϵ_n are not i.i.d. and instead exhibit reversals as per the Gambler's Fallacy. Formally, the process ϵ_n is modeled as

$$\epsilon_n = \omega_n - \alpha \sum_{i=1}^{n} \delta^i \epsilon_{n-1-i}$$

where ω_n is i.i.d. normal. Intuitively, a greater number of recent positive realizations of the error make it more likely that the next realization will be negative. This has some flavor of Local Mean Reversion because the highest weights are on recent outcomes. In our model, beliefs are a deterministic function of past outcomes, whereas in Rabin and Vayanos (2010) the stochasticity of ω_n introduces an addition layer of uncertainty.

Motivated by Sample Size Neglect, Benjamin, Rabin and Raymond (2016) hypothesize that people's beliefs may not respect the Law of Large Numbers and in particular may believe in a sampling

[16] See He (2022) for a recent application of the Gambler's Fallacy.

distribution for large samples that has more spread than it should. They write an exchangeable model where the agent believes that the outcome of a coin toss is generated by an i.i.d. stochastic bias θ that has the true bias θ^* as its mean.[17] Their model produces Sample Size Neglect for large samples, which corresponds to what they refer to as a Non-Belief in the Law of Large Numbers.

Benjamin, Rabin and Raymond (2016) speak to Sample Size Neglect, while Rabin (2002) and Rabin and Vayanos (2010) speak to the Gambler's Fallacy and Excessive Alternation. A paper that connects this evidence is Noor (2022), which models the formation of intuitive beliefs by representing the agent's beliefs as a neural network of associations that is trained by her "experience". Noor (2022) shows that if the agent's experience is defined by the sampling distribution generated by the environment, then properties of large-sample sampling distributions are reflected in the agent's beliefs regarding small samples. As a result the model exhibits the Gambler's Fallacy and Sample Size Neglect.

Part of the psychology literature interprets the Gambler's Fallacy in terms of a belief in a *switching rate* that is higher than 50% (Rapoport and Budescu (1997), Bar Hillel and Wagenaar (1991)). Rapoport and Budescu (1997) informally describe a model, the essence of which we can express as follows: presuming that the bias of the coin is perceived to be $\theta^* = \frac{1}{2}$,

$$P^n(x) = \frac{1}{2} \times \prod_{i=1}^{n} \theta^{|x_i - x_{i-1}|}(1-\theta)^{1-|x_i - x_{i-1}|},$$

where the probability of a switch on the i^{th} toss is $\theta > \frac{1}{2}$.[18] The i^{th} outcome is "rewarded" (in the sense of being attributed a higher belief) if $|x_i - x_{i-1}| > 0$, that is, if it differs from the outcome in the previous toss. This model violates Mean Reversion since Mean Reversion predicts that

$$P(HHHHHHTH) < P(HHHHHHTT),$$

while the model requires the reverse ranking because the former sequence contains more switches than the latter. Given that the model is defined for $\theta^* = \frac{1}{2}$, it satisfies Local Mean Reversion where the relevant segment at each $n \geq 2$ has a fixed length of 2.

A Appendix: Proof of Theorem 1

Lemma 1 *A family of full support beliefs $\{P^n\}$ satisfies Independence iff*

$$P^n(x^n) = \prod_{i=1}^{n} (\theta_i)^{x_i}(\gamma_i)^{1-x_i}.$$

In addition, Marginal Consistency holds iff $\theta_n + \gamma_n = 1$ for all n. Furthermore, given Marginal Consistency, Time-Invariant Bias holds iff $\theta_n = \theta_1$ for all n. Finally, given Marginal Consistency and Time-Invariant Bias, Knowledge of Bias holds iff $\theta_n = \theta^*$ for all n.

Proof. Let $\theta_1 = P^1(1)$. By Independence, $\frac{P^2(1,1)}{P^2(0,1)} = \frac{P^1(1)}{P^1(0)} = \frac{\theta^1}{1-\theta^1}$ and $\frac{P^2(1,0)}{P^2(0,0)} = \frac{P^1(1)}{P^1(0)} = \frac{\theta^1}{1-\theta^1}$ Define $\theta_2 := \frac{P^2(1,1)}{\theta^1} = \frac{P^2(0,1)}{1-\theta^1}$ and $\gamma_2 := \frac{P^2(1,0)}{\theta^1} = \frac{P^2(0,0)}{1-\theta^1}$. Then we have

$$P^2(x_1x_2) = (\theta^1)^{x_1}(1-\theta^1)^{1-x_1} \times (\theta_2)^{x_2}(\gamma_2)^{1-x_2}.$$

Moreover by Marginal Consistency, $P^2(1,1) + P^2(1,0) = P^1(1)$ and so $\theta^1\theta_2 + \theta^1\gamma_2 = \theta^1$ and in particular $\theta_2 + \gamma_2 = 1$ given that $\theta^1 > 0$ by the full support assumption. Proceed inductively.

[17] The counterexample constructed in our Theorem 5 also takes this form.

[18] Similar to Rabin (2002), Rapoport and Budescu (1997)'s model switching probability may depend on whether n is even or odd and is characterized by a parameter m that governs the length of throws in which the agent behaves in the standard way. The model we are describing corresponds to their model when $m = 1$.

Assume that the representation holds for n. Invoking Independence as above, $\theta_{n+1} := \frac{P^{n+1}(x^n,1)}{P^n(x^n)}$ and $\gamma_{n+1} := \frac{P^{n+1}(x^n,0)}{P^n(x^n)}$ are independent of x^n, and the same argument establishes that P^{n+1} has the desired representation. Moreover, $\theta_{n+1} + \gamma_{n+1} = 1$ holds iff Marginal Consistency holds.

By the representation above and by repeated application of Time-Invariant Bias, for any $x^{n-1} \in \Omega^{n-1}$,

$$\frac{\theta_n}{1-\theta_n} = \frac{P^n(x^{n-1}1)}{P^n(x^{n-1}0)} = \frac{P^1(1)}{P^1(0)} = \frac{\theta^1}{1-\theta^1}.$$

It follows that $\theta_n = \theta^1$ for all $n \geq 1$.

Finally under Knowledge of Bias, $\theta_1 = \theta^*$. ■

B Appendix: Proof of Theorem 4 and Proposition 1

For each n suppose there is some set $W_n \subset \{1, .., n\}$ of contiguous indices that include n. Say that a family of segments $\{W_n\}_{n\geq 1}$ is regular if $W_{n+1}\backslash\{n+1\} \subseteq W_n$ for all n. While the axioms for the general model are defined with respect to relevant segments, our first lemma considers the counterparts of these axioms for *any* regular family of segments.

Lemma 2 *A family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ satisfies Local Mean Reversion, Local MR Independence and Segment Regularity wrt some regular family of segments $\{W_n\}_{n\geq 1}$ iff for each $n \geq 1$ there exists a continuous $g^n : [0,1]^2 \to (0,1]$ that is weakly decreasing in its first argument such that for any N and $x^N \in \Omega^N$,*

$$P^N(x^N) = \prod_{n=1}^{N} g^n(d_{W_n}(x^n), \overline{x}^{n-1}(W_{n-1})).$$

Proof. Consider a family of full support beliefs. For each n and r, fix some $y^{n,r} \in \Omega^n$ with $\sum_{i\in W_n} y_i^{n,r} = r$. By the full support assumption, $P^n(y^{n,r} > 0)$.

Step 1: For any r and $x^n x_{n+1} \in \Omega^n$with $\overline{x}^n(W_n) = \frac{r}{|W_n|}$,

$$\frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} = \frac{P^{n+1}(y^{n,r} x_{n+1})}{P^n(y^{n,r})}.$$

This just relates the conclusion of Local MR Independence.

Step 2: Show that there exists a function g^{n+1} on $[0,1]^2$that is weakly decreasing in its first argument and for any x^{n+1},

$$g^{n+1}(d_{W_{n+1}}(x^{n+1}), \overline{x}^{n-1}(W_{n-1})) = \frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)}.$$

By Local Mean Reversion, there exists f^ns.t.

$$P^n(x^n) = f^n(d_{W_1}(x^1), ..., d_{W_n}(x^n)|\overline{x}^{n-1}(W_{n-1})),$$

where f^n is weakly decreasing in its first n arguments.[19] Take any n, r and any $x_{n+1} \in \Omega$. Take the corresponding $y^{n,r}$but suppress superscript r in the notation for exposition as needed. Then by the representation,

$$\frac{P^{n+1}(y^{n,r} x_{n+1})}{P^n(y^{n,r})} = \frac{f^{n+1}(d_{W_1}(y^1), ..., d_{W_n}(y^{n+1}), d_{W_{n+1}}(y^{n,r} x_{n+1})|\overline{y}^{n,r}(W_n))}{f^n(d_{W_1}(y^1), ..., d_{W_n}(y^n)|\overline{y}^{n,r-1}(W_{n-1}))}.$$

[19] See Lemma 4 for some of the missing details.

Since n, r and thus $y^{n,r}$ are given, we can write the RHS ratio as a function g^{n+1} so that

$$\frac{P^{n+1}(y^{n,r}x_{n+1})}{P^n(y^{n,r})} = g^{n+1}(d_{W_{n+1}}(y^{n,r}x_{n+1}), \overline{y}^{n,r}(W_n)) > 0.$$

By definition of d_W and by Local Mean Reversion, for any realizations $x_{n+1}, x'_{n+1} \in \Omega$,

$$d_{W_{n+1}}(y^{n,r}x_{n+1}) \geq d_{W_{n+1}}(y^{n,r}x'_{n+1}) \implies P^{n+1}(y^{n,r}x_{n+1}) \leq P^{n+1}(y^{n,r}x'_{n+1})$$
$$\implies g^{n+1}(d_{W_{n+1}}(y^{n,r}x_{n+1}), d_{W_n}(y^{n,r})) \leq g^{n+1}(d_{W_{n+1}}(y^{n,r}x'_{n+1}), \overline{y}^{n,r}(W_n)).$$

Therefore, g^{n+1} is weakly decreasing in its first argument.

To complete the step, take any $x^n x_{n+1} \in \Omega^{n+1}$ with $\overline{x}^n(W_n) = r$. Then, by Segment Regularity, $\overline{x}^{n+1}(W_{n+1}) = \overline{y^{n,r}x_{n+1}}(W_{n+1})$. Step 1 yields

$$g^{n+1}(d_{W_{n+1}}(x^n x_{n+1}), \overline{x}^{n,}(W_n)) = g^{n+1}(d_{W_{n+1}}(y^{n,r}x_{n+1}), \overline{y}^{n,r}(W_n)) = \frac{P^{n+1}(y^{n,r}x_{n+1})}{P^n(y^{n,r})} = \frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)}.$$

Step 3. Complete the proof of sufficiency.

By Segment Regularity, $W_{n+1}\backslash n+1 \subseteq W_n$. Define $g^1(d(x^1)) := P^1(x_1)$. Apply Step 2 iteratively to obtain that for any N and $x^N \in \Omega^N$,

$$P^N(x^N) = \prod_{n=1}^{N} g^n(d_{W_n}(x^n), \overline{x}^{n-1}(W_{n-1})),$$

yielding the desired functional form.

Finally, observe that g^n is defined over a finite subset of $[0,1]^2$ for each $n > 1$ but is weakly decreasing in its first argument for each given value of the second argument. Moreover, by the full support assumption, it takes on strictly positive values. For each possible value of the second argument, the function can clearly be extended to a continuous weakly decreasing function in the first argument. Moreover, it can be continuously extended in its second argument by exploiting the fact that the mixture of decreasing functions is decreasing.

Step 4: Proof of Necessity.

The necessity of Local Mean Reversion is obvious, and Segment Regularity is asserted in the representation. To show that Local MR Independence holds, take any N and $x^N, y^N \in \Omega^N$ s.t. $\overline{x}^N(W_N) = \overline{y}^N(W_N)$. Then $d_{W_{N+1}}(x^N x_{N+1}) = d_{W_{N+1}}(y^N x_{N+1})$ by the Segment Regularity condition. By the representation,

$$P^{N+1}(x^N x_{N+1}) = \left[\prod_{n=1}^{N} g^n(d_{W_n}(x^n), \overline{x}^{n-1}(W_{n-1}))\right] \times g^{N+1}(d_{W_{N+1}}(x^N x_{N+1}), \overline{x}^N(W_N))$$

and similarly

$$P^{N+1}(y^N x_{N+1}) = \left[\prod_{n=1}^{N} g^n(d_{W_n}(y^n), \overline{y}^{n-1}(W_{n-1}))\right] \times g^{N+1}(d_{W_{N+1}}(y^N x_{N+1}), \overline{y}^N(W_N))$$

$$= \left[\prod_{n=1}^{N} g^n(d_{W_n}(y^n), \overline{y}^{n-1}(W_{n-1}))\right] \times g^{N+1}(d_{W_{N+1}}(x^N x_{N+1}), \overline{x}^N(W_N))$$

where the last equality uses $d_{W_{N+1}}(y^N x_{N+1}) = d_{W_{N+1}}(x^N x_{N+1})$ and $\overline{x}^n(W_N) = \overline{y}^n(W_N)$. Therefore, since $P^n(y^n) > 0$ and $P^{n+1}(y^n x_{n+1}) > 0$ by the full support assumption,

$$\frac{P^{N+1}(x^N x_{N+1})}{P^{N+1}(y^N x_{N+1})} = \frac{\prod_{n=1}^{N} g^n(d_{W_n}(x^n), \overline{x}^{n-1}(W_{n-1}))}{\prod_{n=1}^{N} g^n(d_{W_n}(y^n), \overline{y}^{n-1}(W_{n-1}))} = \frac{P^N(x^N)}{P^N(y^N)},$$

as desired. ■

Lemma 3 *Suppose that a family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ admits the representation in Lemma 2. Then beliefs satisfy Marginal Consistency iff*

$$\sum_{x_n} g^n(d_{W_n}(x^{n-1}x_n), \overline{x}^{n-1}(W_{n-1})) = 1.$$

Proof. Given the representation, compute that for any x^N,

$$\frac{\sum_{x_{N+1}} P^{N+1}(x^N x_{N+1})}{P^N(x^N)}$$

$$= \frac{\prod_{n=1}^{N} g^n(d_{W_n}(x^n), \overline{x}^{n-1}(W_{n-1})) \times \sum_{x_{N+1}} g^{N+1}(d_{W_{N+1}}(x^N x_{N+1}), \overline{x}^N(W_N))}{\prod_{n=1}^{N} g^n(d_{W_n}(x^n), \overline{x}^{n-1}(W_{n-1}))}$$

$$= \sum_{x_{N+1}} g^{N+1}(d_{W_{N+1}}(x^N x_{N+1}), \overline{x}^N(W_N)).$$

It follows that Marginal Consistency holds iff $\frac{\sum_{x_{N+1}} P^{N+1}(x^N x_{N+1})}{P^N(x^N)} = 1$ iff $\sum_{x_{N+1}} g^{N+1}(d_{W_{N+1}}(x^N x_{N+1}), \overline{x}^N(W_N))$, as was to be shown. ■

To complete the proof of Theorem 4, define

$$\theta_{n,\overline{x}^{n-1}(W_{n-1})} := g^n(d_{W_n}(x^{n-1}1), \overline{x}^{n-1}(W_{n-1}))$$

$$1 - \theta_{n,\overline{x}^{n-1}(W_{n-1})} := g^n(d_{W_n}(x^{n-1}0), \overline{x}^{n-1}(W_{n-1}))$$

to obtain the desired functional form.

To prove Proposition 1, suppose not. Then there is n s.t. $W_n \neq \hat{W}_n$. Since $W_1 = \{1\}$ in the model, it must be that $n \geq 2$. In this case it is easy to see that we can always find x^n, y^n s.t. $d_{\hat{W}_n}(x^n) = d_{\hat{W}_n}(y^n)$ but $d_{W_n}(x^n) \neq d_{W_n}(y^n)$. But then, given the condition in the proposition, the model fails Local MR Independence wrt $\{\hat{W}_n\}_{n=1}^{\infty}$ since $d_{\hat{W}_n}(x^n) = d_{\hat{W}_n}(y^n)$ but $\frac{P^{n+1}(x^n x_{n+1})}{P^n(x^n)} \neq \frac{P^{n+1}(y^n x_{n+1})}{P^n(y^n)}$.

C Appendix: Proof of Theorems 2 and 3

Lemma 4 *P^n satisfies Weak Mean Reversion iff for each $r \in [0,1]$ there exists a weakly decreasing function $f^n(\cdot|r)$ on $[0,1]^n$ such that for any $x^n \in \Omega^n$,*

$$P^n(x^n) = f^n(d(x^1), ..., d(x^n)|\overline{x}^{n-1}).$$

P^n satisfies Mean Reversion iff the dependence of f^n on $\overline{x}^{n-1}$ is replaced with dependence on $d(x^{n-1})$. P^n satisfies Strong Mean Reversion iff for each f^n is constant in its last argument.

Proof. For any $x, y \in \Omega^n$ s.t. $d(x^{n-1}) = d(y^{n-1})$ and $d(x^i) \leq d(y^i)$ for all $i \leq n$, Mean Reversion implies $P^n(x^n) \geq P^n(x^n)$, with $d(x^i) < d(y^i)$ for some $i \leq n$ implying $P^n(x^n) \geq P^n(x^n)$. Therefore there exists a function $f^n : [0,1]^{n+1} \to [0,1]$ s.t.

$$P^n(x^n) = f^n(d(x^1), ..., d(x^n)|d(x^{n-1})),$$

and f^n is weakly decreasing in all arguments $d(x^i)$ for $i \neq n-1$. Wlog $f^n(\cdot|d(x^{n-1}))$ can be presumed weakly decreasing in all arguments. Conversely, if P^n admits such a representation, then Mean Reversion is implied. A similar argument establishes the desired characterization of Weak Mean Reversion and Strong Mean Reversion. ■

The proof of the general representation result in Theorem 3 obtains from the following Lemma.

Lemma 5 *A family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ satisfies Weak Mean Reversion and MR Independence iff for each $n \geq 1$ there exists a continuous $g^n : [0,1]^2 \to [0,1]$ that is weakly decreasing in its first argument such that for any n and $x^n \in \Omega^n$,*

$$P^n(x^n) = \prod_{i=1}^{n} g^i(d(x^i), \overline{x}^{i-1}).$$

If Mean Reversion is replaced with Mean Reversion (resp. Strong Mean Reversion) then g^n can be written $g^n(d(x^n), d(x^{n-1}))$ (respectively, $g^i(d(x^i))$).

Proof. In the case of Weak Mean Reversion, this is a special case of Lemma 2. If Mean Reversion holds then there exists f^n s.t.

$$P^n(x^n) = f^n(d(x^1), ..., d(x^n)|d(x^{n-1})),$$

and if Strong Mean holds then f^n is constant in the last argument. In either case, the steps of Lemma 2 go through to yield the desired result. ■

Theorem 2 is a corollary of the following Lemma.

Lemma 6 *A family of full support beliefs $\{P^n\}_{n=1}^{\infty}$ satisfies Weak Mean Reversion and MR Independence iff it admits a self-correcting bias representation:*

$$P^n(x^n) = \prod_{i=1}^{n} (\theta_{i,\overline{x}^{i-1}})^{x_i} (\gamma_{i,\overline{x}^{i-1}})^{1-x_i},$$

where $\theta_{i,\overline{x}^{i-1}}, \gamma_{i,\overline{x}^{i-1}} \in (0,1)$ are such that for any $x^{i-1}, y^{i-1} \in \Omega^{i-1}$ s.t. $\overline{x}^{i-1} = \overline{y}^{i-1}$,

$$d(x^{i-1}1) \leq d(y^{i-1}0) \implies \theta_{i,\overline{x}^{i-1}} \geq \gamma_{i,\overline{y}^{i-1}}.$$

Moreover, $\{P^n\}$ also satisfies Strong Mean Reversion (resp. Mean Reversion) iff the representation satisfies: for any $x^{i-1}, y^{i-1} \in \Omega^{i-1}$ (resp. for any $x^{i-1}, y^{i-1} \in \Omega^{i-1}$ s.t. $d(x^{i-1}) = d(y^{i-1})$),

$$d(x^{i-1}1) \leq d(y^{i-1}0) \implies \theta_{i,\overline{x}^{i-1}} \geq \gamma_{i,\overline{y}^{i-1}}$$

$$d(x^{i-1}1) \leq d(y^{i-1}1) \implies \theta_{i,\overline{x}^{i-1}} \geq \theta_{i,\overline{y}^{i-1}}$$

$$d(x^{i-1}0) \leq d(y^{i-1}0) \implies \gamma_{i,\overline{x}^{i-1}} \geq \gamma_{i,\overline{y}^{i-1}}.$$

Finally, $\gamma_{i,\overline{x}^{i-1}} = 1 - \theta_{i,\overline{x}^{i-1}}$ iff $\{P^n\}$ satisfy Marginal Consistency.

Proof. Begin with the representation in Lemma 5 for Weak Mean Reversion and MR Independence. Denote the bias towards heads on the i^{th} throw given a sample mean $\overline{x}^{i-1}$ by

$$\theta_{i,\overline{x}^{i-1}} := g^i(d(x^{i-1}1), \overline{x}^{i-1}) \in [0,1]$$

and similarly for the bias towards tails:

$$\gamma_{i,\overline{x}^{i-1}} := g^i(d(x^{i-1}0), \overline{x}^{i-1}) \in [0,1].$$

The representation can then be written

$$P^n(x^n) = \prod_{i=1}^{n} (\theta_{i,\overline{x}^{i-1}})^{x_i} (\gamma_{i,\overline{x}^{i-1}})^{1-x_i},$$

Since g^i is weakly decreasing in its first argument, the functions θ, γ must have the desired monotonicity property that for any $x^{i-1}, y^{i-1} \in \Omega^{i-1}$ s.t. $\overline{x}^{i-1} = \overline{y}^{i-1}$,

$$d(x^{i-1}1) \leq d(y^{i-1}0) \implies \theta_{i,\overline{x}^{i-1}} \geq \gamma_{i,\overline{y}^{i-1}}.$$

Conversely, if we have such a representation, then for any sequences $x^{i-1}1, y^{i-1}1 \in \Omega^{i-1}$ ending in a heads we have

$$\overline{x}^{i-1} = \overline{y}^{i-1} \text{ and } d(x^{i-1}1) \leq d(y^{i-1}1) \implies \theta_{i,\overline{x}^{i-1}} \geq \theta_{i,\overline{y}^{i-1}},$$

and for any sequences $x^{i-1}0, y^{i-1}0 \in \Omega^{i-1}$ ending in tails we have

$$\overline{x}^{i-1} = \overline{y}^{i-1} \text{ and } d(x^{i-1}0) \leq d(y^{i-1}0) \implies \gamma_{i,\overline{x}^{i-1}} \geq \gamma_{i,\overline{y}^{i-1}}.$$

So we can define functions $f^i(d(x^{i-1}1), \overline{x}^{i-1}) := \theta_{i,\overline{x}^{i-1}}$ and $h^i(d(y^{i-1}0), \overline{y}^{i-1}) := \gamma_{i,\overline{y}^{i-1}}$ that are both weakly decreasing in their first argument. These functions are connected by the condition that

$$\overline{x}^{i-1} = \overline{y}^{i-1} \text{ and } d(x^{i-1}1) = d(y^{i-1}0) \implies \theta_{i,\overline{x}^{i-1}} = \gamma_{i,\overline{y}^{i-1}},$$

in which case $f^i(d(y^{i-1}0), \overline{y}^{i-1}) = f^i(d(x^{i-1}1), \overline{x}^{i-1}) = \theta_{i,\overline{x}^{i-1}} = \gamma_{i,\overline{y}^{i-1}} = h^i(d(y^{i-1}0), \overline{y}^{i-1})$. Therefore $f(\cdot, \overline{x}^{i-1})$ and $h(\cdot, \overline{x}^{i-1})$ coincide on the intersection of their domains. Consequently, together the functions define a weakly decreasing function $g(\cdot, \overline{x}^{i-1})$ on the union of the domains, and we can write

$$P^n(x^n) = \prod_{i=1}^{n} (\theta_{i,\overline{x}^{i-1}})^{x_i} (\gamma_{i,\overline{x}^{i-1}})^{1-x_i} = \prod_{i=1}^{n} g^i(d(x^i), \overline{x}^{i-1}).$$

By Lemma 5, beliefs satisfy Weak Mean Reversion and MR Independence.

The corresponding arguments for Mean Reversion and Strong Mean Reversion are analogous.

Finally, the argument for Marginal Consistency is the same as in Lemma 3. We reproduce it here for the convenience of the reader. Compute that for any x^N,

$$\begin{aligned}
&\frac{\sum_{x_{N+1}} P^{N+1}(x^N x_{N+1})}{P^N(x^N)} \\
&= \frac{\prod_{n=1}^{N} g^n(d(x^n), d(x^{n-1})) \times \sum_{x_{N+1}} \left[g^{N+1}(d(x^N x_{N+1}), d(x^N)) \right]}{\prod_{n=1}^{N} g^n(d(x^n), d(x^{n-1}))} \\
&= \sum_{x_{N+1}} \left[g^{N+1}(d(x^N x_{N+1}), d(x^N)) \right] \\
&= \theta_{i,d(x^{i-1}1),d(x^{i-1})} + \gamma_{i,d(x^{i-1}0),d(x^{i-1})}.
\end{aligned}$$

It follows that Marginal Consistency holds iff $\frac{\sum_{x_{N+1}} P^{N+1}(x^N x_{N+1})}{P^N(x^N)} = 1$ iff $\gamma_{i,d(x^{i-1}0),d(x^{i-1})} = 1 - \theta_{i,d(x^{i-1}1),d(x^{i-1})}$, as was to be shown. ■

The proof of part (ii) of Theorem 3 follows readily by applying the representation to see that

$$\begin{aligned}
P^{n+1}(x^n 1) &= g^{n+1}(d(x^n 1), d(x^n)) \times \prod_{i=1}^{n} g^i(d(x^i), d(x^{i-1})) \\
&= g^{n+1}(d(x^n 1), d(x^n)) P^n(x^n),
\end{aligned}$$

and noting that the marginal belief is given by $P^{n+1}(\Omega^n 1) = \sum_{x^n \in \Omega^n} P^{n+1}(x^n 1) = \sum_{x^n \in \Omega^n} g^{n+1}(d(x^n 1), d(x^n)) P^n(x^n)$.

Finally, for the proof of part (iii) of Theorem 3 assume Strong Mean Reversion and Marginal Consistency. Fix $n > 3$. For Marginal Consistency to hold, it has to be the case that $g^n(d(x^{n-1}1)) +$

$g^n(d(x^{n-1}0)) = 1$ for each x^{n-1}. Take x^{n-1} for which $\min\{d(x^{n-1}1), d(x^{n-1}0)\}$ is closest to 0 and take y^{n-1} for which $\max\{d(y^{n-1}1), d(y^{n-1}0)\}$ is closest to 1. Since $n > 3$, $\max\{d(x^{n-1}1), d(x^{n-1}0)\} \leq \min\{d(y^{n-1}1), d(y^{n-1}0)\}$. Since g^n is weakly decreasing, we therefore cannot have $g^n(d(x^{n-1}1)) + g^n(d(x^{n-1}0)) = 1 = g^n(d(y^{n-1}1)) + g^n(d(y^{n-1}0))$ unless $g^n(d(x^{n-1}1)) = g^n(d(x^{n-1}0)) = g^n(d(y^{n-1}1)) = g^n(d(y^{n-1}0)) = \frac{1}{2}$.

D Proof of Theorem 5

Proof. Assume Weak Mean Reversion, Marginal Consistency and Exchangeability. Given Marginal Consistency, de Finetti's theorem (see for instance Diaconis and Freedman (1980)) ensures that Exchangeability implies the representation

$$P(x^n) = \int_{[0,1]} \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} d\mu(\theta),$$

for some prior over the bias, $\mu \in \Delta[0,1]$. We derive the condition on the prior that characterizes Weak Mean Reversion. Due to Exchangeability, Weak Mean Reversion is equivalent to the condition that for any n and $x, y \in \Omega^n$ s.t. $\overline{x}^{n-1} = \overline{y}^{n-1}$,

$$d(x^n) \leq d(y^n) \implies P^n(x) \geq P^n(y).$$

Suppose x^n, y^n are such that that there are k heads in the first $n-1$ tosses (so that $\overline{x}^{n-1} = \overline{y}^{n-1} = \frac{k}{(n-1)}$) and moreover $x_n = 1$ and $y_n = 0$. Compute that

$$P(x^n) = \int_{[0,1]} \theta^{k+1}(1-\theta)^{n-1-k} d\mu(\theta) = \int_{[0,1]} \frac{\theta}{(1-\theta)} \theta^k (1-\theta)^{n-k} d\mu(\theta)$$

$$\text{and } P(y^n) = \int_{[0,1]} \theta^k (1-\theta)^{n-k} d\mu(\theta).$$

Then

$$P(x^n) \geq P(y^n)$$

$$\iff \int_{[0,1]} \frac{\theta}{(1-\theta)} \theta^k (1-\theta)^{n-k} d\mu(\theta) \geq \int_{[0,1]} \theta^k (1-\theta)^{n-k} d\mu(\theta)$$

$$\iff \int_{[0,1]} \frac{2\theta - 1}{1-\theta} \theta^k (1-\theta)^{n-k} d\mu(\theta) \geq 0.$$

Therefore Weak Mean Reversion holds iff it is the case that

$$d(x^n) \leq d(y^n) \implies \int_{[0,1]} \frac{2\theta - 1}{1-\theta} \theta^k (1-\theta)^{n-k} d\mu(\theta) \geq 0.$$

and

$$d(x^n) \geq d(y^n) \implies \int_{[0,1]} \frac{2\theta - 1}{1-\theta} \theta^k (1-\theta)^{n-k} d\mu(\theta) \leq 0.$$

To give an example of such a model suppose that μ has support $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ with $0 < \mu(\frac{1}{4}) = \mu(\frac{3}{4}) < \mu(\frac{1}{2})$. Then

$$\int_{[0,1]} \frac{2\theta - 1}{1-\theta} \theta^k (1-\theta)^{n-k} d\mu(\theta) \geq 0$$

$$\iff \left[-\frac{2}{3} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{n-k} + 2 \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{n-k} \right] \mu(\frac{1}{4}) \geq 0$$

$$\iff 2\left(\frac{3}{4}\right)^k\left(\frac{1}{4}\right)^{n-k} \geq \frac{2}{3}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k}$$

$$\iff 3^{k+1} \geq 3^{n-k}$$

$$\iff 2k \geq n-1$$

$$\iff \frac{k}{(n-1)} \geq \frac{1}{2}$$

$$\iff \overline{x}^{n-1} \geq \frac{1}{2}.$$

To complete the proof. Recall that x^n, y^n are such that $x_n = 1$ and $y_n = 0$. Then $d(x^n) \leq d(y^n)$ (resp. $d(x^n) \geq d(y^n)$) then it must be that $\overline{x}^{n-1} \geq \frac{1}{2}$ (resp. $\overline{x}^{n-1} \leq \frac{1}{2}$), and thus we obtain $\int_{[0,1]} \frac{2\theta-1}{1-\theta}\theta^k(1-\theta)^{n-k}d\mu(\theta) \geq 0$ (resp. $\int_{[0,1]} \frac{2\theta-1}{1-\theta}\theta^k(1-\theta)^{n-k}d\mu(\theta) \leq 0$). ■

E Appendix: Proof of Theorems 6 and 7

Lemma 7 *Assume $\mu \in \Delta(\Theta)$ and each $P^n \in \Delta(\Omega^n)$ have full support. Then,*

$$\liminf_n P^n(\theta^*|x^n) > 0 \text{ a.s-}p^{\theta^*}.$$

Proof. Fix any sequence such that $\lim_n \overline{x}^n = \theta^*$ and let x^{n_k} denote a subsequence that converges to lim inf of $P^n(\theta^*|x^n)$:

$$lim_{k\to\infty} P^{n_k}_{\theta^*}(x^{n_k}) = \liminf_n P^n(\theta^*|x^n).$$

For any $\theta \neq \theta^*$, this subsequence generates a sequence $\{P^{n_k}_\theta(x^{n_k})\}$ in $[0,1]$, and a further subsequence must lead to convergence of $P^{n_k}_\theta(x^{n_k})$. Since there are finitely many θ, we can wlog suppose that $P^{n_k}_\theta(x^{n_k})$ are convergent for all $\theta \in \Theta$. Due to the full support assumptions, it must be that $\sum_{\theta\in\Theta} P^{n_k}_\theta(x^{n_k})\mu(\theta) > 0$ and in particular the posteriors $P^{n_k}(\theta|x^{n_k})$ are well-defined.

Suppose by way of contradiction that $\liminf_n P^n(\theta^*|x^n) = 0$. Thus $P^{n_k}_{\theta^*}(x^{n_k}) \to 0$. It cannot be that $P^{n_k}_\theta(x^{n_k}) \to 0$ for all $\theta \in \Theta$, otherwise we obtain the contradiction that $1 = \sum_{\theta\in\Theta} P^n(\theta|x^n) \to 0$. Let $\theta \in \Theta$ be such that $lim_{k\to\infty} P^{n_k}_\theta(x^{n_k}) > 0$ and consider the likelihood ratio of θ and θ^*,

$$\frac{P^n_\theta(x^n)}{P^n_{\theta^*}(x^n)} = \frac{\prod_{i=1}^n g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{\prod_{i=1}^n g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} = \prod_{i=1}^n \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}.$$

By the Law of Large Numbers, $p^{\theta^*}(x^\infty|\lim_{n\to\infty}\overline{x}^n = \theta^*) = 1$. Hence, it is enough to consider such sequences. Fix $\epsilon = \min_{\theta\neq\theta'}|\theta - \theta'|$ and $x \in \Omega^\infty$ such that $\lim_{n\to\infty}\overline{x}^n = \theta^*$. Let N be such that for all $n > N$, $|\overline{x}^n - \theta^*| < \frac{\epsilon}{4}$. Then, $|\overline{x}^n - \theta| > \frac{\epsilon}{2}$. Further,

$$\prod_{i=1}^n \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} = \prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} \times \prod_{i=N}^n \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}.$$

Because g is weakly decreasing in its first argument, and since $|\overline{x}^n - \theta^*| < \frac{\epsilon}{4}$ and $|\overline{x}^n - \theta| > \frac{\epsilon}{2}$, then

$$g^i(|\overline{x}^i - \theta^*|, \overline{x}^i) \geq g^i(|\overline{x}^i - \theta|, \overline{x}^i)$$

for all $i > N$. Hence,

$$\lim_{n_k\to\infty} \frac{P^{n_k}_\theta(x^{n_k})}{P^{n_k}_{\theta^*}(x^{n_k})} \leq \prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}$$

which contradicts the hypothesis that $P^{n_k}_{\theta^*}(x^{n_k}) \to 0$ and in particular contradicts $\liminf_n P^n(\theta^*|x^n) = 0$. ■

Lemma 8 *Suppose $\mu \in \Delta(\Theta)$ and each $P^n \in \Delta(\Omega^n)$ have full support. Assume Mean Reversion and MR Independence, and consider a representation where g^i is strictly decreasing in its first argument for each i.*

1. If $g^i = g$ for all $i >$ and g is strictly decreasing in its first argument and continuous in its second, then $p^{\theta^*}(\lim_{n\to\infty} P^n(\theta^*|x^n) \to 1) = 1$, that is,

$$\lim_{n\to\infty} P^n(\theta^*|x^n) \to 1 \text{ a.s.- } p^{\theta^*}.$$

2. If $g^i \to c$ uniformly faster than $\frac{1}{n^2} \to 0$ for all θ, where g^i is strictly decreasing in its first argument and continuous in its second argument for all i and $c > 0$ is a constant function, then

$$0 < \lim_{n\to\infty} P^n(\theta^*|x^n) \neq 1 \text{ a.s.- } p^{\theta^*}.$$

Proof. Because we are only considering finitely many θ's, it is enough to show that $\frac{P^n_\theta(x^n)}{P^n_{\theta^*}(x^n)} \to 0$ a.s.- p^{θ^*} for all $j \neq i$.

By an identical argument to the one in Lemma 7, for $\epsilon = \min_{\theta\in\Theta\setminus\{\theta^*\}} |\theta - \theta^*|$, there exists N such that for all $i > N, |\overline{x^i} - \theta^*| < \frac{\epsilon}{4}$, and

$$\frac{P^n_\theta(x^n)}{P^n_{\theta^*}(x^n)} = \prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} \times \prod_{i=N}^{n} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}.$$

Further, because g is strictly decreasing in its first argument, and since $|\overline{x}^n - \theta^*| < \frac{\epsilon}{4}$ and $|\overline{x}^n - \theta| > \frac{\epsilon}{2}$,

$$\prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} \times \prod_{i=N}^{n} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} < \prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} \times \prod_{i=N}^{n} \frac{g^i(\frac{\epsilon}{2}, \overline{x}^i)}{g^i(\frac{\epsilon}{4}, \overline{x}^i)}.$$

Notice that $\frac{g^i(\frac{\epsilon}{2}, \overline{x}^i)}{g^i(\frac{\epsilon}{4}, \overline{x}^i)} < 1$ for all i and the term $\prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}$ does not depend on n.

Therefore, to prove the result, it suffices to show that $\prod_{i=N}^{n} \frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} \to 0$.

Step 1: Establish the result under the first assumption in the lemma.

A sufficient condition for the result $\lim_{n\to\infty} P^n(\theta^*|x^n) \to 1$ *a.s.-*p^{θ^*} is that $\prod_{i=N}^{n} \frac{g^i(\frac{\epsilon}{2}, \overline{x}^i)}{g^i(\frac{\epsilon}{4}, \overline{x}^i)} \to 0$. To see this observe that, given the preceding, $\prod_{i=N}^{n} \frac{g^i(\frac{\epsilon}{2}, \overline{x}^i)}{g^i(\frac{\epsilon}{4}, \overline{x}^i)} \to 0$ implies

$$\frac{P^n_\theta(x^n)}{P^n_{\theta^*}(x^n)} < \prod_{i=1}^{N-1} \frac{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)} \times \prod_{i=N}^{n} \frac{g^i(\frac{\epsilon}{2}, \overline{x}^i)}{g^i(\frac{\epsilon}{4}, \overline{x}^i)} \to 0$$

and so $\frac{P^n_\theta(x^n)}{P^n_{\theta^*}(x^n)} \to 0$, and in particular $\lim_{n\to\infty} P^n(\theta^*|x^n) \to 1$ *a.s.-*p^{θ^*}.

So consider the first assumption in the lemma. Since the assumption restricts $g^i = g$ for $i > 1$, we take $N > 1$. We show that $\prod_{i=N}^{n} \frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} \to 0$. Since g is continuous in its second argument and since $\overline{x}^i \to \theta^*$, we have $\frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} \to \frac{g(\frac{\epsilon}{2}, \theta^*)}{g(\frac{\epsilon}{4}, \theta^*)} < 1$. In particular there exists M and $\overline{\epsilon} > 0$ s.t. $\frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} < 1 - \overline{\epsilon}$ for all $i > M$. But then

$$\prod_{i=N}^{n} \frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} = \prod_{i=N}^{max\{M,N\}} \frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} \times \prod_{i=max\{M,N\}+1}^{n} \frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)}$$

$$< \prod_{i=N}^{max\{M,N\}} \frac{g(\frac{\epsilon}{2}, \overline{x}^i)}{g(\frac{\epsilon}{4}, \overline{x}^i)} \times \prod_{i=max\{M,N\}+1}^{n} (1 - \overline{\epsilon}) \to 0,$$

as desired.

Step 2: Establish the result under the second assumption in the lemma.

Assume that $g^n \to c > 0$ uniformly faster than $\frac{1}{n^2} \to 0$. Then there exists K such that $\inf_{n>K} g^n > 0$ for all $n > K$. Moreover, for any $a, b \in [0,1]$ and sample mean θ such that $a > b > 0$, it must be that for all $n > K$, and

$$|\frac{g^n(a,\theta)}{g^n(b,\theta)} - 1| = |\frac{g^n(a,\theta) - g^n(b,\theta)}{g^n(b,\theta)}| < |\frac{g^n(a,\theta) - c}{g^n(b,\theta)}| + |\frac{g^n(b,\theta) - c}{g^n(b,\theta)}|$$

$$< \frac{1}{n^2}\frac{2}{g^n(b,\theta)} < \frac{k}{n^2}$$

for some constant $k = \frac{2}{\inf_{n>K} g^n}$. In fact $k > 1$ since $g \leq 1$. Hence, for all $n > K$,

$$\frac{g^n(a,\theta)}{g^n(b,\theta)} < 1 + \frac{k}{n^2}.$$

Fix any $x \in \Omega^\infty$ and consider

$$\lim_{n\to\infty} \frac{P^{\theta^*}(x^n)}{P^\theta(x^n)} = \lim_{n\to\infty} \prod_{i=1}^{n} \frac{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta|, \overline{x}^i)} = \prod_{i=1}^{\infty} \frac{g^i(|\overline{x}^i - \theta^*|, \overline{x}^i)}{g^i(|\overline{x}^i - \theta|, \overline{x}^i)}.$$

This product exists if and only if there exists N such that for all $m > N$,

$$\sum_{n=m}^{\infty} ln(\frac{g^n(|\overline{x}^n - \theta^*|, \overline{x}^n)}{g^n(|\overline{x}^n - \theta|, \overline{x}^n)}) < \infty,$$

which we shall proof happens a.s.-p^{θ^*}. Since the law of large numbers implies $\overline{x}^n \to \theta^*$, there is M s.t. $|\overline{x}^n - \theta^*| < |\overline{x}^n - \theta|$ and thus $\frac{g^n(|\overline{x}^n - \theta^*|, \overline{x}^n)}{g^n(|\overline{x}^n - \theta|, \overline{x}^n)} > 1$ (since g^n is strictly decreasing in its first argument) for all $n \geq M$. Also, as we saw earlier, there is K such that $\frac{g^n(a,\theta)}{g^n(b,\theta)} < 1 + \frac{k}{n^2}$ for all $n > K$ and any a, b, θ. It follows that for all $n > N := max\{M, K\}$,[20]

$$\sum_{n=N+1}^{\infty} ln(\frac{g^n(|\overline{x}^n - \theta^*|, \overline{x}^n)}{g^n(|\overline{x}^n - \theta|, \overline{x}^n)}) < \sum_{n=N+1}^{\infty} ln(1 + \frac{k}{n^2}) < \infty.$$

Therefore, we establish $\lim_{n\to\infty} \frac{P^{\theta^*}(x^n)}{P^\theta(x^n)} < \infty$, and in particular $P(\theta^*|x^n) \not\to 1$ a.s.-p^{θ^*}. Moreover, since $\frac{P^{\theta^*}(x^n)}{P^\theta(x^n)} > 0$ for any n by the full support assumption, and since we have shown that $\frac{g^n(|\overline{x}^n - \theta^*|, \overline{x}^n))}{g^n(|\overline{x}^n - \theta|, \overline{x}^n)} > 1$ for all $n > N$, it must be that $\lim_{n\to\infty} \frac{P^{\theta^*}(x^n)}{P^\theta(x^n)} = \prod_{i=1}^{\infty} \frac{g^i(u(|\sum_{j\leq i} \frac{x_i}{i} - \theta^*|))}{g^i(u(|\sum_{j\leq i} \frac{x_i}{i} - \theta))} > 0$. Thus, $P(\theta^*|x^n) > 0$ a.s.-p^{θ^*}. ■

F Appendix: Evolution

F.0.1 Proof of Proposition 3

We start with a convenient observation about the normalizing constants Z^i_θ and $Z^i_{1-\theta}$ in the representation.

[20] To see why the inequality $\sum_{n=1}^{\infty} ln(1 + \frac{k}{n^2}) < \infty$ in the expression holds, let $f(x) = ln(1 + \frac{k}{x^2})$ and note that it is decreasing on $(0, \infty)$. Then

$$\sum_{n=1}^{\infty} ln(1 + \frac{k}{n^2}) < f(1) + \int_1^{\infty} f(x) = f(1) + (2\sqrt{k})tan^{-1}(\sqrt{k}) - ln(k+1) < \infty.$$

Lemma 9 *Let Z^i_θ be the constant associated with the representation 1 for parameter θ. Then, $Z^i_\theta = Z^i_{1-\theta}$ for all $i > 1$ and all $\theta \in [0,1]$.*

Proof. WLOG fix $\theta \geq \frac{1}{2}$. For each sequence $x^n \in \Omega^n$,let $y(x^n)$ be the sequence obtained by replacing the ones in $x^n$with zeros and the zeros with ones. It is easy to see that $|\overline{x}^n - \theta| = |\overline{y(x^n)} - (1-\theta)|$. Hence, $|\overline{x}^n - \theta| \leq \epsilon^* \iff |\overline{y(x^n)} - (1-\theta)| \leq \epsilon^*$ which implies $P^n_\theta(x^n) = P^n_{1-\theta}(y(x^n))$ by the representation (1). In particular, $P^n_\theta(1^n) = P^n_{1-\theta}(0^n)$. We show that for any $n \geq 2$,

$$\prod_{i=2}^{n} \frac{1}{Z^i_\theta} = \prod_{i=2}^{n} \frac{1}{Z^i_{1-\theta}}.$$

Consider two cases:

(i) $|1-\theta| > \epsilon^*$.

Then $\theta(1-\alpha)^{n-1}\prod_{i=2}^n \frac{1}{Z^i_\theta} = P^n_\theta(1^n) = P^n_{1-\theta}(0^n) = (1-(1-\theta))(1-\alpha)^{n-1}\prod_{i=2}^n \frac{1}{Z^i_{1-\theta}} \implies \prod_{i=2}^n \frac{1}{Z^i_\theta} = \prod_{i=2}^n \frac{1}{Z^i_{1-\theta}}$.

(ii) $|1-\theta| \leq \epsilon^*$.

Then $\theta\alpha^{n-1}\prod_{i=2}^n \frac{1}{Z^i_\theta} = P^n_\theta(1^n) = P^n_{1-\theta}(0^n) = (1-(1-\theta))\alpha^{n-1}\prod_{i=2}^n \frac{1}{Z^i_{1-\theta}} \implies \prod_{i=2}^n \frac{1}{Z^i_\theta} = \prod_{i=2}^n \frac{1}{Z^i_{1-\theta}}$.

Using the equalities $\prod_{i=2}^n \frac{1}{Z^i_\theta} = \prod_{i=2}^n \frac{1}{Z^i_{1-\theta}}$ for all $n \geq 2$, a proof by induction yields $Z^i_\theta = Z^i_{1-\theta}$ for all $i > 1$. ■

Next, observe that by Bayesian updating, if there is N s.t. $\overline{x}^n > \frac{1}{2}$ and $|\overline{x}^n - \theta| < \epsilon^*$ for all $n > N$, then

$$P^n_{MR}(\theta|x^n) = \frac{P^n_{MR}(\theta|x^N)\alpha^{n-N}}{P^n_{MR}(\theta|x^N)\alpha^{n-N} + P^n_{MR}(1-\theta|x^N)(1-\alpha)^{n-N}}$$

and as usual,

$$P^n_{IID}(\theta|x^n) = \frac{P^n_{IID}(\theta|x^N)(\theta)^{k_{n-N}}(1-\theta)^{n-N-k_{n-N}}}{P^n_{IID}(\theta|x^N)(\theta)^{k_{n-N}}(1-\theta)^{n-N-k_{n-N}} + P^n_{IID}(1-\theta|x^N)(1-\theta)^{k_{n-N}}(\theta)^{n-N-k_{n-N}}},$$

where k_{n-N} is the number of heads that occur after the N^{th} throw (which satisfies $2k_{n-N} > n-N$ since that $\overline{x}^n > \frac{1}{2}$ for all $n > N$). Now we are ready to prove the proposition.

Proof of (i): We only establish the case in which $\theta = \hat{\theta} > \frac{1}{2}$ since the proof $\theta = 1-\hat{\theta}$ is analogous. Let $P^n_{MR}(\theta|x^n)$ and $P^n_{IID}(\theta|x^n)$ be the posterior beliefs after observing signals x^n of the MR and IID agents respectively.

By LLN, there exists N such that $\overline{x}^n > \frac{1}{2}$ and $|\overline{x}^n - \theta| < \epsilon^*$ for all $n > N$. Moreover, by Bayesian updating,

$$\frac{P^n_{MR}(\theta|x^n)}{P^n_{MR}(1-\theta|x^n)} = \frac{P^n_{MR}(\theta|x^N)}{P^n_{MR}(1-\theta|x^N)}\frac{\alpha^{n-N}}{(1-\alpha)^{n-N}} \text{ and } \frac{P^n_{IID}(\theta|x^n)}{P^n_{IID}(1-\theta|x^n)} = \frac{P^n_{IID}(\theta|x^N)}{P^n_{IID}(1-\theta|x^N)}\frac{(\theta)^{k_{n-N}}(1-\theta)^{n-N-k_{n-N}}}{(1-\theta)^{k_{n-N}}(\theta)^{n-N-k_{n-N}}},$$

where k_{n-N} is the number of heads that occur after the N^{th} throw. Then

$$\frac{P^n_{MR}(\theta|x^n)}{P^n_{MR}(1-\theta|x^n)} \Big/ \frac{P^n_{IID}(\theta|x^n)}{P^n_{IID}(1-\theta|x^n)} = \frac{P^n_{MR}(\theta|x^N)}{P^n_{MR}(1-\theta|x^N)}\frac{\alpha^{n-N}}{(1-\alpha)^{n-N}} \Big/ \frac{P^n_{IID}(\theta|x^N)}{P^n_{IID}(1-\theta|x^N)}\frac{(\theta)^{k_{n-N}}(1-\theta)^{n-N-k_{n-N}}}{(1-\theta)^{k_{n-N}}(\theta)^{n-N-k_{n-N}}}$$

$$= \left[\frac{P^n_{MR}(\theta|x^N)}{P^n_{MR}(1-\theta|x^N)} \Big/ \frac{P^n_{IID}(\theta|x^N)}{P^n_{IID}(1-\theta|x^N)}\right]\left[\frac{\alpha}{1-\alpha}\right]^{n-N}\left[\frac{\theta}{1-\theta}\right]^{n-N-2k_{n-N}} \to \infty,$$

since $\frac{\alpha}{1-\alpha}, \frac{\theta}{1-\theta} > 1$.

Hence, there exists N' such that for all $n > N'$,

$$\frac{P^n_{MR}(\theta|x^n)}{P^n_{MR}(1-\theta|x^n)} > \frac{P^n_{IID}(\theta|x^n)}{P^n_{IID}(1-\theta|x^n)},$$

which implies $P^n_{MR}(\theta|x^n) > P^n_{IID}(\theta|x^n)$ for all $n > N'$, as desired.

Proof of (ii): Let $P^n_{MR}(\theta|x^n)$ and $P^n_{IID}(\theta|x^n)$ be the posterior beliefs after observing signals x^n of MR and IID agents respectively.

Step 1: Show that

$$P^n_{MR}(\theta|x^n) \geq P^n_{IID}(\theta|x^n) \iff k^{MR}_{x^n} \geq k^{IID}_{x^n}.$$

The optimal choice of agent A solves

$$U_{x^i}(k) = [\theta P^n(\theta|x^n)+(1-\theta)P^n(1-\theta|x^n)]u(k2+(1-k))+[(1-\theta)P^n(\theta|x^n)+\theta P^n(1-\theta|x^n)]u(1-k).$$

The FOC is therefore

$$2[\theta P^n(\theta|x^n) + (1-\theta)P^n(1-\theta|x^n)]u'(k+1) = [(1-\theta)P^n(\theta|x^n) + \theta P^n(1-\theta|x^n)]u'(1-k),$$

which rearranges to

$$2\frac{u'(k+1)}{u'(1-k)} = \frac{1-\theta+\frac{1}{2}P^n(1-\theta|x^n)}{1-\theta+\frac{1}{2}P^n(\theta|x^n)}.$$

Since u is strictly concave, the LHS is strictly decreasing in k. Therefore $P^n_{MR}(\theta|x^n) \geq P^n_{IID}(\theta|x^n) \iff k^{MR}_{x^n} \geq k^{IID}_{x^n}$.

Step 2: Prove the result.

By step 1 and part (i) of the proposition, the MR agent will eventually spend more time on the risky ground than the safe ground a.s. Furthermore, when the true parameter is $\theta = \hat{\theta} > \frac{1}{2}$ (the argument for $\theta = 1 - \hat{\theta}$ is analogous), there will eventually be more heads than tails a.s. by LLN. Hence, a.s., the MR agent will eventually have a higher return than the IID agent.

F.0.2 Proof of Proposition 4

The proof relies on the following statistical fact: in an infinite sequence of coin tosses, the probability of observing more heads than tails at every toss is equal to $2\theta - 1$ whenever $\theta > \frac{1}{2}$.

Lemma 10 *Assume $\theta > \frac{1}{2}$ and let $E = \{x^\infty | \overline{x}^n > \frac{1}{2}$ for all $n\}$. Then $p^\theta(E) = 2\theta - 1$.*

Proof. From the *Gambler's Ruin Problem,* we know that conditional on the first throw being heads, the probability of always having more heads than tails is equal to $1 - \frac{1-\theta}{\theta}$.[21] Hence, the probability of *always* having more heads than tails is equal to $\theta(1 - \frac{1-\theta}{\theta}) = 2\theta - 1$. ■

Suppose that $\theta = \hat{\theta}$ (the argument is analogous for $\theta = 1 - \hat{\theta}$). Consider

$$E = \{x^\infty | \overline{x^n} > \frac{1}{2} \text{ for all n} \in \mathbb{N}\},$$

the event where the sample mean exceeds $\frac{1}{2}$ at every i. Under our assumptions on $\hat{\theta}$ and ϵ^*, for any sequence $x \in E$, we will have that $|\overline{x}^n - \theta| < \epsilon^*$ for all n. Hence from the proof of Proposition 3, we see that for each $x \in E$, the MR population grow at a faster rate than the IID population in every period, that is $\frac{\lambda^{c^{MR}_{x^i}}}{\lambda^{c^{IID}_{x^i}}} > 1$ for each i. Therefore $\lim_{n\to\infty} \prod_{i=1}^n \frac{\lambda^{c^{MR}_{x^i}}}{\lambda^{c^{IID}_{x^i}}} > 1$.

[21] In the Gambler's Ruin Problem literature, this probability is referred to as the probability of "never going broke" and $1 - \frac{1-\theta}{\theta}$ is the value for the case in which the gambler starts with one dollar.

By lemma 10, $p^\theta(E) = 2\theta - 1$. Consider also the event $F = \{1\} \times \{0\} \times E$, where the first toss yields a head, the second a tails and the subsequent stream belongs to E. This event has probability $(1-\theta)\theta p^\theta(E) > 0$. Note that after the heads on the first toss, $P^{MR}(\theta|1) = P^{IID}(\theta|1)$, after the tails on the second toss, $P^{MR}(\theta|1,0) = \frac{\theta\alpha}{\theta\alpha+\theta(1-\alpha)} = \theta > \frac{1}{2} = P^{IID}(\theta|1,0)$, and then for all subsequent tosses we have $P^{MR}(\theta|x^i) > P^{IID}(\theta|x^i)$ since $\overline{x}^n > \frac{1}{2}$ for all $n > 2$. That is, the MR agents are weakly more confident about the true parameter than the IID agents for $i = 1$ and strictly so for all $i > 1$. Consequently, by step 1, $\lim_{n\to\infty} \prod_{i=1}^n \frac{\lambda^{c^{LSN}}_{x^i}}{\lambda^{c^{IID}}_{x^i}} > 1$ for each $x \in F$. Conclude that $p^\theta(lim_{n\to\infty} \prod_{i=1}^n \frac{\lambda^{c^{LSN}}_{x^i}}{\lambda^{c^{IID}}_{x^i}} > 1) \geq p^\theta(E \cup F) > \frac{1}{2}$.

References

[1] Acemoglu, D., V. Chernozhukov and M. Yildez (2016): "Fragility of Asymptotic Agreement under Bayesian Learning", *Theoretical Economics* 11, pp. 1555-7561.

[2] Bar-Hillel, M. and W. Wagenaar (1991): "The Perception of Randomness", *Advances in Applied Mathematics* 12(4), pp. 428–454.

[3] Benjamin, D. (2019): "Errors in Probabilistic Reasoning and Judgment Biases." Chapter for the *Handbook of Behavioral Economics* (edited by Doug Bernheim, Stefano DellaVigna, and David Laibson). Elsevier Press.

[4] Benjamin, D., D. Moore and M. Rabin (2018): "Biased Beliefs About Random Samples: Evidence from Two Integrated Experiments", working paper.

[5] Benjamin, D., M. Rabin and C. Raymond (2016): "A Model of Non-Belief in the Law of Large Numbers", *Journal of the European Economic Association* 14(2), pp 515-544.

[6] Berk, R. (1966): "Limiting Behavior of Posterior Distributions when the Model is Incorrect", *Annals of Mathematical Statistics* 37, pp. 51-58.

[7] Chen, D., T. Moskowitz and K. Shue (2016): "Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers and Baseball Umpires", *Quarterly Journal of Economics* 131, pp. 1181-1242.

[8] Diaconis, P., and D. Freedman (1980): "Finite Exchangeable Sequences", *The Annals of Probability* 8(4), pp. 745-764.

[9] Epstein, L. (2006): "An Axiomatic Model of Non-Bayesian Updating", *Review of Economic Studies* 73, pp. 413-436.

[10] Epstein, L., J. Noor and A. Sandroni (2008): "Non-Bayesian Updating: A Theoretical Framework", *Theoretical Economics* 3(2), pp. 193-229.

[11] Epstein, L., J. Noor and A. Sandroni (2010): "Non-Bayesian Learning", *B.E. Journal of Theoretical Economics (Advances)* 10(1), article 3.

[12] Freedman, D. (1965): "Bernard Friedman's Urn", *The Annals of Mathematical Statistics* 36(3), pp. 956-970.

[13] Friedman, B. (1949): "A Simple Urn Model", *Communications on Pure and Applied Math* 2, pp. 59-70.

[14] Gigerenzer, G. (1996): "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky" Psychological Review 103, pp. 592-596.

[15] Gilovich, T.; Tversky, A. and Vallone, R. (1985): "The Hot Hand in Basketball: On the Misperception of Random Sequences", *Cognitive Psychology* 17 (3), pp. 295–314.

[16] He, K. (2022): "Mislearning from Censored Data: The Gambler's Fallacy and Other Correlational Mistakes in Optimal-Stopping Problems", *Theoretical Economics* 17(3), pp. 1269–1312.

[17] Jin, L. and C. Peng (2022): "Trading under the Law of Small Numbers", mimeo.

[18] Kahneman, D. and A. Tversky (1972): "Subjective Probability: A Judgement of Representativeness", *Cognitive Psychology* 3(3), pp.430-454.

[19] Noor, J. (2022): "Intuitive Priors", mimeo.

[20] Oppenheimer, D. and B. Monin (2009): "The Retrospective Gambler's Fallacy: Unlikely Events, Constructing the Past, and Multiple Universes", *Judgement and Decision-Making* 4(5), pp. 326-334.

[21] Rabin, M. (2002), "Inference by Believers in the Law of Small Numbers", *Quarterly Journal of Economics* 117, pp. 775–816.

[22] Rapoport, A. and D. Budescu (1997): "Randomization in Individual Choice Behavior", *Psychological Review* 104(3), pp. 603–617.

[23] Rabin, M. and D. Vayanos (2010): "The Gambler's and Hot-Hand Fallacies: Theory and Applications", *Review of Economic Studies* 77, pp. 730–778.

[24] Sandroni, A. (2000): "Do Markets Favor Agents able to Make More Accurate Predictions?" *Econometrica* 8(6), pp. 1303-1341.

[25] Sedlmeier, P. and G. Gigerenzer (1997): "Intuitions about Sample Size: The Empirical Law of Large Numbers", *Journal of Behavioral Decision Making* 10(1), pp. 33-51.

[26] Suetens, S., C. Galbo-Jorgensen and J. Tyran (2016): "Predicting Lotto Numbers: a Natural Experiment on the Gambler's Fallacy and the Hot-Hand Fallacy", *Journal of the European Economic Association* 14, pp. 584-607.

[27] Terrell, D. (1994): "A Test of the Gambler's Fallacy: Evidence from Pari-Mutuel Games", *Journal of Risk and Uncertainty* 8, pp. 309-317.

[28] Tversky, A. and D. Kahneman (1971), "Belief in the Law of Small Numbers", *Psychological Bulletin* 76(2), pp. 105–110.

[29] Tversky, A. and D. Kahneman (1974), "Judgement under Uncertainty: Heuristics and biases", *Science* 185, pp. 1124–1131.

[30] Wagenaar, W. (1970): "Subjective Randomness and the Capacity to Generate Information" in A. F. Sanders (Ed.) *Attention and performance III, Acta Psychologica* 33, pp. 233-242.