

# Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers

Joshua B. Miller and Adam Sanjurjo <sup>\*†‡</sup>

October 17, 2018

## Abstract

We prove that a subtle but substantial bias exists in a common measure of the conditional dependence of present outcomes on streaks of past outcomes in sequential data. The magnitude of this *streak selection bias* generally decreases as the sequence gets longer, but increases in streak length, and remains substantial for a range of sequence lengths often used in empirical work. We observe that the canonical study in the influential hot hand fallacy literature, along with replications, are vulnerable to the bias. Upon correcting for the bias we find that the long-standing conclusions of the canonical study are reversed.

**JEL Classification Numbers:** C12; C14; C18;C19; C91; D03; G02.

**Keywords:** Law of Small Numbers; Alternation Bias; Negative Recency Bias; Gambler’s Fallacy; Hot Hand Fallacy; Hot Hand Effect; Sequential Decision Making; Sequential Data; Selection Bias; Finite Sample Bias; Small Sample Bias.

---

\*Universidad de Alicante. Financial support from the Department of Decision Sciences at Bocconi University, and the Spanish Ministries of Education and Science and Economics and Competitiveness (ECO2015-65820-P) and Generalitat Valenciana (Research Projects Grupo3/086 and PROMETEO/2013/037) is gratefully acknowledged.

†Both authors contributed equally, with names listed in alphabetical order.

‡This draft has benefitted from helpful comments and suggestions from the editor and anonymous reviewers, as well as Jason Abaluck, Jose Apesteguia, David Arathorn, Jeremy Arkes, Maya Bar-Hillel, Phil Birnbaum, Daniel Benjamin, Marco Bonetti, Colin Camerer, Juan Carrillo, Gary Charness, Ben Cohen, Vincent Crawford, Martin Dufwenberg, Jordan Ellenberg, Florian Ederer, Jonah Gabry, Andrew Gelman, Ben Gillen, Tom Gilovich, Maria Glymour, Uri Gneezy, Daniel Goldstein, Daniel Houser, Richard Jagacinski, Daniel Kahan, Daniel Kahneman, Erik Kimbrough, Dan Levin, Elliot Ludvig, Mark Machina, Daniel Martin, Filippo Massari, Guy Molyneux, Gidi Nave, Muriel Niederle, Christopher Olivola, Andreas Ortmann, Ryan Oprea, Carlos Oyarzun, Judea Pearl, David Rahman, Justin Rao, Alan Reifman, Pedro Rey-Biel, Yosef Rinott, Aldo Rustichini, Ricardo Serrano-Padial, Bill Sandholm, Vernon Smith, Lones Smith, Connan Snider, Joel Sobel, Charlie Sprenger, Daniel Stone, Sigrid Suetens, Dmitry Taubinsky, Richard Thaler, Michael J. Wiener, Nat Wilcox, and Bart Wilson. We would also like to thank seminar participants at Caltech, City U. London, Chapman U., Claremont Graduate School, Columbia U., Drexel U., George Mason U., New York University, NHH Norway, Max Planck Institute for Human Development (ABC), Microsoft Research, U. of Minnesota, Naval Postgraduate School, the Ohio State U., Santa Clara U., Stanford U., Tilburg U., U. de Alicante, U. del País Vasco, U. of Amsterdam, UC Berkeley, UC Irvine, U. of Pittsburg, UC Santa Cruz, UC San Diego, U. New South Wales, U. Southern California, U. of Queensland, U. of Wellington, U. of Wisconsin, U. of Zurich, Washington State U., WZB Social Science Center, as well as conference participants at Gary’s Conference, IMEBE Rome 2016, M-BEES Maastricht 2015, SITE Stanford U. 2016, 11th World Congress of The Econometric Society, The 30th Annual Congress of the European Economic Association, and the 14th TIBER Symposium on Psychology and Economics. All mistakes and omissions remain our own.

# 1 Introduction

Jack the researcher takes a coin from his pocket and decides to flip it, say, one hundred times. As he is curious about what outcome typically follows a heads, whenever he flips a heads he commits to writing down the outcome of the next flip on the scrap of paper next to him. Upon completing the one hundred flips, Jack of course expects the proportion of heads written on the scrap of paper to be one-half. Shockingly, Jack is wrong. For a fair coin, the expected proportion of heads is smaller than one-half.

We prove that for any finite sequence of binary data, in which each outcome of “success” or “failure” is determined by an i.i.d. random variable, the proportion of successes among the outcomes that immediately follow a streak of consecutive successes is expected to be strictly less than the underlying (conditional) probability of success.<sup>1</sup> While the magnitude of this *streak selection bias* generally decreases as the sequence gets longer, it increases in streak length, and remains substantial for a range of sequence lengths often used in empirical work.

We observe that the canonical study in the influential *hot hand fallacy* literature,<sup>2</sup> Gilovich, Vallone, and Tversky (1985), along with replications, have mistakenly employed a biased selection procedure that is analogous to Jack’s.<sup>3</sup> Upon conducting a de-biased analysis we find that the long-standing conclusions of the canonical study are reversed.

To illustrate how the selection procedure that Jack uses in the opening example leads to a bias, consider the simplest case in which he decides to flip the coin three times, rather than 100. In this case there are only  $2^3 = 8$  possibilities for the *single* three-flip sequence that Jack will observe. Column one of Table 1 lists these, with the respective flips that Jack would record (write down) underlined for each possible sequence. Column two gives the respective proportion of heads on recorded flips for each possible sequence. As Jack is equally likely to encounter each sequence, one can see that the expected proportion is strictly less than  $1/2$ , and in this case is  $5/12$ .<sup>4</sup> Notice that because the sequence (rather than the flip) is the primitive outcome, the weight that the (conditional) expectation places on each sequence’s associated proportion is independent of the number of recorded flips.<sup>5</sup>

In Section 2 we prove the existence of the streak selection bias for the general case, then quantify it with a formula that we provide. In the case of streaks of length  $k = 1$  (as in the examples discussed

---

<sup>1</sup>The expectation is conditional on the appearance of at least one streak of  $k$  consecutive heads within the first  $n - 1$  trials, where  $n \geq 3$  and  $1 \leq k < n - 1$ .

<sup>2</sup>The hot hand fallacy has been given considerable weight as a candidate explanation for various puzzles and behavioral anomalies identified in the domains of financial markets, sports wagering, casino gambling, and lotteries (Arkes 2011; Avery and Chevalier 1999; Barberis and Thaler 2003; Brown and Sauer 1993; Camerer 1989; Croson and Sundali 2005; De Bondt 1993; De Long, Shleifer, Summers, and Waldmann 1991; Durham, Hertz, and Martin 2005; Galbo-Jørgensen, Suetens, and Tyran 2016; Guryan and Kearney 2008; Kahneman and Riepe 1998; Lee and Smith 2002; Loh and Warachka 2012; Malkiel 2011; Narayanan and Manchanda 2012; Paul and Weinbach 2005; Rabin and Vayanos 2010; Sinkey and Logan 2013; Smith, Levere, and Kurtzman 2009; Sundali and Croson 2006; Xu and Harvey 2014;

**Table 1:** Column one lists the eight sequences that are possible for three flips of a fair coin. The proportion of heads on the flips that immediately follow one or more heads is reported in Column two, for each sequence that has at least one such flip. The (conditional) expectation of the proportion, which is simply its arithmetic average across the six equally likely sequences for which it is defined, is reported in the bottom row.

Three flip sequence	Proportion of Hs on recorded flips
TTT	-
TTH	-
THT	0
HTT	0
THH	1
HTH	0
HHT	$\frac{1}{2}$
HHH	1
Expectation:	$\frac{5}{12}$

above) the formula admits a simple representation, and the bias is tightly related to a form of finite sample bias that shows up in autoregressive coefficient estimators (Shaman and Stine 1988; Yule 1926).<sup>6</sup> By contrast, for the more general case of  $k > 1$  the streak selection bias is typically of larger magnitude, and the formula does not appear to admit a simple representation.<sup>7</sup> In this case we provide a formula for the bias that is numerically tractable for sequence lengths commonly used in the literature that we discuss.

The bias has important implications for the analysis of streak effects in the hot hand fallacy literature. The fallacy refers to the conclusion of the seminal work of Gilovich, Vallone, and Tversky (1985; henceforth GVT), in which the authors found that despite the near ubiquitous belief among basketball fans and experts that there is momentum in shooting performance (“hot

---

Yuan, Sun, and Siu 2014).

<sup>3</sup>For an extensive survey of the hot hand fallacy literature see (Miller and Sanjurjo 2017a).

<sup>4</sup>The expectation is conditional on Jack having at least one flip to record.

<sup>5</sup>By contrast, if Jack were instead to observe *multiple* sequences generated from the same coin, then he could weight each proportion according to its number of recorded flips when taking the average proportion across sequences. This would result in a relatively smaller bias that vanishes in the limit (see Appendix A.2).

<sup>6</sup>In the context of time series regression this bias is known as the *Hurwicz bias* (Hurwicz 1950), which is exacerbated when one introduces fixed effects into a time series model with few time periods (Nerlove 1967, 1971; Nickell 1981). In Web Appendix F.1, we use a sampling-without-replacement argument to show that in the case of  $k = 1$  the streak selection bias, along with finite sample bias for autocorrelation (and time series), are essentially equivalent to: (i) a form of selection bias known in the statistics literature as Berkson’s bias, or Berkson’s paradox (Berkson 1946; Roberts, Spitzer, Delmore, and Sackett 1978), and (ii) several classic conditional probability puzzles.

<sup>7</sup>In Web Appendix D we show that the bias can be decomposed into two factors: a form of sampling-without-replacement, and a stronger bias driven by the overlapping nature of the selection procedure. In Web Appendix F.2 we show how the bias due to the overlapping nature of the selection procedure is related to the *overlapping words paradox* (Guibas and Odlyzko 1981).

hand” or “streak” shooting) the conclusion from their statistical analyses was that momentum did not exist.<sup>8</sup> The result has long been considered a surprising and stark exhibit of irrational behavior, as professional players and coaches have consistently rejected the conclusion, and its implications for their decision making. Indeed, in the years since the seminal paper was published a consensus has emerged that the hot hand is a “myth,” and the associated belief a “massive and widespread cognitive illusion” (Kahneman 2011; Thaler and Sunstein 2008).

We find that GVT’s critical test of hot hand shooting is vulnerable to the bias for the following simple reason: just as it is (surprisingly) incorrect to expect a fair coin flipped 100 times to yield heads half of the time on those flips that immediately follow three consecutive heads, it is incorrect to expect a consistent 50 percent (Bernoulli i.i.d.) shooter who has taken 100 shots to make half of the shots that immediately follow a streak of three hits. Thus, after first replicating the original results using GVT’s: (i) raw data, (ii) biased measures, and (iii) statistical tests, we perform a bias correction to GVT’s measures, then repeat their statistical tests. We also run some additional (unbiased) tests as robustness checks. In contrast with GVT’s results, the bias-corrected re-analysis reveals significant evidence of streak shooting, with large effect sizes.

In a brief discussion of the related literature in Section 3, we first observe that the two replications of GVT (Avugos, Bar-Eli, Ritov, and Sher 2013a; Koehler and Conley 2003) are similarly vulnerable to the bias. We illustrate how the results of Avugos et al. (2013a), a close replication of GVT, similarly reverse when the bias is corrected for. Miller and Sanjurjo (2015b) show that the results of Koehler and Conley (2003), which has been referred to as “an ideal situation in which to study the hot hand” (Thaler and Sunstein 2008), reverse when an unbiased (and more powered) analysis is performed. These results in turn agree with the unbiased analyses performed on all remaining extant controlled shooting datasets in Miller and Sanjurjo (2014). Conservative estimates of hot hand effect sizes are consistently moderate to large across studies.

It follows from these results that the hot hand is not a myth, and that the associated belief is not a cognitive illusion. In addition, because researchers have: (i) accepted the null hypothesis that players have a fixed probability of success, and (ii) treated the *mere* belief in the hot hand as a cognitive illusion, the hot hand fallacy itself can be viewed as a fallacy.<sup>9</sup>

Finally, because the bias is subtle and (initially) surprising, even for people well-versed in probability and statistics, those unaware of it may be susceptible to being misled, or exploited.<sup>10</sup>

---

<sup>8</sup>In particular, they observed that basketball shooting is “analogous to coin tossing” and “adequately described by a simple binomial model.” From this, they concluded that the belief in the hot hand was both “erroneous” and “a powerful and widely shared cognitive illusion” (Gilovich et al. 1985, pp.312–313)

<sup>9</sup>While our evidence reveals that belief in the hot hand is not a fallacy, it remains possible that those who believe in the hot hand hold beliefs that are too strong (or too weak), or cannot accurately detect the hot hand when it occurs. In Section 3.5 we briefly discuss existing evidence on beliefs.

<sup>10</sup>In informal conversations with researchers, and surveys of students, we have found a near-universal belief that the sample proportion should be equal to the underlying probability, in expectation. The conviction with which these beliefs are often held is notable, and reminiscent of the arguments that surrounded the classic Monty Hall problem

On the most basic level, it is possible that a naïve observer could be convinced that negative sequential dependence exists in an i.i.d. random process if sample size information (i.e. the number of flips that Jack records) is obscured.<sup>11</sup> More subtly, the bias can be leveraged to manipulate people into believing that the outcomes of an unpredictable process can be predicted at rates better than chance.<sup>12</sup> Lastly, the bias can be applied in a straightforward way to construct gambling games that appear actuarially fair, but are not.<sup>13</sup>

## 2 The Streak Selection Bias

Let  $\mathbf{X} = \{X_i\}_{i=1}^n$  be a sequence of binary random variables, with  $X_i = 1$  a “success” and  $X_i = 0$  a “failure.” A natural procedure for estimating the probability of success on trial  $t$ , conditional on trial  $t$  immediately following  $k$  consecutive successes, is to first select the subset of trials that immediately follow  $k$  consecutive successes  $I_k(\mathbf{X}) := \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \dots, n\}$ , then calculate the proportion of successes on these trials.<sup>14</sup> The following theorem establishes that when  $\{X_i\}_{i=1}^n$  is a sequence of i.i.d random variables, with probability of success  $p$  and fixed length  $n$ , this procedure yields a biased estimator of the conditional probability,  $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) \equiv p$ .

**Theorem 1** *Let  $\mathbf{X} = \{X_i\}_{i=1}^n$ ,  $n \geq 3$ , be a sequence of independent Bernoulli trials, each with probability of success  $0 < p < 1$ . Let  $\hat{P}_k(\mathbf{X})$  be the proportion of successes on the subset of trials  $I_k(\mathbf{X})$  that immediately follow  $k$  consecutive successes, i.e.  $\hat{P}_k(\mathbf{X}) := \sum_{i \in I_k(\mathbf{X})} X_i / |I_k(\mathbf{X})|$ .  $\hat{P}_k$  is a biased estimator of  $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) \equiv p$  for all  $k$  such that  $1 \leq k \leq n - 2$ . In particular,*

$$E \left[ \hat{P}_k(\mathbf{X}) \mid I_k(\mathbf{X}) \neq \emptyset \right] < p \tag{1}$$

**Outline of Proof:** In the proof contained in Appendix A we begin by showing that the conditional expectation  $E[\hat{P}_k(\mathbf{X}) \mid I_k(\mathbf{X}) \neq \emptyset]$  is equal to the conditional probability  $\mathbb{P}(X_\tau = 1 \mid I_k(\mathbf{X}) \neq \emptyset)$ ,

(Friedman 1998; Nalebuff 1987; Selvin 1975; Vos Savant 1990). See Miller and Sanjurjo (2015a) for more details on the connection between the selection bias, the Monty Hall problem, and other conditional probability puzzles.

<sup>11</sup>In particular, Miller and Sanjurjo (2016) show that the bias introduced here, in conjunction with a quasi-Bayesian model of decision making under sample size neglect (Benjamin, Rabin, and Raymond 2014; Griffin and Tversky 1992; Kahneman and Tversky 1972), provides a novel structural candidate explanation for the persistence of gambler’s fallacy beliefs.

<sup>12</sup>For example, suppose that a predictor observes successive realizations from a binary (or binarized) i.i.d. random process (e.g. daily stock price movements), and is evaluated according to the *success rate* of her predictions over, say, three months. If the predictor is given the freedom of *when* to predict, then she can exceed chance in her expected success rate simply by predicting a reversal whenever there is a streak of consecutive outcomes of the same kind.

<sup>13</sup>A simple example is to sell the following lottery ticket for \$5. A fair coin will be flipped 4 times. For each flip the outcome will be recorded if and only if the previous flip is a heads. If the proportion of recorded heads is strictly greater than one-half then the ticket pays \$10; if the proportion is strictly less than one-half then the ticket pays \$0; if the proportion is exactly equal to one-half, or if no flip is immediately preceded by a heads, then a new sequence of 4 flips is generated. While, intuitively, it seems that the expected value of the lottery must be \$5, it is instead \$4.

<sup>14</sup>In fact, this procedure yields the maximum likelihood estimate for  $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1)$ .

where  $\tau$  is a trial drawn (uniformly) at random from the set of selected trials  $I_k(\mathbf{X})$ . Next, we show that for all eligible trials  $t \in I_k(\mathbf{X})$  we have that  $\mathbb{P}(X_t = 1 | \tau = t, I_k(\mathbf{X}) \neq \emptyset) \leq p$ , with the inequality strict for  $t < n$ , which implies that  $\mathbb{P}(X_\tau = 1 | I_k(\mathbf{X}) \neq \emptyset) < p$ . The strict inequality for  $t < n$  follows from an application of Bayes' rule. In particular, we observe that  $\mathbb{P}(X_t = 1 | \tau = t, I_k(\mathbf{X}) \neq \emptyset) = \mathbb{P}(X_t = 1 | \tau = t, \prod_{i=t-k}^{t-1} X_i = 1) \propto \mathbb{P}(\tau = t | X_t = 1, \prod_{i=t-k}^{t-1} X_i = 1) \times \mathbb{P}(X_t = 1 | \prod_{i=t-k}^{t-1} X_i = 1) = \mathbb{P}(\tau = t | X_t = 1, \prod_{i=t-k}^{t-1} X_i = 1) \times p$ , and then argue that  $\mathbb{P}(\tau = t | X_t = 1, \prod_{i=t-k}^{t-1} X_i = 1) < \mathbb{P}(\tau = t | X_t = 0, \prod_{i=t-k}^{t-1} X_i = 1)$  for  $t < n$ , which guarantees that the likelihood ratio (updating factor) is less than one, and yields  $\mathbb{P}(X_t = 1 | \tau = t, \prod_{i=t-k}^{t-1} X_i = 1) < p$  for  $t < n$ . The intuition for why  $\tau = t$  is more likely when  $X_t = 0$  is the following: because the streak of ones ( $\prod_{i=t-k}^{t-1} X_i = 1$ ) is interrupted by  $X_t = 0$ , the next  $k$  trials are necessarily excluded from the set  $I_k(\mathbf{X})$ . This means that when  $X_t = 0$  there are, on average, fewer eligible trials in  $I_k(\mathbf{X})$  from which to draw (relative to when  $X_t = 1$ ), which implies that any single trial is more likely to be drawn.

In Web Appendix D we show that the downward bias can be decomposed into two factors: (i) *sampling-without-replacement*: the restriction that the finite number of available successes places on the procedure for selecting trials into  $I_k(\mathbf{X})$ , and (ii) *streak overlap*: the additional, and stronger, restriction that the arrangement of successes and failures in the sequence places on the procedure for selecting trials into  $I_k(\mathbf{X})$ .

Though  $\hat{P}_k(\mathbf{X})$  is biased, it is straightforward to show that it is a consistent estimator of  $\mathbb{P}(X_t = 1 | \prod_{j=t-k}^{t-1} X_j = 1)$ .<sup>15,16</sup>

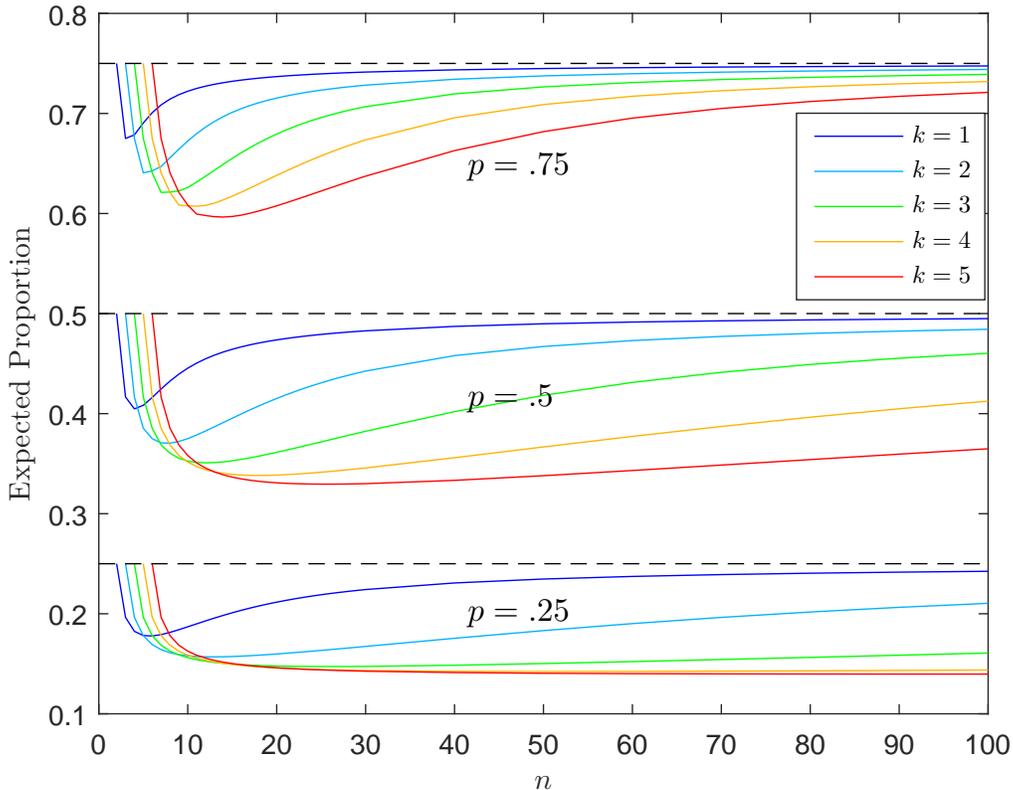
## 2.1 Quantifying the bias.

In Web Appendix E.2 we provide a formula that can be used to calculate  $E[\hat{P}_k(\mathbf{X}) | I_k(\mathbf{X}) \neq \emptyset]$ . For the special case of  $k = 1$  a closed form exists, which we provide in Appendix A.3. There does not appear to be a simple representation for  $k > 1$ .

Figure 1 contains a plot of  $E[\hat{P}_k(\mathbf{X}) | I_k(\mathbf{X}) \neq \emptyset]$ , as a function of the number of trials in the sequence  $n$ , and for different values of  $k$  and  $p$ . The dotted lines in the figure represent the true probability of success for  $p = 0.25, 0.50$ , and  $0.75$ , respectively. The five solid lines immediately below each dotted line represent the respective expected proportions for each value of  $k = 1, 2, \dots, 5$ . Observe that while the bias does generally decrease as  $n$  increases, it can remain substantial even

<sup>15</sup>See Appendix A.2 for a proof.

<sup>16</sup>It is possible to devise alternative estimators of the conditional probability that are unbiased. To illustrate, if the researcher were instead to control the number of selected trials by repeating the experiment until he generates exactly  $m$  trials that immediately follow  $k$  consecutive successes, then the proportion would be unbiased. Alternatively, if the researcher were to eliminate the overlapping nature of the measure, there would be no bias, even though the number of selected trials would still be random. In particular, for a sequence of  $n$  trials, one can take each run of successes, and if it is of even length  $2\ell$ , divide it into blocks of two trials; if it is of odd length  $2\ell - 1$  include the right adjacent tails and divide it into blocks of two trials. In each case, the run of successes contributes  $\ell$  observations.



**Figure 1:** *The expected value of the proportion of successes on trials that immediately follow  $k$  consecutive successes,  $\hat{P}_k(\mathbf{X})$ , as a function of the total number of trials  $n$ , for different values of  $k$  and probabilities of success  $p$ , using the formula provided in Web Appendix E.*

for long sequences. For example, in the case of  $n = 100$ ,  $p = 0.5$ , and  $k = 5$ , the magnitude of the bias is  $.35 - .50 = -0.15$ , and in the case of  $n = 100$ ,  $p = 0.25$ , and  $k = 3$ , the magnitude of the bias is  $.16 - .25 = -0.09$ .<sup>17</sup>

### 3 Application to the Hot Hand Fallacy

*This account explains both the formation and maintenance of the erroneous belief in the hot hand: if random sequences are perceived as streak shooting, then no amount of exposure to such sequences will convince the player, the coach, or the fan that the sequences are in fact random. (Gilovich, Vallone, and Tversky [GVT] 1985)*

<sup>17</sup>The non-monotonicity in  $n$  of the curves presented in Figure 1 arises because for any streak length  $k$  there is no bias when  $n = k + 1$  (because there are only two feasible sequences, which are equally likely), or in the limit (see Appendix A.2).

In their seminal paper GVT find no evidence of hot hand shooting in their analysis of basketball shooting data, despite the near-unanimous belief in the hot hand among players, coaches, and fans. As a result, they conclude that belief in the hot hand is a “powerful and widely shared cognitive illusion.” (p. 313).

### 3.1 GVT’s analysis

#### *Empirical approach*

GVT’s “Analysis of Conditional Probabilities” is their main test of hot hand shooting, and provides their only measure of the magnitude of the hot hand effect. The goal of their analysis is to determine whether a player’s hit probability is higher following a streak of hits than it is following a streak of misses.<sup>18</sup> To this end, GVT reported each player  $i$ ’s shooting percentage conditional on having: (1) hit the last  $k$  shots,  $\hat{P}^i(\text{hit}|k \text{ hits})$ , and (2) missed the last  $k$  shots,  $\hat{P}^i(\text{hit}|k \text{ misses})$ , for streak lengths  $k = 1, 2, 3$  (Table 4, p. 307).<sup>19</sup> After informally comparing these shooting percentages for individual players, GVT performed a paired t-test of whether  $E[\hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})] = 0$ , for  $k = 1, 2, 3$ .<sup>20,21</sup>

In the remainder of this section, we focus our discussion on streaks of length three (or more), as in, e.g. Koehler and Conley (2003); Rao (2009b), given that: (i) shorter streak lengths exacerbate attenuation bias due to measurement error (see Footnote 21 and Appendix B), and (ii) people typically perceive streaks as beginning with the third successive event (Carlson and Shu 2007). In any case, robustness checks using different streak lengths yield similar results (see Footnotes 29 and 33 in Section 3.3).

<sup>18</sup>GVT explicitly treat hot hand and streak shooting as synonymous (Gilovich et al. 1985, pp. 296-297). Miller and Sanjurjo (2014) provide an analysis that distinguishes between hot hand and cold hand shooting, and find hot hand shooting across all extant controlled shooting datasets, but little in the way of cold hand shooting. Thus, in the present analysis we use the terms streakiness and hot hand shooting interchangeably.

<sup>19</sup>We abuse our notation from Section 2 here in order to facilitate comparison with GVT’s analysis: we use  $\hat{P}^i(\text{hit}|k \text{ hits})$  for both the random variable  $\hat{P}_k(\mathbf{X})$  and its realization  $\hat{P}_k(\mathbf{x})$ . Similarly, we use  $\hat{P}^i(\text{hit}|k \text{ misses})$  for the proportion of successes on trials that immediately follow  $k$  consecutive failures.

<sup>20</sup>Under the null hypothesis the difference between each  $i$ ’s pair of shooting percentages is drawn from a normal distribution with mean zero.

<sup>21</sup>While GVT’s analysis of conditional probabilities provides their only measure of the magnitude of the hot hand, they also analyze the number of runs, serial correlation, and variation of shooting percentage in 4-shot windows. Miller and Sanjurjo (2014) show that the runs and serial correlation tests, along with the conditional probability test for  $k = 1$ , all amount to roughly the same test, and moreover, that they are not sufficiently powered to identify hot hand shooting. The reason why is due to measurement error: the act of hitting a single shot is only a weak signal of a change in a player’s underlying probability of success, which leads to an attenuation bias in the estimate of the increase in the probability of success associated with entering the hot state (see Appendix B and Stone (2012)’s work on measurement error when estimating autocorrelation in ability). The test of variation in 4-shot windows is even less powered than the aforementioned tests (Miller and Sanjurjo 2014; Wardrop 1999).

## Data

GVT analyze shot sequences from basketball players in three contexts: NBA field goal data, NBA free-throw data, and a controlled shooting experiment with NCAA collegiate players. The shooting experiment was GVT’s controlled test of hot hand shooting, designed for the purpose of “eliminating the effects of shot selection and defensive pressure” (p. 34), which makes it central to their main conclusions. Thus, we focus on this data below when discussing the relevance of the bias to GVT’s results.<sup>22,23</sup>

In GVT’s controlled shooting experiment 26 players from the Cornell University Mens’ (14) and Womens’ (12) basketball teams participated in an incentivized shooting task. Each player shot 100 times at a distance from which the experimenters determined he/she would make around 50 percent of the shots. Following each shot the player had to change positions along two symmetric arcs—one facing the basket from the left, and the other from the right.

## Results

In Columns 4 and 5 of Table 2 we use the raw data from GVT to reproduce the shooting percentages,  $\hat{P}^i(\text{hit}|3 \text{ hits})$  and  $\hat{P}^i(\text{hit}|3 \text{ misses})$ , for each of the 26 players (these are identical to Columns 2 and 8 of Table 4 in GVT). As indicated in GVT, players on average hit .49 when on a hit streak, versus .45 when on a miss streak. GVT’s paired t-test finds the difference to be statistically indistinguishable from zero, and we replicate this result ( $p = .49$ ).

### 3.2 The bias in GVT’s analysis

While GVT’s null hypothesis that  $E[\hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})] = 0$  seems intuitively correct for a consistent shooter with a fixed probability of success  $p^i$  (i.i.d. Bernoulli), Theorem 1 reveals a

<sup>22</sup>From the statistical point of view, the in-game field goal data that GVT analyze (Study 2: 76ers, 1980-81 season: 9 players, 48 home games) is not ideal for the study of hot hand shooting for reasons that are unrelated to the bias. The most notable concern with in-game field goal data is that the opposing team has incentive to make *costly* strategic adjustments to mitigate the impact of the “hot” player (Dixit and Nalebuff 1991, p. 17). This concern has been emphasized by researchers in the hot hand literature (Aharoni and Sarig 2011; Green and Zwiebel 2017), and is not merely theoretical, as it has a strong empirical basis. While GVT observed that a shooter’s field goal percentage is lower after consecutive successes, subsequent studies have shown that with even partial controls for defensive pressure (and shot location), this effect is eliminated (Bocskocsky, Ezekowitz, and Stein 2014; Rao 2009a). Further, evidence of specific forms of strategic adjustment has been documented (Aharoni and Sarig 2011; Bocskocsky et al. 2014). See Miller and Sanjurjo (2014) for further details.

<sup>23</sup>The in-game free throw data that GVT analyze (Study 3: Celtics, 1980-81, 1981-82 seasons: 9 players), while arguably controlled, is not ideal for the study of hot hand shooting for a number of reasons: (i) hitting the first shot in a pair of isolated shots is not typically regarded by fans and players as hot hand shooting (Koehler and Conley 2003), presumably due to the high prior probability of success ( $\approx .75$ ), (ii) hitting a single shot is a weak signal of a player’s underlying state, which can lead to severe measurement error (Arkes 2013; Stone 2012), (iii) it is vulnerable to an omitted variable bias, as free throw pairs are relatively rare, and shots must be aggregated across games and seasons in order to have sufficient sample size (Miller and Sanjurjo 2014). In any event, subsequent studies of free throw data have found evidence that is inconsistent with the conclusions that GVT drew from the Celtics’ data (Aharoni and Sarig 2011; Arkes 2010; Goldman and Rao 2012; Miller and Sanjurjo 2014; Wardrop 1995; Yaari and Eisenmann 2011).

**Table 2:** Columns 4 and 5 reproduce the shooting percentages and number of shots that appear in Table 4, Columns 2 and 8, from Gilovich et al. (1985) (note: 3 hits (misses) includes streaks of 3, 4, 5, etc.). Column 6 reports the difference between the proportions (using the raw data), and column 7 adjusts for the bias (mean correction), based on each player’s shooting percentage (probability in this case) and number of shots.

Player	# shots	$\hat{P}(\text{hit})$	$\hat{P}(\text{hit} 3 \text{ hits})$	$\hat{P}(\text{hit} 3 \text{ misses})$	$\hat{D}_3 := \hat{P}(\text{hit} 3 \text{ hits}) - \hat{P}(\text{hit} 3 \text{ misses})$		
					GVT est.	bias adj.	
Males							
1	100	.54	.50 (12)	.44 (9)	.06	.14	
2	100	.35	.00 (3)	.43 (28)	-.43	-.33	
3	100	.60	.60 (25)	.67 (6)	-.07	.02	
4	90	.40	.33 (3)	.47 (15)	-.13	-.03	
5	100	.42	.33 (6)	.75 (12)	-.42	-.33	
6	100	.57	.65 (23)	.25 (12)	.40	.48	
7	75	.56	.65 (17)	.29 (7)	.36	.47	
8	50	.50	.57 (7)	.50 (6)	.07	.24	
9	100	.54	.83 (30)	.35 (20)	.48	.56	
10	100	.60	.57 (21)	.57 (7)	.00	.09	
11	100	.58	.62 (21)	.57 (7)	.05	.14	
12	100	.44	.43 (7)	.41 (17)	.02	.10	
13	100	.61	.50 (18)	.40 (5)	.10	.19	
14	100	.59	.60 (20)	.50 (6)	.10	.19	
Females							
1	100	.48	.33 (9)	.67 (9)	-.33	-.25	
2	100	.34	.40 (5)	.43 (28)	-.03	.07	
3	100	.39	.50 (8)	.36 (25)	.14	.23	
4	100	.32	.33 (3)	.27 (30)	.07	.17	
5	100	.36	.20 (5)	.22 (27)	-.02	.08	
6	100	.46	.29 (7)	.54 (11)	-.26	-.18	
7	100	.41	.62 (13)	.32 (25)	.30	.39	
8	100	.53	.73 (15)	.67 (9)	.07	.15	
9	100	.45	.50 (8)	.46 (13)	.04	.12	
10	100	.46	.71 (14)	.32 (19)	.40	.48	
11	100	.53	.39 (13)	.50 (10)	-.12	-.04	
12	100	.25	.- (0)	.32 (37)	.	.	
Average		.47	.49	.45	.03	.13	

flaw in this reasoning. In particular, we have established that  $\hat{P}^i(\textit{hit}|k \textit{ hits})$  is expected to be less than  $p^i$ , and  $\hat{P}^i(\textit{hit}|k \textit{ misses})$  greater than  $p^i$  (by symmetry). In fact, in Appendix A.4 we show that the difference  $\hat{P}^i(\textit{hit}|k \textit{ hits}) - \hat{P}^i(\textit{hit}|k \textit{ misses})$  is not only expected to be negative, but that its magnitude is more than double the bias in either of the respective proportions.<sup>24</sup>

Under GVT’s design target of each player taking  $n = 100$  shots and making half ( $p = .5$ ) of them, we use the results from Section 2 and Appendix A.4 to find that the expected difference (and the strength of the bias) is -8 percentage points.<sup>25</sup> Therefore, the difference between the average proportion of +4 percentage points observed by GVT is actually +12 percentage points higher than the difference that would be expected from a Bernoulli i.i.d. shooter. Thus, the bias has long disguised evidence in GVT’s data that may well indicate hot hand shooting.

### 3.3 A bias-corrected statistical analysis of GVT

A straightforward way to adjust for the bias in GVT’s analysis is simply to shift the difference for each shooter by the amount of the corresponding bias, then repeat their paired t-test. While this test yields a statistically significant result ( $p < .05$ ), the paired t-test limits statistical power because it reduces each player’s performance to a single number, ignoring the number of shots that the player attempted in each category, i.e. “3 hits” and “3 misses.” In addition, adjusting for the bias based on the assumption that  $p = .5$  assumes that GVT’s design target was met precisely.

As a result, for each player we again compute the bias under the null hypothesis that trials are i.i.d. Bernoulli (i.e. “consistent” shooting) but now with a probability of success equal to the player’s observed shooting percentage (Column 3 of Table 2), and using the number of shots taken in each category to inform our standard errors. With this approach the average difference goes from +3 to a considerable +13 percentage points ( $p < .01$ ,  $S.E. = 4.7\text{pp}$ ).<sup>26,27</sup> To put the magnitude of +13 percentage points into perspective, the difference between the median three point shooter and the top three point shooter in the 2015-2016 NBA season was 12 percentage points.<sup>28</sup> Further, this is a *conservative* estimate because in practice the data generating processes (i.e. shooters) clearly

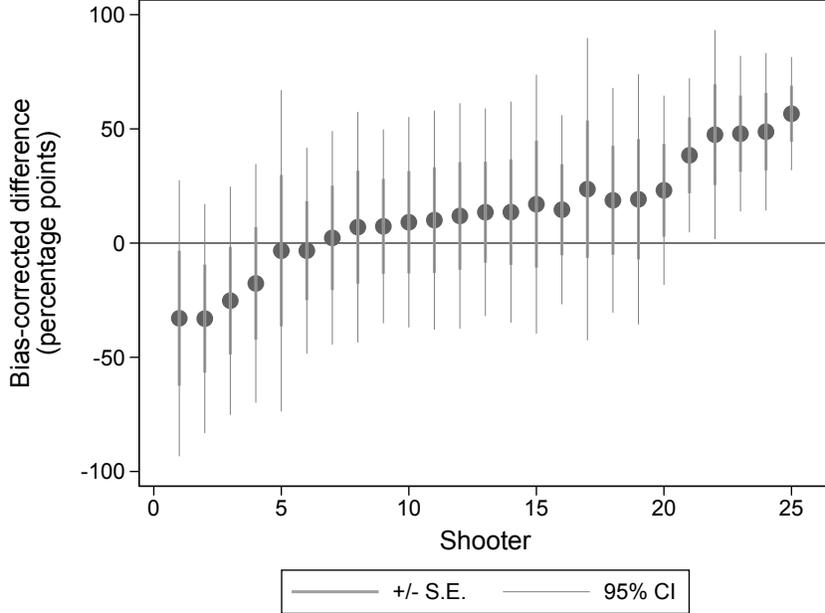
<sup>24</sup>That the difference is expected to be negative does not follow immediately from Theorem 1, as the set of sequences for which the difference is well-defined is a strict subset of the set corresponding to either of the respective proportions. Nevertheless, the reasoning of the proof is similar. See Theorem 3 of Appendix A.4.

<sup>25</sup>See Figure 4 in Appendix A.4 for the bias in the difference as  $n, p$  and  $k$  vary.

<sup>26</sup>The standard error is computed based on the assumption of independence across the 2600 trials, and normality. In particular, defining player  $i$ ’s difference  $\hat{D}_k^i := \hat{P}^i(\textit{hit}|k \textit{ hits}) - \hat{P}^i(\textit{hit}|k \textit{ misses})$ , the variance satisfies  $\widehat{Var}(\hat{D}_k^i) = \widehat{Var}(\hat{P}^i(\textit{hit}|k \textit{ hits})) + \widehat{Var}(\hat{P}^i(\textit{hit}|k \textit{ misses}))$  for each player  $i$ . Simulations reveal that the associated  $(1 - \alpha) \times 100\%$  confidence intervals with radius  $z_{\alpha/2} \times \widehat{Var}(\hat{D}_k^i)^{1/2}$  (where the mean difference is given by  $\bar{D}_k := (1/n) \sum_{i=1}^n \hat{D}_k^i$ ) have the appropriate coverage—i.e.  $(1 - \alpha/2) \times 100\%$  of the time the true difference is greater than  $\bar{D}_k - z_{\alpha/2} \times \widehat{Var}(\hat{D}_k^i)^{1/2}$ , for both Bernoulli trials and the positive feedback model discussed in Section B.

<sup>27</sup>For an alternative approach that involves pooling shots across players, and yields similar results, see Appendix C.

<sup>28</sup>ESPN, “NBA Player 3-Point Shooting Statistics - 2015-16.” [http://www.espn.com/nba/statistics/player/\\_/stat/3-points](http://www.espn.com/nba/statistics/player/_/stat/3-points) [accessed September 24, 2016].

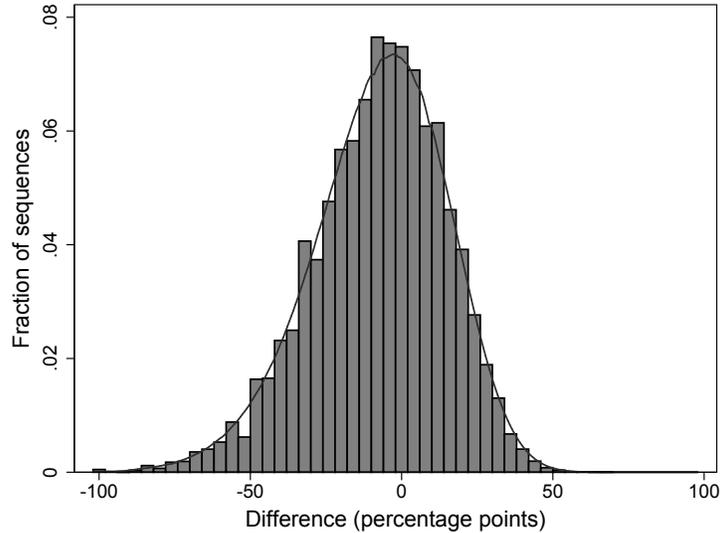


**Figure 2:** The bias-corrected difference  $\hat{D}_3^i = \hat{P}^i(\text{hit}|3 \text{ hits}) - \hat{P}^i(\text{hit}|3 \text{ misses})$  for each player, under the assumption that his/her probability of success is equal to his/her overall shooting percentage.

differ from i.i.d. Bernoulli trials, and the bias becomes much larger under various models of hot hand shooting because of measurement error (see Appendix B).

GVT also informally discussed the heterogeneity across players, and asserted that most players shot relatively better when on a streak of misses than when on a streak of hits. By contrast, Figure 2 shows that once the bias correction is made to the differences 19 of the 25 players directionally exhibit hot hand shooting, which is itself significant ( $p < .01$ , binomial test).<sup>29</sup> Further, as indicated by the confidence intervals, t-tests reveal that 5 of the players exhibit statistically significant evidence of hot hand shooting ( $p < .05$ , t-test), which, for a set of 25 independent tests, is itself significant ( $p < .01$ , binomial test).

<sup>29</sup>Repeating the tests for longer ( $k = 4$ ) or shorter ( $k = 2$ ) streak lengths yields similar results that are consistent with the attenuation bias in estimated effect sizes discussed in Footnote 21. In particular, If we instead define a streak as beginning with 4 consecutive hits, which is a stronger signal of hot hand shooting, then the average bias-adjusted difference in proportions is 10 percentage points ( $p = .07$ ,  $S.E. = 6.9$ , one-sided test), and four players exhibit statistically significant hot hand shooting ( $p < .05$ ), which is itself significant ( $p < .01$ , binomial test). On the other hand, if we define a streak as beginning with 2 consecutive hits, which is a weaker signal of hot hand shooting, then the average bias-adjusted difference in proportions is 5.4 percentage points ( $p < .05$ ,  $S.E. = 3$ , one-sided test), and four players exhibit statistically significant hot hand shooting ( $p < .05$ ), which is itself significant ( $p < .01$ , binomial test).



**Figure 3:** The histogram and kernel density plot of the (exact) discrete probability distribution of  $\hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})$ , for a single player  $i$  with  $n = 100$  and  $n_1 = 50$ , using a variant of the formula for the distribution provided in Web Appendix E.2 (see supplementary materials).<sup>31</sup>

#### Non-parametric robustness test

As a robustness check we perform permutation tests, which are (by construction) invulnerable to the bias. The null hypothesis for a permutation test is that a player is a consistent shooter, i.e. has an i.i.d. fixed (unknown) probability of success. The first step to test for streak shooting in player  $i$  is to observe his/her shot sequence and compute the difference in proportions,  $\hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})$ . The second step is to compute this difference for each unique rearrangement of the observed sequence; each of these *permutations* is equally likely because player  $i$ 's probability of success is fixed under the null hypothesis.<sup>30</sup> The set of unique differences computed in the second step, along with their associated relative frequencies, constitutes the exact sampling distribution of the difference under the null hypothesis (conditional on the observed number of hits). This distribution can then be used for statistical testing (See Appendix C.2 for details). The distribution is negative-skewed, and can be represented by histograms such as the one shown in Figure 3, which provides the *exact* distribution for a player who has hit 50 out of 100 shots.

Results of the permutation tests agree with those of the bias-corrected tests reported above. In particular, the average difference across shooters indicates hot hand shooting with a similar level of

<sup>30</sup>Thus, the permutation procedure directly implements GVT's idea of comparing a "player's performance... to a sequence of hits and misses generated by tossing a coin" (Gilovich et al. 1985, p. 296)

<sup>31</sup>Uses a bin width of 4 percentage points. The values for the difference are grouped based on the first 6 decimal digits of precision. For this precision, the more than  $10^{29}$  distinguishable sequences take on 19,048 distinct values. In the computation of the expected value in Figures 1 and 4, each difference is instead represented with the highest floating point precision available.

significance ( $p < .01$ ).<sup>32</sup> Also as before, 5 individual players exhibit significant hot hand shooting ( $p < .01$ , binomial test).<sup>33</sup>

### 3.4 The hot hand (and bias) in other controlled and semi-controlled studies

A close replication of GVT’s controlled shooting experiment is found in Avugos et al. (2013a), a study that mimics GVT’s design and analysis, but with olympian rather than collegiate players, and fewer shots ( $n = 40$ ) per player. From the authors’ Table 1 (p. 6), one can derive the average  $\hat{p}(\text{hit}|3 \text{ hits})$  and  $\hat{p}(\text{hit}|3 \text{ misses})$  across players, which are roughly .52 and .54, respectively, yielding an average difference in shooting percentages of -2 percentage points.<sup>34</sup> However, Figure 4 in Appendix A.4 shows that the strength of the bias for  $n = 40$  shots and  $p = .5$  (the design target) is -.20. Thus, once the bias is corrected for in this small sample the average difference across shooters becomes roughly +18 percentage points.<sup>35</sup>

Koehler and Conley (2003) test for the hot hand in the NBA three point shooting contest, which has been described as an ideal setting in which to study the hot hand (Thaler and Sunstein 2008). The authors find no evidence of hot hand shooting in their analysis of four years of data. However, as in GVT and Avugos et al. (2013a), the conditional probability tests that the authors conduct are vulnerable to the bias. By contrast, Miller and Sanjurjo (2015b) collect 28 years of data, which yields 33 players that have taken at least 100 shots; using this dataset, we find that the average bias-corrected difference across players is +8 percentage points ( $p < .01$ ).<sup>36</sup> Further, 8 of the 33 players exhibit significant hot hand shooting ( $p < .05$ ), which itself is statistically significant ( $p < .001$ , binomial test).

The only other controlled shooting studies that we are aware of are Jagacinski, Newell, and Isaac (1979) and Miller and Sanjurjo (2014).<sup>37,38</sup> Both studies have few shooters (6 and 8, respectively) but many shots across multiple shooting sessions for each player (540 and 900+ shots, respectively).

<sup>32</sup>The procedure in this pooled test involves stratifying the permutations by player. In particular, we conduct a test of the average of the standardized difference, where for each player the difference is standardized by shifting its mean and scaling its variance under  $H_0$ . In this case  $H_0: \mathbb{P}(\text{success on trial } t \text{ for player } i) = p^i$  for all  $t, i$ .

<sup>33</sup>As in Footnote 29, the results of the permutation test are robust to varying streak length  $k$ .

<sup>34</sup>We could not analyze the raw data because the authors declined to provide it to us. The data that represents a close replication of GVT is from the betting game phase. Using Table 1, we have  $\hat{p}(\text{hit}|3 \text{ hits}) = (.56 + .52)/2$  and  $\hat{p}(\text{hit}|3 \text{ misses}) = (.54 + .49)/2$ , which is the average of the shooting percentage of Group A in Phase 1 with that of Group B from Phase 2.

<sup>35</sup>The authors also had another treatment, in which they had shooters rate, before each shot, from 0-100% on a certainty scale whether they would hit the next shot. If we repeat the analysis on the data from this treatment then the average  $\hat{p}(\text{hit}|3 \text{ hits})$  and  $\hat{p}(\text{hit}|3 \text{ misses})$  across players are roughly .56 and .65, respectively, yielding an average difference of -9 percentage points, and a bias-adjusted difference of +11 percentage points.

<sup>36</sup>Miller and Sanjurjo (2015b) also implemented the unbiased permutation test procedure of Section 3.3.

<sup>37</sup>The one exception is a controlled shooting study that involved a single shooter Wardrop (1999). After personal communication with the shooter, who conducted the study herself (recording her own shots), we viewed it as not having sufficient control to warrant analysis.

<sup>38</sup>We thank Tom Gilovich for bringing the study of Jagacinski et al. to our attention. It had gone uncited in the hot hand literature until Miller and Sanjurjo (2014).

The bias-adjusted average difference in the studies are +7 and +4 percentage points, respectively. In addition, Miller and Sanjurjo (2014) find substantial and persistent evidence of hot hand shooting in individual players.<sup>39</sup>

Thus, once the bias is accounted for, *conservative* estimates of hot hand effect sizes across all extant controlled and semi-controlled shooting studies are consistently moderate to large.<sup>40</sup>

### 3.5 Belief in the Hot Hand

The results of our reanalysis of GVT’s data lead us to a conclusion that is the opposite of theirs: belief in the hot hand is not a cognitive illusion. Nevertheless, it remains possible, perhaps even likely, that professional players and coaches sometimes infer the presence of a hot hand when it does not exist. Similarly, even when in the presence of the hot hand, players may overestimate its influence and respond too strongly to it. By contrast, a hot hand might also go undetected, or be underestimated (Stone and Arkes 2017). These questions are important because understanding the extent to which decision makers’ beliefs and behavior do not correspond to the actual degree of hot hand shooting could have considerable implications for decision-making more generally.

While GVT’s main conclusion was of a binary nature, i.e. based on the question of whether belief in the hot hand is either fallacious or not, they explored hot hand beliefs via a survey of player and coach beliefs, and an incentivized betting task with the Cornell players. In the survey they find that the near universal beliefs in the hot hand do not accord with the lack of hot hand shooting evidence that resulted from their analysis of the shooting data, and in the betting task they found that players were incapable of predicting upcoming shot outcomes successfully, which suggests that even if there were a hot hand, it could not be detected successfully.

However, in light of the results presented in the present paper subjects’ responses in GVT’s un-incentivized survey are actually qualitatively consistent with the evidence presented above.<sup>41</sup> More substantively, GVT’s statistical analysis of betting data has recently been shown to be considerably underpowered, as the authors conduct many separate individual bettor level tests rather than pooling the data across bettors (Miller and Sanjurjo 2017b). In addition, GVT misinterpret their

<sup>39</sup>See Avugos, Köppen, Czienskowski, Raab, and Bar-Eli (2013b) for a meta-analysis of the hot hand, which includes sports besides basketball. Tversky and Gilovich (1989) argue that evidence for the hot hand in other sports is not relevant to their main conclusion because so long as the hot hand does not exist in basketball, then the perception of the hot hand by fans, players and coaches must necessarily be a cognitive illusion (see also Alter and Oppenheimer [2006]).

<sup>40</sup>The magnitudes of all estimated effect sizes are conservative for two reasons: (i) if a player’s probability of success is not driven merely by feedback from previous shots, but also by other time-varying player (and environment) specific factors, then the act of hitting consecutive shots will serve as only a noisy proxy of the hot state, resulting in measurement error, and an attenuation bias in the estimate (see Appendix B), and (ii) if the effect of consecutive successes on subsequent success is heterogenous in magnitude (and sign) across players, then an average measure will underestimate how strong the effect can be in certain players.

<sup>41</sup>See Appendix B of Miller and Sanjurjo (2017b) for details.

measures of bettors' ability to predict. In light of these limitations, Miller and Sanjurjo (2017b) reanalyze GVT's betting data, and find that players on average shoot around +7 percentage points higher when bettors have predicted that the shot will be a hit, rather than a miss ( $p < .001$ ). This increase is comparable in magnitude to an NBA shooter going from slightly above average to elite in three point percentage.<sup>42</sup>

Miller and Sanjurjo (2014) present complementary evidence on beliefs, in which semi-professional players rank their teammates' respective increases in shooting percentage when on a streak of three hits (relative to their base rates) in a shooting experiment that the rankers do not observe. Players' rankings are found to be highly correlated with their teammates' actual increases in shooting percentage in this out-of-sample test, yielding an average correlation of -0.60 ( $p < .0001$ ; where 1 is the rank of the shooter with the perceived largest percentage point increase).

In sum, while it remains possible that professional players' and coaches' hot hand beliefs are poorly calibrated, this claim is not clearly supported by the existing body of evidence.

## 4 Conclusion

We prove the existence of, and quantify, a novel form of selection bias that counter-intuitively arises in some particularly simple analyses of sequential data. A key implication of the bias is that the empirical approach of the canonical hot hand fallacy paper, and its replications, are incorrect. Upon correcting for the bias we find that the data that had previously been interpreted as demonstrating that belief in the hot hand is a fallacy, instead provides substantial evidence that it is not a fallacy to believe in the hot hand.

## References

- AHARONI, G. AND O. H. SARIG (2011): "Hot hands and equilibrium," *Applied Economics*, 44, 2309–2320.
- ALTER, A. L. AND D. M. OPPENHEIMER (2006): "From a fixation on sports to an exploration of mechanism: The past, present, and future of hot hand research," *Thinking & Reasoning*, 12, 431–444.
- ARKES, J. (2010): "Revisiting the Hot Hand Theory with Free Throw Data in a Multivariate Framework," *Journal of Quantitative Analysis in Sports*, 6.
- (2011): "Do Gamblers Correctly Price Momentum in NBA Betting Markets?" *Journal of Prediction Markets*, 5, 31–50.

---

<sup>42</sup>ESPN, "NBA Player 3-Point Shooting Statistics - 2015-16." [http://www.espn.com/nba/statistics/player/\\_/stat/3-points](http://www.espn.com/nba/statistics/player/_/stat/3-points) [accessed September 24, 2016].

- (2013): “Misses in ‘Hot Hand’ Research,” *Journal of Sports Economics*, 14, 401–410.
- AVERY, C. AND J. CHEVALIER (1999): “Identifying Investor Sentiment from Price Paths: The Case of Football Betting,” *Journal of Business*, 72, 493–521.
- AVUGOS, S., M. BAR-ELI, I. RITOV, AND E. SHER (2013a): “The elusive reality of efficacy performance cycles in basketball shooting: analysis of players’ performance under invariant conditions,” *International Journal of Sport and Exercise Psychology*, 11, 184–202.
- AVUGOS, S., J. KÖPPEN, U. CZIENSKOWSKI, M. RAAB, AND M. BAR-ELI (2013b): “The “hot hand” reconsidered: A meta-analytic approach,” *Psychology of Sport and Exercise*, 14, 21–27.
- BAI, D. S. (1975): “Efficient Estimation of Transition Probabilities in a Markov Chain,” *The Annals of Statistics*, 3, 1305–1317.
- BARBERIS, N. AND R. THALER (2003): “A survey of behavioral finance,” *Handbook of the Economics of Finance*, 1, 1053–1128.
- BENJAMIN, D. J., M. RABIN, AND C. RAYMOND (2014): “A Model of Non-Belief in the Law of Large Numbers,” Working Paper.
- BERKSON, J. (1946): “Limitations of the application of fourfold table analysis to hospital data,” *Biometrics Bulletin*, 47–53.
- BOCSKOSKY, A., J. EZEKOWITZ, AND C. STEIN (2014): “The Hot Hand: A New Approach to an Old ‘Fallacy’,” 8th Annual Mit Sloan Sports Analytics Conference.
- BROWN, W. A. AND R. D. SAUER (1993): “Does the Basketball Market Believe in the Hot Hand? Comment,” *American Economic Review*, 83, 1377–1386.
- CAMERER, C. F. (1989): “Does the Basketball Market Believe in the ‘Hot Hand,’?” *American Economic Review*, 79, 1257–1261.
- CARLSON, K. A. AND S. B. SHU (2007): “The rule of three: how the third event signals the emergence of a streak,” *Organizational Behavior and Human Decision Processes*, 104, 113–121.
- CROSON, R. AND J. SUNDALI (2005): “The Gamblers Fallacy and the Hot Hand: Empirical Data from Casinos,” *Journal of Risk and Uncertainty*, 30, 195–209.
- DE BONDT, W. P. (1993): “Betting on trends: Intuitive forecasts of financial risk and return,” *International Journal of Forecasting*, 9, 355–371.
- DE LONG, J. B., A. SHLEIFER, L. H. SUMMERS, AND R. J. WALDMANN (1991): “The Survival of Noise Traders In Financial-markets,” *Journal of Business*, 64, 1–19.
- DIXIT, A. K. AND B. J. NALEBUFF (1991): *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*, W.W. Norton & Company.
- DURHAM, G. R., M. G. HERTZEL, AND J. S. MARTIN (2005): “The Market Impact of Trends and Sequences in Performance: New Evidence,” *Journal of Finance*, 60, 2551–2569.

- FRIEDMAN, D. (1998): “Monty Hall’s Three Doors: Construction and Deconstruction of a Choice Anomaly,” *American Economic Review*, 88, 933–946.
- GALBO-JØRGENSEN, C. B., S. SUETENS, AND J.-R. TYRAN (2016): “Predicting Lotto Numbers A natural experiment on the gamblers fallacy and the hot hand fallacy,” *Journal of the European Economic Association*, 14, working Paper.
- GILOVICH, T., R. VALLONE, AND A. TVERSKY (1985): “The Hot Hand in Basketball: On the Misperception of Random Sequences,” *Cognitive Psychology*, 17, 295–314.
- GOLDMAN, M. AND J. M. RAO (2012): “Effort vs. Concentration: The Asymmetric Impact of Pressure on NBA Performance,” 6th Annual Mit Sloan Sports Analytics Conference.
- GREEN, B. S. AND J. ZWIEBEL (2017): “The Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments?” *Management Science*, working Paper.
- GRIFFIN, D. AND A. TVERSKY (1992): “The weighing of evidence and the determinants of confidence,” *Cognitive Psychology*, 24, 411–435.
- GUIBAS, L. J. AND A. M. ODLYZKO (1981): “String overlaps, pattern matching, and nontransitive games,” *Journal of Combinatorial Theory, Series A*, 30, 183–208.
- GURYAN, J. AND M. S. KEARNEY (2008): “Gambling at Lucky Stores: Empirical Evidence from State Lottery Sales,” *American Economic Review*, 98, 458–473.
- HURWICZ, L. (1950): “Least squares bias in time series,” *Statistical inference in dynamic economic models*, 10, 365–383.
- IMBENS, G. W. AND M. KOLESAR (2016): “Robust Standard Errors in Small Samples: Some Practical Advice,” Working Paper, March.
- JAGACINSKI, R. J., K. M. NEWELL, AND P. D. ISAAC (1979): “Predicting the Success of a Basketball Shot at Various Stages of Execution,” *Journal of Sport Psychology*, 1, 301–310.
- KAHNEMAN, D. (2011): *Thinking, Fast and Slow*, Farrar, Straus and Giroux.
- KAHNEMAN, D. AND M. W. RIEPE (1998): “Aspects of Investor Psychology: Beliefs, preferences, and biases investment advisors should know about,” *Journal of Portfolio Management*, 24, 1–21.
- KAHNEMAN, D. AND A. TVERSKY (1972): “Subjective Probability: A Judgement of Representativeness,” *Cognitive Psychology*, 3, 430–454.
- KOEHLER, J. J. AND C. A. CONLEY (2003): “The “hot hand” myth in professional basketball,” *Journal of Sport and Exercise Psychology*, 25, 253–259.
- KONOLD, C. (1995): “Confessions of a coin flipper and would-be instructor,” *The American Statistician*, 49, 203–209.
- LANCASTER, T. (2000): “The incidental parameter problem since 1948,” *Journal of Econometrics*, 95, 391–413.

- LEE, M. AND G. SMITH (2002): “Regression to the mean and football wagers,” *Journal of Behavioral Decision Making*, 15, 329–342.
- LOH, R. K. AND M. WARACHKA (2012): “Streaks in Earnings Surprises and the Cross-Section of Stock Returns,” *Management Science*, 58, 1305–1321.
- MALKIEL, B. G. (2011): *A random walk down Wall Street: the time-tested strategy for successful investing*, New York: W. W. Norton & Company.
- MILLER, J. B. AND A. SANJURJO (2014): “A Cold Shower for the Hot Hand Fallacy,” Working Paper. Available at SSRN: <https://ssrn.com/abstract=2450479>.
- (2015a): “A Bridge from Monty Hall to the (Anti-)Hot Hand: Restricted Choice, Selection Bias, and Empirical Practice,” Working Paper. Available at SSRN: <https://ssrn.com/abstract=2709837>.
- (2015b): “Is the Belief in the Hot Hand a *Fallacy* in the NBA Three Point Shootout?” Working Paper. Available at SSRN: <https://ssrn.com/abstract=2611987>.
- (2016): “Selection Bias, Sample-size Neglect, and the Persistence of Gambler’s Fallacy Beliefs,” *Working Paper*.
- (2017a): “The Hot Hand, and the Hot Hand *Fallacy*: A Critical Survey of the Literature,” Working Paper. Available at SSRN: XXto be updatedXX.
- (2017b): “A Visible Hand? Betting on the hot hand in Gilovich, Vallone, and Tversky (1985),” Working Paper. Available at SSRN: <https://ssrn.com/abstract=3032826>.
- NALEBUFF, B. (1987): “Puzzles: Choose a curtain, duel-ity, two point conversions, and more,” *Journal of Economic Perspectives*, 157–163.
- NARAYANAN, S. AND P. MANCHANDA (2012): “An empirical analysis of individual level casino gambling behavior,” *Quantitative Marketing and Economics*, 10, 27–62.
- NERLOVE, M. (1967): “Experimental Evidence On The Estimation Of Dynamic Economic Relations From A Time Series Of Cross-Section,” *The Economic Studies Quarterly (Tokyo. 1950)*, 18, 42–74.
- (1971): “Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections,” *Econometrica*, 39, 359–382.
- NEYMAN, J. AND E. L. SCOTT (1948): “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, 16, 1–32.
- NICKELL, S. (1981): “Biases in Dynamic Models with Fixed Effects,” *Econometrica*, 49, 1417–1426.
- NICKERSON, R. S. (2007): “Penney Ante: Counterintuitive Probabilities in Coin Tossing,” *The UMAP Journal*.
- PAUL, R. J. AND A. P. WEINBACH (2005): “Bettor Misperceptions in the NBA: The Overbetting of Large Favorites and the ‘Hot Hand’,” *Journal of Sports Economics*, 6, 390–400.

- RABIN, M. (2002): “Inference by Believers in the Law of Small Numbers,” *Quarterly Journal of Economics*, 117, 775–816.
- RABIN, M. AND D. VAYANOS (2010): “The Gamblers and Hot-Hand Fallacies: Theory and Applications,” *Review of Economic Studies*, 77, 730–778.
- RAO, J. M. (2009a): “Experts’ Perceptions of Autocorrelation: The Hot Hand Fallacy Among Professional Basketball Players,” Working Paper.
- (2009b): “When the Gambler’s Fallacy becomes the Hot Hand Fallacy: An Experiment with Experts and Novices,” Working Paper.
- RINOTT, Y. AND M. BAR-HILLEL (2015): “Comments on a ‘Hot Hand’ Paper by Miller and Sanjurjo,” Federmann Center For The Study Of Rationality, The Hebrew University Of Jerusalem. Discussion Paper # 688 (August 11).
- ROBERTS, R. S., W. O. SPITZER, T. DELMORE, AND D. L. SACKETT (1978): “An empirical demonstration of Berkson’s bias,” *Journal of Chronic Diseases*, 31, 119–128.
- SELVIN, S. (1975): “A Problem in Probability (letter to the editor),” *The American Statistician*, 29, 67.
- SHAMAN, P. AND R. A. STINE (1988): “The bias of autoregressive coefficient estimators,” *Journal of the American Statistical Association*, 83, 842–848.
- SINKEY, M. AND T. LOGAN (2013): “Does the Hot Hand Drive the Market?” *Eastern Economic Journal*, Advance online publication, doi:10.1057/ej.2013.33.
- SMITH, G., M. LEVERE, AND R. KURTZMAN (2009): “Poker Player Behavior After Big Wins and Big Losses,” *Management Science*, 55, 1547–1555.
- STONE, D. F. (2012): “Measurement error and the hot hand,” *The American Statistician*, 66, 61–66, working paper.
- STONE, D. F. AND J. ARKES (2017): “March Madness? Underreaction to Hot and Cold Hands in NCAA Basketball,” *Working Paper*, –.
- SUNDALI, J. AND R. CROSON (2006): “Biases in casino betting: The hot and the gamblers fallacy,” *Judgement and Decision Making*, 1, 1–12.
- THALER, R. H. AND C. R. SUNSTEIN (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press.
- TVERSKY, A. AND T. GILOVICH (1989): “The cold facts about the “hot hand” in basketball,” *Chance*, 2, 16–21.
- VOS SAVANT, M. (1990): “Ask Marilyn,” *Parade Magazine*, 15.
- WARDROP, R. L. (1995): “Simpson’s Paradox and the Hot Hand in Basketball,” *The American Statistician*, 49, 24–28.

- (1999): “Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting,” *Unpublished manuscript*, 1, 1–20.
- WHITE, H. (1999): *Asymptotic Theory for Econometricians*, Academic Press: London.
- XU, J. AND N. HARVEY (2014): “Carry on winning: The gambler’s fallacy creates hot hand effects in online gambling,” *Cognition*, 131, 173 – 180.
- YAARI, G. AND S. EISENMANN (2011): “The Hot (Invisible?) Hand: Can Time Sequence Patterns of Success/Failure in Sports Be Modeled as Repeated Random Independent Trials?” *PLoS One*, 6, 1–10.
- YUAN, J., G.-Z. SUN, AND R. SIU (2014): “The Lure of Illusory Luck: How Much Are People Willing to Pay for Random Shocks,” *Journal of Economic Behavior & Organization*, forthcoming.
- YULE, G. U. (1926): “Why do we Sometimes get Nonsense-Correlations between Time-Series?—A Study in Sampling and the Nature of Time-Series,” *Journal of the Royal Statistical Society*, 89, 1–63.

## A Appendix: Section 2 Proofs

### A.1 Proof of Theorem 1 (Section 2)

Define  $F := \{\mathbf{x} \in \{0, 1\}^n : I_k(\mathbf{x}) \neq \emptyset\}$  to be the sample space of sequences for which  $\hat{P}_k(\mathbf{X})$  is well defined. The probability distribution over  $F$  is given by  $\mathbb{P}(A|F) := \mathbb{P}(A \cap F)/\mathbb{P}(F)$  for  $A \subseteq \{0, 1\}^n$ , where  $\mathbb{P}(\mathbf{X} = \mathbf{x}) = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}$ .

Let the random variable  $X_\tau$  represent the outcome of the randomly “drawn” trial  $\tau$ , which is selected as a result of the two-stage procedure that: (i) draws a sequence  $\mathbf{x}$  at random from  $F$ , according to the distribution  $\mathbb{P}(\mathbf{X} = \mathbf{x}|F)$ , and (ii) draws a trial  $\tau$  at random from  $\{k+1, \dots, n\}$ , according to the distribution  $\mathbb{P}(\tau = t|\mathbf{X} = \mathbf{x})$ . Let  $\tau$  be a uniform draw from the trials in sequence  $\mathbf{X}$  that immediately follow  $k$  consecutive successes, i.e. for  $\mathbf{x} \in F$ ,  $\mathbb{P}(\tau = t|\mathbf{X} = \mathbf{x}) = 1/|I_k(\mathbf{x})|$  for  $t \in I_k(\mathbf{x})$ , and  $\mathbb{P}(\tau = t|\mathbf{X} = \mathbf{x}) = 0$  for  $t \in I_k(\mathbf{x})^C \cap \{k+1, \dots, n\}$ .<sup>43</sup> It follows that the unconditional probability distribution of  $\tau$  over all trials that can possibly follow  $k$  consecutive successes is given by  $\mathbb{P}(\tau = t|F) = \sum_{\mathbf{x} \in F} \mathbb{P}(\tau = t|\mathbf{X} = \mathbf{x}, F) \mathbb{P}(\mathbf{X} = \mathbf{x}|F)$ , for  $t \in \{k+1, \dots, n\}$ . The probability that this randomly drawn trial is a success,  $\mathbb{P}(X_\tau = 1|F)$ , must be equal to the expected proportion,  $E[\hat{P}_k(\mathbf{X})|F]$ .<sup>44</sup>

Note that  $\mathbb{P}(X_\tau = 1|F) = \sum_{t=k+1}^n \mathbb{P}(X_t = 1|\tau = t, F) \mathbb{P}(\tau = t|F)$ , and  $\mathbb{P}(\tau = t|F) > 0$  for  $t \in \{k+1, \dots, n\}$ . Below, we demonstrate that  $\mathbb{P}(X_t = 1|\tau = t, F) < p$  when  $t < n$ , and that  $\mathbb{P}(X_t = 1|\tau = n, F) = p$ , which, taken together, guarantee that  $\mathbb{P}(X_\tau = 1|F) < p$ .

First we observe that  $\mathbb{P}(X_t = 1|\tau = t, F) = \mathbb{P}(X_t = 1|\tau = t, F_t)$ , where  $F_t := \{\mathbf{x} \in \{0, 1\}^n : \prod_{i=t-k}^{t-1} x_i = 1\}$ . Bayes Rule then yields:

$$\begin{aligned} \frac{\mathbb{P}(X_t = 1|\tau = t, F_t)}{\mathbb{P}(X_t = 0|\tau = t, F_t)} &= \frac{\mathbb{P}(\tau = t|X_t = 1, F_t) \mathbb{P}(X_t = 1|F_t)}{\mathbb{P}(\tau = t|X_t = 0, F_t) \mathbb{P}(X_t = 0|F_t)} \\ &= \frac{\mathbb{P}(\tau = t|X_t = 1, F_t)}{\mathbb{P}(\tau = t|X_t = 0, F_t)} \frac{p}{1-p}. \end{aligned}$$

Therefore, for the case of  $t \in \{k+1, \dots, n-1\}$ , in order to show that  $\mathbb{P}(X_t = 1|\tau = t, F) = \mathbb{P}(X_t = 1|\tau = t, F_t) < p$  it suffices to show that  $\mathbb{P}(\tau = t|X_t = 1, F_t) < \mathbb{P}(\tau = t|X_t = 0, F_t)$ , which follows below:

<sup>43</sup>For  $\mathbf{x} \in F^C$  no trial is drawn, which we can represent as  $\mathbb{P}(\tau = 1|\mathbf{X} = \mathbf{x}) = 1$  (for example).

<sup>44</sup>The identity follows by the law of total probability, with the key observation that  $\hat{P}_k(\mathbf{x}) = \sum_{t \in I_k(\mathbf{x})} x_t \cdot \frac{1}{|I_k(\mathbf{x})|} = \sum_{t=k+1}^n \mathbb{P}(X_t = 1|\tau = t, \mathbf{X} = \mathbf{x}, F) \mathbb{P}(\tau = t|\mathbf{X} = \mathbf{x}, F)$ .

$$\begin{aligned}
\mathbb{P}(\tau = t | X_t = 0, F_t) &= \sum_{\mathbf{x} \in F_t: x_t=0} \mathbb{P}(\tau = t | X_t = 0, \mathbf{X} = \mathbf{x}, F_t) \mathbb{P}(\mathbf{X} = \mathbf{x} | X_t = 0, F_t) \\
&= \sum_{\mathbf{x} \in F_t: x_t=0} \mathbb{P}(\tau = t | X_t = 0, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t) \mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 0, F_t) \quad (2) \\
&> \sum_{\mathbf{x} \in F_t: x_t=0} \mathbb{P}(\tau = t | X_t = 1, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t) \mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 1, F_t) \quad (3) \\
&= \sum_{\mathbf{x} \in F_t: x_t=1} \mathbb{P}(\tau = t | X_t = 1, \mathbf{X} = \mathbf{x}, F_t) \mathbb{P}(\mathbf{X} = \mathbf{x} | X_t = 1, F_t) \\
&= \mathbb{P}(\tau = t | X_t = 1, F_t)
\end{aligned}$$

where in (2), given  $\mathbf{x}$ , we define  $\mathbf{x}_{-t} := (x_1, \dots, x_{t-1}, x_{t+1}, \dots, x_n)$ . To obtain the inequality in (3) we observe that: (i)  $\mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 0, F_t) = \mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 1, F_t)$  because  $\mathbf{X}$  is a sequence of i.i.d. Bernoulli trials, and (ii)  $\mathbb{P}(\tau = t | X_t = 1, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t) < \mathbb{P}(\tau = t | X_t = 0, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t)$  because  $\tau$  is drawn at random (uniformly) from the set  $I_k(\mathbf{x})$ , which contains at least one more element (trial  $t + 1$ ) if  $x_t = 1$  rather than  $x_t = 0$ .

For the case of  $t = n$  we follow the above steps until (3), at which point an equality now emerges as  $X_n = 1$  no longer yields an additional trial from which to draw, because trial  $n$  is terminal. This implies that  $\mathbb{P}(\tau = n | X_n = 1, F_n) = \mathbb{P}(\tau = n | X_n = 0, F_n)$ .

Taking these two facts together: (i)  $\mathbb{P}(X_t = 1 | \tau = t, F) < p$ , for  $k + 1 \leq t < n$ , and (ii)  $\mathbb{P}(X_n = 1 | \tau = n, F) = p$ , it immediately follows that  $\mathbb{P}(X_\tau = 1 | F) < p$ .<sup>45</sup>

■

## A.2 Asymptotic Unbiasedness

*Proof that the proportion is asymptotically unbiased*

To demonstrate that  $\hat{P}_k(\mathbf{X})$  is a consistent estimator of  $\mathbb{P}(X_t = 1 | \prod_{j=t-k}^{t-1} X_j = 1)$ , first define  $Y_{k,i} := \prod_{j=i-k+1}^i X_j$  for  $i \geq k$ . With this,  $\hat{P}_k(\mathbf{X}) = \sum_{i=k+1}^n Y_{k+1,i} / \sum_{i=k}^{n-1} Y_{k,i}$ . Note that each of the respective sequences  $\{Y_{k,i}\}, \{Y_{k+1,i}\}$  are asymptotically uncorrelated ( $k$  fixed). Therefore, their time averages converge to their respective means almost surely, i.e.  $1/(n-k) \sum_{i=k}^{n-1} Y_{k,i} \xrightarrow{a.s.} E[Y_{k,i}] = p^k$ , and  $1/(n-k) \sum_{i=k+1}^n Y_{k+1,i} \xrightarrow{a.s.} E[Y_{k+1,i}] = p^{k+1}$ .<sup>46</sup> The continuous mapping theorem implies that  $\hat{P}_k(\mathbf{X}) \xrightarrow{a.s.} p = \mathbb{P}(X_t = 1 | \prod_{j=t-k}^{t-1} X_j = 1)$ , which in turn implies consistency.

<sup>45</sup>Note that the proof does not require that the Bernoulli trials be identically distributed. Instead, we could allow the probability distribution to vary, with  $\mathbb{P}(X_i = 1) = p^i$  for  $i = 1, \dots, n$ , in which case our result would be that  $\mathbb{P}(X_\tau = 1 | F) < E[p_\tau | F]$ .

<sup>46</sup>See Definition 3.55 and Theorem 3.57 from White (1999).

*Proof that weighted proportions are asymptotically unbiased*

In order to prove the assertion made in Footnote 5 that the weighted average proportion over multiple realized sequences is a consistent estimator of  $\mathbb{P}(X_t = 1 | \prod_{j=t-k}^{t-1} X_j = 1)$ , we first define  $Y_{k,i} := \prod_{j=i-k+1}^i X_j$  for  $i \geq k$ , just as we did in the previous proof. Then, we note that: (i) the number of trials that follow  $k$  consecutive successes in the weighted proportion taken over  $T$  sequences is given by  $\sum_{t=1}^T Z_{k,t}$ , where  $Z_{k,t} = \sum_{i=n(t-1)+k}^{nt-1} Y_{k,i}$ , and (ii) the number of successes on these trials is given by  $\sum_{t=1}^T Z_{k+1,t}$ , where  $Z_{k+1,t} = \sum_{i=n(t-1)+k+1}^{nt} Y_{k+1,i}$ . Because  $Z_{k,t}$  are i.i.d. with  $E[Z_{k,t}] = (n-k)p^k$ , it follows that  $1/T \sum_{t=1}^T Z_{k,t} \xrightarrow{a.s.} E[Z_{k,t}] = (n-k)p^k$ ; similarly,  $1/T \sum_{t=1}^T Z_{k+1,t} \xrightarrow{a.s.} E[Z_{k+1,t}] = (n-k)p^{k+1}$ . Then, the continuous mapping theorem yields the desired consistency of the weighted proportion (after sequence  $T$ ), i.e.  $\sum_{t=1}^T Z_{k+1,t} / \sum_{t=1}^T Z_{k,t} \xrightarrow{a.s.} p = \mathbb{P}(X_t = 1 | \prod_{j=t-k}^{t-1} X_j = 1)$ .

### A.3 Formula for the expected proportion (special case of $k = 1$ )

The following lemma shows that the expected proportion  $\hat{P}_1(\mathbf{X})$ , conditional on a known number of successes  $N_1(\mathbf{X}) = n_1$ , satisfies the sampling-without-replacement formula, which for any given trial is less than the probability of success  $\mathbb{P}(X_i | N_1(\mathbf{X}) = n_1) = \frac{n_1}{n}$ .

**Lemma 1** *Let  $n > 1$ . Then*

$$E \left[ \hat{P}_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq \emptyset, N_1(\mathbf{X}) = n_1 \right] = \frac{n_1 - 1}{n - 1} \quad (4)$$

for  $0 \leq n_1 \leq n$ .

**Proof:** As in the proof of Theorem 1, let  $\tau$  be drawn at random from  $I_1(\mathbf{X})$ , which is non-empty when  $N_1(\mathbf{X}) = n_1 \geq 2$  (the result is trivial when  $n_1 = 1$ ). In order to ease notation we let probability  $\mathbb{P}(\cdot)$  represent the conditional probability  $\mathbb{P}(\cdot | N_1(\mathbf{X}) = n_1)$ , which is defined over the subsets of  $\{\mathbf{x} \in \{0, 1\}^n : N_1(\mathbf{x}) = n_1\}$ .

$$E[\hat{P}_1(\mathbf{X}) | N_1(\mathbf{X}) = n_1, I_1(\mathbf{X}) \neq \emptyset] = \mathbb{P}(X_\tau = 1) \quad (5)$$

$$= \mathbb{P}(X_\tau = 1 | \tau < n) \mathbb{P}(\tau < n) + \mathbb{P}(X_n = 1 | \tau = n) \mathbb{P}(\tau = n)$$

$$= \sum_{t=2}^{n-1} \mathbb{P}(X_t = 1 | \tau = t) \frac{1}{n-1} + \mathbb{P}(X_n = 1 | \tau = n) \frac{1}{n-1} \quad (6)$$

$$= \frac{n-1}{n-2} \left( \frac{n_1}{n} - \frac{1}{n-1} \right) \frac{n-2}{n-1} + \frac{n_1}{n} \frac{1}{n-1} \quad (7)$$

$$= \frac{n_1 - 1}{n - 1}$$

In (5), equality follows by an argument analogous to that provided in the proof of Theorem 1. In (6), equality follows from the fact that  $\mathbb{P}(\tau = t) = 1/(n-1)$  for all  $t \in \{2, 3, \dots, n\}$ .<sup>47</sup> In (7), equality follows from using an application of Bayes rule to derive  $\mathbb{P}(X_t = 1|\tau = t)$ , which satisfies:

$$\mathbb{P}(X_t = 1|\tau = t) = \begin{cases} \frac{n-1}{n-2} \left( \frac{n_1}{n} - \frac{1}{n-1} \right) & \text{for } t = 2, \dots, n-1 \\ \frac{n_1}{n} & \text{for } t = n \end{cases} \quad (8)$$

In particular,

$$\begin{aligned} \mathbb{P}(X_t = 1|\tau = t) &= \frac{\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)\mathbb{P}(X_{t-1} = 1|X_t = 1)\mathbb{P}(X_t = 1)}{\mathbb{P}(\tau = t)} \\ &= \mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1) \frac{n_1(n_1 - 1)}{n} \end{aligned} \quad (9)$$

where for all  $t$ ,  $\mathbb{P}(X_{t-1} = 1|X_t = 1) = (n_1 - 1)/(n - 1)$ , which is the likelihood that relates to sampling-without-replacement. For  $t < n$ ,  $\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)$ , which is the likelihood that relates to the arrangement of successes and failures, satisfies:

$$\begin{aligned} \mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1) &= E \left[ \frac{1}{M} \mid X_{t-1} = 1, X_t = 1 \right] \\ &= \sum_{x \in \{0,1\}} E \left[ \frac{1}{M} \mid X_{t-1} = 1, X_t = 1, X_n = x \right] \mathbb{P}(X_n = x|X_{t-1} = 1, X_t = 1) \\ &= \frac{1}{n_1} \frac{n_0}{n-2} + \frac{1}{n_1-1} \frac{n_1-2}{n-2} \\ &= \frac{1}{n-2} \left( \frac{n_0}{n_1} + \frac{n_1-2}{n_1-1} \right) \end{aligned}$$

where  $M := |I_1(\mathbf{X})|$ , i.e.  $M = n_1 - X_n$ . In the case that  $t = n$ , clearly  $\mathbb{P}(\tau = n|X_{n-1} = 1, X_n = 1) = \frac{1}{n_1-1}$ .

■

#### *Formulae for expected value of the proportion*

The conditional expectation in Lemma 1 can be combined with  $\mathbb{P}(N_1(\mathbf{X}) = n_1|I_1(\mathbf{X}) \neq \emptyset)$  to express the expected proportion in terms of just  $n$  and  $p$ .<sup>48</sup>

<sup>47</sup>Note that  $\mathbb{P}(\tau = t) = \sum_{\mathbf{x}: N_1(\mathbf{x})=n_1} \mathbb{P}(\tau = t|\mathbf{X} = \mathbf{x})\mathbb{P}(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}: N_1(\mathbf{x})=n_1, x_{t-1}=1} \frac{1}{n_1-x_n} \frac{1}{\binom{n}{n_1}} = \frac{1}{\binom{n}{n_1}} \left[ \binom{n-2}{n_1-1} \frac{1}{n_1} + \binom{n-2}{n_1-2} \frac{1}{n_1-1} \right] = \frac{1}{n-1}$ .

<sup>48</sup>In a comment written about this paper, Rinott and Bar-Hillel (2015) provide an alternative proof for this theorem.

**Theorem 2** *Let  $n > 2$  and  $0 < p < 1$ . Then*

$$E \left[ \hat{P}_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq \emptyset \right] = \frac{\left[ p - \frac{1-(1-p)^n}{n} \right] \frac{n}{n-1}}{1 - (1-p)^{n-1}} < p \quad (10)$$

**Proof:** We first observe that in light of Lemma 1, Equation 10 can be written as follows:

$$\begin{aligned} E \left[ \hat{P}_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq \emptyset \right] &= E \left[ E \left[ \hat{P}_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq \emptyset, N_1(\mathbf{X}) = n_1 \right] \right] \\ &= E \left[ \frac{N_1(\mathbf{x}) - 1}{n - 1} \mid I_1(\mathbf{X}) \neq \emptyset \right] \end{aligned}$$

The expected value can then be computed using the binomial distribution, which yields:

$$\begin{aligned} E \left[ \frac{N_1(\mathbf{x}) - 1}{n - 1} \mid I_1(\mathbf{X}) \neq \emptyset \right] &= C \sum_{n_1=1}^n p^{n_1} (1-p)^{n-n_1} \left[ \binom{n}{n_1} - U(n, n_1) \right] \cdot \frac{n_1 - 1}{n - 1} \\ &= \frac{\sum_{n_1=2}^n \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1} \frac{n_1-1}{n-1}}{1 - (1-p)^n - p(1-p)^{n-1}} \\ &= \frac{\frac{1}{n-1} [(np - np(1-p)^{n-1}) - (1 - (1-p)^n - np(1-p)^{n-1})]}{1 - (1-p)^n - p(1-p)^{n-1}} \\ &= \frac{\left[ p - \frac{1-(1-p)^n}{n} \right] \frac{n}{n-1}}{1 - (1-p)^{n-1}} \end{aligned}$$

where  $U(n, n_1)$  is the number of sequences with  $n_1$  successes for which the proportion is undefined, and  $C$  is the constant that normalizes the total probability to 1. The second line follows because  $U_1(n, n_1) = 0$  for  $n_1 > 1$ ,  $U_1(n, 0) = U_1(n, 1) = 1$ , and  $C = 1/[1 - (1-p)^n - p(1-p)^{n-1}]$ .

Finally, by letting  $q := 1 - p$  it is straightforward to show that the bias in  $\hat{P}_1(\mathbf{X})$  is negative:

$$\begin{aligned} E \left[ \hat{P}_1(\mathbf{X}) - p \mid I_1(\mathbf{X}) \neq \emptyset \right] &= \frac{\left[ p - \frac{1-q^n}{n} \right] \frac{n}{n-1}}{1 - q^{n-1}} - p \\ &= \frac{(n-1)(q^{n-1} - q^n) - (q - q^n)}{(n-1)(1 - q^{n-1})} \\ &< 0 \end{aligned}$$

The inequality follows from  $f(x) = q^x$  being strictly decreasing and convex, which implies that  $q - q^n > (n-1)(q^{n-1} - q^n)$ .

■

## A.4 Expected difference in proportions

Let  $D_k$  be the difference in the probability of success when comparing trials that immediately follow  $k$  consecutive successes with trials that immediately follow  $k$  consecutive failures. That is,  $D_k := \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) - \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} (1 - X_j) = 1)$ . An estimator of  $D_k$  that is used in the hot hand fallacy literature (see Section 3) is  $\hat{D}_k(\mathbf{x}) := \hat{P}_k(\mathbf{x}) - [1 - \hat{Q}_k(\mathbf{X})]$ , where  $\hat{Q}_k(\mathbf{X})$  is the proportion of failures on the subset of trials that immediately follow  $k$  consecutive failures,  $J_k(\mathbf{X}) := \{j : \prod_{i=j-k}^{j-1} (1 - X_i) = 1\} \subseteq \{k+1, \dots, n\}$ .

### A.4.1 Proof of the bias in the difference

We extend the proof of Theorem 1 to show that  $\hat{D}_k(\mathbf{X})$  is a biased estimator of  $D_k$ . Recall that  $I_k(\mathbf{X})$  is the subset of trials that immediately follow  $k$  consecutive successes, i.e.  $I_k(\mathbf{X}) := \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \dots, n\}$ . Analogously, let  $J_k(\mathbf{X})$  be the subset of trials that immediately follow  $k$  consecutive failures, i.e.  $J_k(\mathbf{X}) := \{j : \prod_{i=j-k}^{j-1} (1 - X_i) = 1\} \subseteq \{k+1, \dots, n\}$ .

**Theorem 3** *Let  $\mathbf{X} = \{X_i\}_{i=1}^n$ ,  $n \geq 3$ , be a sequence of independent Bernoulli trials, each with probability of success  $0 < p < 1$ . Let  $\hat{P}_k(\mathbf{X})$  be the proportion of successes on the subset of trials  $I_k(\mathbf{X})$  that immediately follow  $k$  consecutive successes, and  $\hat{Q}_k(\mathbf{X})$  be the proportion of failures on the subset of trials  $J_k(\mathbf{X})$  that immediately follow  $k$  consecutive failures.  $\hat{D}_k(\mathbf{x}) := \hat{P}_k(\mathbf{x}) - [1 - \hat{Q}_k(\mathbf{X})]$  is a biased estimator of  $D_k := \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) - \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} (1 - X_j) = 1) \equiv 0$  for all  $k$  such that  $1 \leq k < n/2$ . In particular,*

$$E[\hat{D}_k(\mathbf{X}) \mid I_k(\mathbf{X}) \neq \emptyset, J_k(\mathbf{X}) \neq \emptyset] < 0 \quad (11)$$

**Proof:** Following the notation used in the proof of Theorem 1, let  $F := \{\mathbf{x} \in \{0, 1\}^n : I_k(\mathbf{x}) \neq \emptyset\}$  and  $G := \{\mathbf{x} \in \{0, 1\}^n : J_k(\mathbf{x}) \neq \emptyset\}$ . We will show the following:

$$\begin{aligned} E[\hat{D}_k(\mathbf{x})|F, G] &= E[\hat{P}_k(\mathbf{X})|F, G] + E[\hat{Q}_k(\mathbf{X})|F, G] - 1 \\ &= \mathbb{P}(X_\tau = 1|F, G) + \mathbb{P}(X_\sigma = 0|F, G) - 1 \end{aligned} \quad (12)$$

$$< p + (1 - p) - 1 \quad (13)$$

$$= 0 \quad (14)$$

where in (12), as in the proof of Theorem 1,  $\tau$  is a random draw from  $I_k(\mathbf{x})$  and  $\sigma$  is an analogous random draw from  $J_k(\mathbf{x})$ . In particular, we will demonstrate that the inequality in (13) holds by showing that  $\mathbb{P}(X_\tau = 1|F, G) < p$ , which, by symmetry, implies that  $\mathbb{P}(X_\sigma = 0|F, G) < 1 - p$ .

To show that  $\mathbb{P}(X_\tau = 1|F, G) < p$  we use an approach similar to that presented in the proof of Theorem 1. In particular, note that  $\mathbb{P}(X_\tau = 1|F, G) = \sum_{t=k+1}^n \mathbb{P}(X_t = 1|\tau = t, F, G)\mathbb{P}(\tau = t|F, G)$ , and  $\mathbb{P}(\tau = t|F, G) > 0$  for  $t \in \{k+1, \dots, n\}$ . In what follows, we demonstrate that  $\mathbb{P}(X_t = 1|\tau = t, F, G) < p$  when  $t < n$ , and that  $\mathbb{P}(X_t = 1|\tau = n, F, G) = p$ , which, taken together, guarantee that  $\mathbb{P}(X_\tau = 1|F, G) < p$ .

First we observe that  $\mathbb{P}(X_t = 1|\tau = t, F, G) = \mathbb{P}(X_t = 1|\tau = t, F_t, G)$ , where  $F_t := \{\mathbf{x} \in \{0, 1\}^n : \prod_{i=t-k}^{t-1} x_i = 1\}$ . Bayes Rule then yields:

$$\begin{aligned} \frac{\mathbb{P}(X_t = 1|\tau = t, F_t, G)}{\mathbb{P}(X_t = 0|\tau = t, F_t, G)} &= \frac{\mathbb{P}(\tau = t, G|X_t = 1, F_t) \mathbb{P}(X_t = 1|F_t)}{\mathbb{P}(\tau = t, G|X_t = 0, F_t) \mathbb{P}(X_t = 0|F_t)} \\ &= \frac{\mathbb{P}(\tau = t, G|X_t = 1, F_t)}{\mathbb{P}(\tau = t, G|X_t = 0, F_t)} \frac{p}{1-p}. \end{aligned}$$

Therefore, for the case of  $t \in \{k+1, \dots, n-1\}$ , in order to show that  $\mathbb{P}(X_t = 1|\tau = t, F, G) = \mathbb{P}(X_t = 1|\tau = t, F_t, G) < p$  it suffices to show that  $\mathbb{P}(\tau = t, G|X_t = 1, F_t) < \mathbb{P}(\tau = t, G|X_t = 0, F_t)$ ,

which follows below:

$$\begin{aligned}
\mathbb{P}(\tau = t, G | X_t = 0, F_t) &= \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0}} \mathbb{P}(\tau = t, \mathbf{X} = \mathbf{x} | X_t = 0, F_t) \\
&= \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \mathbf{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t, \mathbf{X} = \mathbf{x} | X_t = 0, F_t) \\
&\quad + \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \mathbf{x}_{-t}) \notin F_t \cap G}} \mathbb{P}(\tau = t, \mathbf{X} = \mathbf{x} | X_t = 0, F_t) \\
&\geq \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \mathbf{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t, \mathbf{X} = \mathbf{x} | X_t = 0, F_t) \\
&= \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \mathbf{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t | \mathbf{X} = \mathbf{x}, X_t = 0, F_t) \mathbb{P}(\mathbf{X} = \mathbf{x} | X_t = 0, F_t) \\
&= \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \mathbf{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t | X_t = 0, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t) \mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 0, F_t) \\
&> \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \mathbf{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t | X_t = 1, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t) \mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 1, F_t) \\
&= \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 1 \\ (1, \mathbf{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t | \mathbf{X} = \mathbf{x}, X_t = 1, F_t) \mathbb{P}(\mathbf{X} = \mathbf{x} | X_t = 1, F_t) \\
&= \sum_{\substack{\mathbf{x} \in F_t \cap G: \\ x_t = 1}} \mathbb{P}(\tau = t, \mathbf{X} = \mathbf{x} | X_t = 1, F_t) \\
&= \mathbb{P}(\tau = t, G | X_t = 1, F_t)
\end{aligned} \tag{15}$$

where in (15), given  $\mathbf{x}$ , we define  $\mathbf{x}_{-t} := (x_1, \dots, x_{t-1}, x_{t+1}, \dots, x_n)$ , and  $(b, \mathbf{x}_{-t}) := (x_1, \dots, x_{t-1}, b, x_{t+1}, \dots, x_n)$ .<sup>49</sup> The inequality in (16) follows for the same reason as the inequality in line (3) of Theorem 1. In particular,  $\mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 0, F_t) = \mathbb{P}(\mathbf{X}_{-t} = \mathbf{x}_{-t} | X_t = 1, F_t)$  because  $\mathbf{X}$  is a sequence of i.i.d. Bernoulli trials, and  $\mathbb{P}(\tau = t | X_t = 1, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t) < \mathbb{P}(\tau = t | X_t = 0, \mathbf{X}_{-t} = \mathbf{x}_{-t}, F_t)$  because  $\tau$  is drawn at random (uniformly) from the set  $I_k(\mathbf{x})$ , which contains at least one more element (trial  $t + 1$ ) if  $x_t = 1$  rather than  $x_t = 0$ .

<sup>49</sup>Note that the second sum will have no terms for  $t \geq n - k$ .

For the case of  $t = n$  we follow the above steps until (16), at which point an equality now emerges, as  $X_n = 1$  no longer yields an additional trial from which to draw, because trial  $n$  is terminal. This implies that  $\mathbb{P}(\tau = n|X_n = 1, F_n, G) = \mathbb{P}(\tau = n|X_n = 0, F_n, G)$ .

Taking these two facts together: (i)  $\mathbb{P}(X_t = 1|\tau = t, F, G) < p$ , for  $k + 1 \leq t < n$ , and (ii)  $\mathbb{P}(X_n = 1|\tau = n, F, G) = p$ , it immediately follows that  $\mathbb{P}(X_\tau = 1|F, G) < p$ .

■

#### A.4.2 Formula for the expected difference in proportions (special case of $k = 1$ )

In the case of  $k = 1$  the expected difference in proportions admits a simple representation that is independent of  $p$ .

**Theorem 4** *Let  $\hat{D}_1(\mathbf{X}) := \hat{P}_1(\mathbf{X}) - (1 - \hat{Q}_1(\mathbf{X}))$ . If  $n > 2$  and  $0 < p < 1$  then*

$$E \left[ \hat{D}_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq \emptyset, J_1(\mathbf{X}) \neq \emptyset \right] = -\frac{1}{n-1}$$

**Proof:** The method of proof is to first show that if  $n > 2$  and  $1 \leq n_1 \leq n - 1$  then:

$$E \left[ \hat{D}_1(\mathbf{X}) \mid N_1(\mathbf{X}) = n_1, I_1(\mathbf{X}) \neq \emptyset, J_1(\mathbf{X}) \neq \emptyset \right] = -\frac{1}{n-1}$$

which leaves us just one step from the desired result.

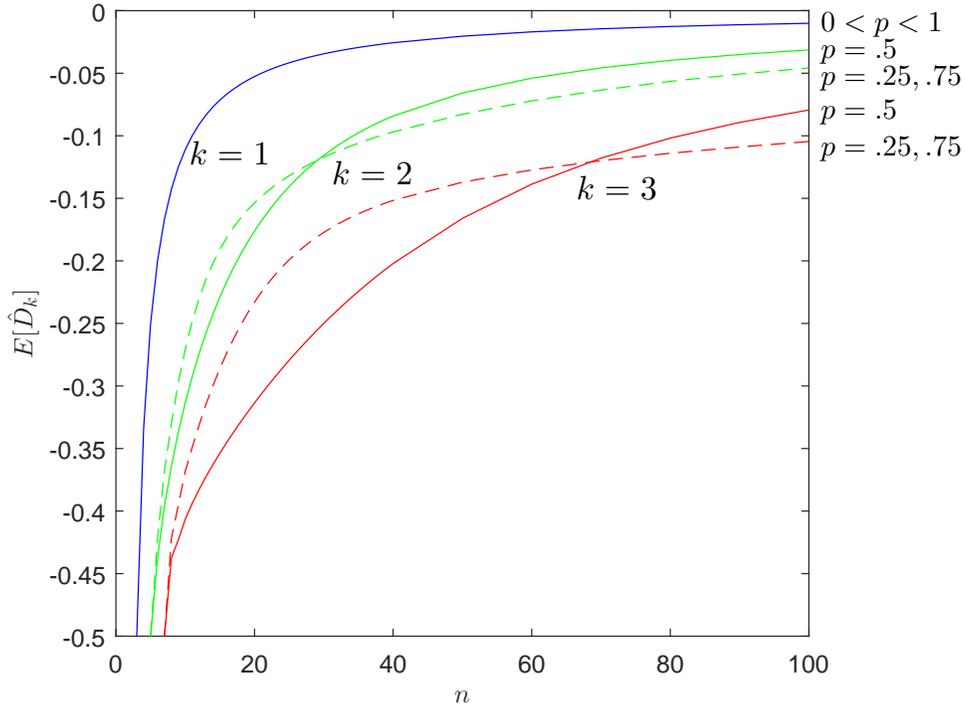
First, consider the case that  $1 < n_1 < n - 1$ . In this case  $\hat{D}_1(\mathbf{x}) := \hat{P}_1(\mathbf{x}) - (1 - \hat{Q}_1(\mathbf{x}))$  is defined for all sequences. Therefore, by the linearity of the expectation, and applying Lemma 1 to  $\hat{Q}_1(\mathbf{X})$  (by symmetry), we have:

$$\begin{aligned} E[\hat{D}_1(\mathbf{X})|N_1(\mathbf{X}) = n_1] &= E[\hat{P}_1(\mathbf{X})|N_1(\mathbf{X}) = n_1] - E(1 - \hat{Q}_1(\mathbf{X})|N_1(\mathbf{X}) = n_1) \\ &= \frac{n_1 - 1}{n - 1} - \left( 1 - \frac{n_0 - 1}{n - 1} \right) \\ &= -\frac{1}{n - 1} \end{aligned}$$

If  $n_1 = 1$  then  $\hat{D}_1$  is defined for all sequences that do not have a 1 in the final position; there are  $n - 1$  such sequences. The sequence with the 1 in the first position yields  $\hat{D}_1 = 0$ , while the other  $n - 2$  sequences yield  $\hat{D}_1 = -1/(n - 2)$ . Therefore,  $E \left[ \hat{D}_1(\mathbf{X}) \mid N_1(\mathbf{X}) = 1, I_1(\mathbf{X}) \neq \emptyset, J_1(\mathbf{X}) \neq \emptyset \right] = -1/(n - 1)$ .

Now consider the case of  $n_1 = n - 1$ . The argument for this case is analogous, with  $\hat{D}_1$  undefined for the sequence with the zero in the last position, equal to 0 for the sequence with the zero in the first position, and equal to  $-1/(n - 2)$  for all other sequences.

Finally, that the conditional expectation is independent of  $N_1(\mathbf{x})$  implies that  $E[D_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq$



**Figure 4:** For the expected difference in the proportion of successes, as a function of  $n$ , three values of  $k$ , and various probabilities of success  $p$ , using the formula provided in Web Appendix E.2.

$\emptyset, J_1(\mathbf{X}) \neq \emptyset$  ] is independent of  $p$ , yielding the result.

■

#### A.4.3 Quantifying the bias for the difference

Figure 4 contains a plot of  $E[\hat{D}_k(\mathbf{X}) \mid I_k(\mathbf{X}) \neq \emptyset, J_k(\mathbf{X}) \neq \emptyset]$  as a function of the number of trials  $n$ , and for  $k = 1, 2$ , and  $3$ . Because the bias is dependent on  $p$  when  $k > 1$ , the difference is plotted for various values of  $p$ . These expected differences are obtained from the formula provided in Web Appendix E.2. The magnitude of the bias is simply the absolute value of the expected difference. As with the bias in the proportion (see Figure 1), the bias in the difference is substantial even when  $n$  is relatively large.

## B Appendix: Size of the bias when the DGP is hot hand/streak shooting

In Section 3.3 the correction to GVT’s estimate of the hot hand effect (and test statistic) is based on the magnitude of the bias under the assumption that the shooter has a fixed probability of success (Bernoulli process). However, if the underlying data generating process (DGP) instead represents hot hand or streak shooting, then the size of the bias changes. While many DGPs can produce hot hand shooting, arguably the most natural are those discussed in Gilovich et al. (1985), as they reflect lay conceptions of the hot hand and streak shooting. While GVT take no particular stance on which lay definition is most appropriate, they do identify hot hand and streak shooting with: (i) “non-stationarity” (the zone, flow, in the groove, in rhythm), and (ii) “positive association” (success breeds success). We label (i) as a *regime shift* model, and interpret it as capturing the idea that a player’s probability of success may increase due to some factor that is unrelated to previous outcomes, so unobservable to the econometrician. This can be modeled naturally as a hidden markov chain over the player’s (hidden) ability state. We label (ii) as a *positive feedback* model, because it can be interpreted as capturing the idea that positive feedback from a player’s previous shot outcomes can affect his/her subsequent probability of success. This can be modeled naturally as an autoregressive process, or equivalently as a markov chain over shot outcomes.<sup>50</sup>

In Figure 5 we plot the bias in the estimator of the difference in probability of success when on a hit streak rather than miss streak,  $\hat{D}_3$ , for three alternative DGPs, each of which admits the Bernoulli process as a special case.<sup>51</sup> The first panel corresponds to the “regime shift” DGP in which the difference in the probability of success between the “hot” state and the “normal” state is given by  $d$  (where  $d = 0$  represents Bernoulli shooting),<sup>52</sup> the second panel corresponds to the “positive feedback” DGP in which hitting (missing) 3 shots in a row increases (decreases) the probability of success by  $d/2$ , and the third panel corresponds to the “positive feedback (for hits)” DGP in which positive feedback operates for hits only, making the probability of success increase by  $d$  whenever

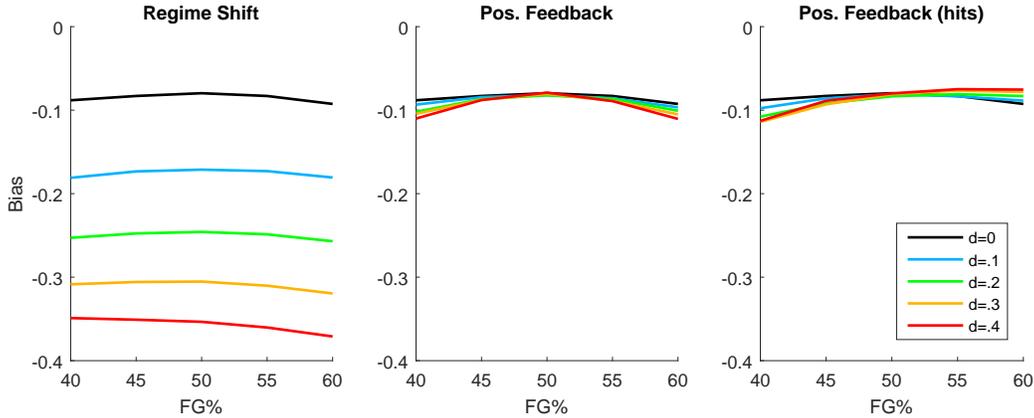
<sup>50</sup>A positive feedback model need not be stationary.

<sup>51</sup>Each point is the output of a simulation with 10,000 repetitions of 100 trials from the DGP.

<sup>52</sup>In particular, let  $Q$  be the hidden markov chain over the “normal” state ( $n$ ) and the “hot” state ( $h$ ), where the probability of success in the normal state is given by  $p_n$ , and the probability of success in the hot state is given by  $p_h$ , with the shift in probability of success given by  $d := p_h - p_n$

$$Q := \begin{pmatrix} q_{nn} & q_{nh} \\ q_{hn} & q_{hh} \end{pmatrix}$$

Where  $q_{nn}$  represents the probability of staying in the “normal” state, and  $q_{nh}$  represents the probability of transitioning from the “normal” to the “hot” state, etc. Letting  $\pi = (\pi_n, \pi_h)$  be the stationary distribution, we find that the magnitude of the bias is not very sensitive to variation in the stationary distribution and transition probabilities within a plausible range (i.e.  $\pi_h \in [.05, .2]$  and  $q_{hh} \in (.8, .98)$ ), while it varies greatly with the difference in probabilities  $d$  and the overall expected shooting percentage  $p = p_n + \pi_h d$ . In the graph, for each  $d$  and  $p$  (FG%), we average across values for the stationary distribution ( $\pi_h$ ) and transition probability ( $q_{hh}$ ).



**Figure 5:** The bias for three types of hot hand and streak shooting data generating processes (DGPs), where  $FG\%$  is the expected overall field goal percentage from the DGP, and  $d$  represents the change in the player’s underlying probability of success. When  $d = 0$  each model reduces to a Bernoulli process. Therefore the black line represents the bias in a Bernoulli process ( $n = 100$  trials,  $k = 3$ ).

3 hits in a row occurs. Within each panel of the figure, the bias, which is the expected difference between  $\hat{D}_3$ , the estimator of the shift in the probability of success, and  $d$ , the true shift in the probability of success, is depicted as a function of the expected overall shooting percentages (from 40 percent to 60 percent), for four true shifts in the underlying probability ( $d \in \{.1, .2, .3, .4\}$ ).<sup>53</sup>

Observe that when the true DGP is a player with a hot hand, the bias is typically more severe, or far more severe, than the bias associated with a Bernoulli DGP. In particular, the bias in the “regime shift” model is particularly severe, which arises from two sources: (i) the bias discussed in Section 2, and (ii) an attenuation bias, due to measurement error, as hitting 3 shots in a row is an imperfect proxy for the “hot state.”<sup>54</sup> The bias in the positive feedback DGP is uniformly below the bias for a Bernoulli shooter. For the DGP in which positive feedback operates only for hits, the bias is stronger than that of Bernoulli shooters for expected shooting percentages below 50 percent (as in GVTs data), and slightly less strong for shooting percentage above 50 percent. As the true DGP is likely some combination of a regime shift and positive feedback, it is reasonable to conclude that the empirical approach in Section 3.3 should be expected to (greatly) understate the true magnitude of any underlying hot hand.

The intuition for why the introduction of regime shift elements increases the strength of the bias so considerably is that if a player’s probability of success is not driven merely by feedback from previous shots, but also by other time-varying player (and environment) specific factors, then the act of hitting consecutive shots will serve as only a noisy proxy of the hot state, resulting in measurement error, and an attenuation bias in the estimate. This type of measurement error is

<sup>53</sup>Results are similar if the DGP instead has negative feedback, i.e.  $d \in \{-.1, -.2, -.3, -.4\}$ .

<sup>54</sup>In practice observers may have more information than the econometrician (e.g. shooting mechanics, perceived confidence, or lack thereof, etc.), so may be subject to less measurement error.

similar to what Stone (2012) uncovered in the relationship between autocorrelation in outcomes and autocorrelation in ability when considering a DGP that contains autocorrelation in ability.

## C Appendix: Additional analyses, and details for Section 3

### C.1 An alternative (pooled) analysis of shooting data

An alternative approach to testing for streak shooting across players is to pool all shots from the “3 hits” and “3 misses” categories (discarding the rest), then use a linear probability model to estimate the effect of a shot falling in the “3-hits” category. If the implementation of GVT’s design met the goal of placing each player in a position in which his or her probability of success is .5, then this approach would be analogous to re-weighting the under-weighted coin flips in Table 1 of Section 1. With 2515 shots, the bias is minimal and the estimate in this case is +17 percentage points ( $p < .01$ ,  $S.E. = 3.7$ ). Because GVT’s design goal is difficult to implement in practice, this approach will introduce an upward bias, due to aggregation, if the probability of success varies across players. Adding fixed effects in a regression will control for this aggregation bias, but strengthens the selection bias related to streaks.<sup>55</sup> As a result, a bias correction is necessary. In this case, the estimated effect is +13.9 percentage points ( $p < .01$ ,  $S.E. = 5.8$ ), which has larger standard errors because the heteroscedasticity under the assumption of different player probabilities necessitates the use of robust variants (in this case, Bell and McCaffrey standard errors, see Imbens and Kolesar [2016]). The magnitude of the estimated effect has a different interpretation than the one given for the estimate of the average difference across players; it should be thought of as the hot hand effect for the average shot rather than the average player. This interpretation arises because pooling shots across players generates an unbalanced panel, which results in the estimate placing greater weight on the players that have taken more shots. As such, in the extreme it is even possible that the majority of players exhibit a tendency to have fewer streaks than expected by chance, yet, because they have generated relatively few observations, their data becomes diluted by many observations from a single streak shooter.

### C.2 Details on the hypothesis testing with the permutation test procedure

Let  $\mathbf{X} \in \{0,1\}^n$  be a sequence of shot outcomes from some player,  $i$ . The null hypothesis is that the shots are i.i.d. with  $\mathbb{P}(X_t = 1) = p^i$ . This implies that conditional on the number of hits,  $N_1(\mathbf{X}) = n_1$ , each rearrangement is equally likely. Considering only sequences for which both  $\hat{P}^i(\text{hit}|k \text{ hits})$  and  $\hat{P}^i(\text{hit}|k \text{ misses})$  are defined, the hot hand hypothesis predicts that the difference  $\hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})$  will be significantly larger than what one would expect by chance. Let  $\hat{D}_k(\mathbf{X})$  be this difference for sequence  $\mathbf{X}$ . For an observed sequence  $\mathbf{x}$ , with  $N_1(\mathbf{x}) = n_1$  hits, to test the null hypothesis at the  $\alpha$  level, one simply checks if  $\hat{D}_k(\mathbf{x}) \geq c_{\alpha, n_1}$ , where the critical

<sup>55</sup>In this panel regression framework, the bias from introducing fixed-effects is an example of an incidental parameter problem of Neyman and Scott (1948), and is essentially equivalent to that discussed in Nerlove (1971) and Nickell (1981), and itself is closely related to the bias in estimates of autocorrelation in time series mentioned in the Introduction.

value  $c_{\alpha, n_1}$  is defined as the smallest  $c$  such that  $\mathbb{P}(D_k(\mathbf{X}) \geq c \mid H_0, N_1(\mathbf{X}) = n_1) \leq \alpha$ , and the distribution  $\mathbb{P}(D_k(\mathbf{X}) \geq c \mid H_0, N_1(\mathbf{X}) = n_1)$  is generated using the formula provided in Web Appendix E.2. For the quantity  $\mathbb{P}(D_k(\mathbf{X}) \geq c \mid H_0, N_1(\mathbf{X}) = n_1)$  it may be the case that for some  $c^*$ , it is strictly greater than  $\alpha$  for  $c \leq c^*$ , and equal to 0 for  $c > c^*$ . In this case, for any sequence with  $N_1(\mathbf{X}) = n_1$  one cannot reject  $H_0$  at an  $\alpha$  level of significance. From the ex ante perspective, a test of the hot hand at the  $\alpha$  level of significance consists of a family of such critical values  $\{c_{\alpha, n_1}\}$ . It follows immediately that  $\mathbb{P}(\text{reject} \mid H_0) \leq \alpha$  because  $\mathbb{P}(\text{reject} \mid H_0) = \sum_{n_1=1}^n \mathbb{P}(D_k(\mathbf{X}) \geq c_{\alpha, n_1} \mid H_0, N_1(\mathbf{X}) = n_1) \mathbb{P}(N_1(\mathbf{X}) = n_1 \mid H_0) \leq \alpha$ . Lastly, for any arbitrary test statistic  $T(\mathbf{X})$ , the fact that the distribution of  $\mathbf{X}$  is *exchangeable* conditional on  $N_1(\mathbf{X}) = n_1$  means that  $\mathbb{P}(T(\mathbf{X}) \geq c \mid H_0, N_1(\mathbf{X}) = n_1)$  can be approximated to appropriate precision with Monte-Carlo permutations of the sequence  $\mathbf{x}$ .

## D Web Appendix: Streak Selection Bias and a Quantitative Comparison to Sampling without Replacement

We show how the downward bias in the estimator  $\hat{P}_k(X)$  is driven by two sources of selection bias. One is related to sampling-without-replacement, and the other to the overlapping nature of streaks.

Recall from the proof of Theorem 1 that  $E \left[ \hat{P}_k(\mathbf{X}) \mid I_k(\mathbf{X}) \neq \emptyset \right] = \mathbb{P}(X_\tau = 1 \mid I_k(\mathbf{X}) \neq \emptyset)$ , where  $\tau$  is drawn (uniformly) at random from  $I_k(\mathbf{X})$ . Because any sequence  $\mathbf{X} \in \{0, 1\}^n$ , such that  $I_k(\mathbf{X}) \neq \emptyset$ , that a researcher encounters will contain a certain number of successes  $N_1(\mathbf{X}) = n_1$  and failures  $n_0 := n - n_1$ , for  $n_1 = k, \dots, n$  we can write  $\mathbb{P}(X_\tau = 1 \mid I_k(\mathbf{X}) \neq \emptyset) = \sum_{n_1=k}^n \mathbb{P}(X_\tau = 1 \mid N_1(\mathbf{X}) = n_1, I_k(\mathbf{X}) \neq \emptyset) \mathbb{P}(N_1(\mathbf{X}) = n_1 \mid I_k(\mathbf{X}) \neq \emptyset)$ . To explore the nature of the downward bias we discuss why  $\mathbb{P}(X_\tau = 1 \mid N_1(\mathbf{X}) = n_1, I_k(\mathbf{X})) < \mathbb{P}(X_t = 1 \mid N_1(\mathbf{X}) = n_1) = n_1/n$ , i.e. why the probability that a randomly drawn trial from  $I_k(\mathbf{X})$  is less than the overall proportion of successes in the sequence  $\hat{p} = n_1/n$ , i.e. the prior probability that a trial is a success when it is drawn (uniformly) at random from  $1, \dots, n$  under the knowledge that  $N_1(\mathbf{X}) = n_1$ .<sup>56</sup>

Suppose that the researcher were to know the overall proportion of successes  $\hat{p} = n_1/n$  in the sequence. Now, consider the following two ways of learning that trial  $t$  immediately follows  $k$  consecutive successes: (i) a trial  $\tau_N$ , drawn uniformly at random from  $\{k+1, \dots, n\}$  is revealed to be trial  $\tau_N = t$ , and preceded by  $k$  consecutive successes, or (ii) a trial  $\tau_I$ , drawn (uniformly) at random from  $I_k(\mathbf{X}) = \{i : \prod_{i=t-k}^{t-1} X_i = 1\} \subseteq \{k+1, \dots, n\}$  is revealed to be trial  $\tau_I = t$ . In each case, the prior probability of success is  $\mathbb{P}(X_t = 1) = n_1/n$ , which can be equivalently represented with the odds ratio  $\mathbb{P}(X_t = 1)/\mathbb{P}(X_t = 0) = n_1/n_0$ , indicates the  $n_1/n_0 : 1$  prior odds in favor of  $X_t = 1$  (relative to  $X_t = 0$ ).

In the first case the probability distribution for  $\tau_N$  is given by  $\mathbb{P}(\tau_N = t) = 1/(n-k)$  for all  $t \in \{k+1, \dots, n\}$ , and is independent of  $\mathbf{X}$ . Upon finding out that  $\tau_N = t$  one then learns that  $\prod_{i=t-k}^{t-1} X_i = 1$ . As a result, the posterior odds can be represented by a sampling-without-replacement formula, via Bayes rule:

<sup>56</sup>Note that  $\mathbb{P}(N_1(\mathbf{X}) = n_1 \mid I_k(\mathbf{X}) \neq \emptyset) > \mathbb{P}(N_1(\mathbf{X}) = n_1)$  because the exclusion of sequences without a streak of  $k$  successes in the first  $n-1$  trials biases upwards the number of successes. We do not consider this upward bias here as Theorem 1 shows that the downward biases predominate.

$$\begin{aligned}
\frac{\mathbb{P}(X_t = 1 | \tau_N = t)}{\mathbb{P}(X_t = 0 | \tau_N = t)} &= \frac{\mathbb{P}(X_t = 1, \prod_{i=t-k}^{t-1} X_i = 1 | \tau_N = t)}{\mathbb{P}(X_t = 0, \prod_{i=t-k}^{t-1} X_i = 1 | \tau_N = t)} \\
&= \frac{\mathbb{P}(\tau_N = t | X_t = 1, \prod_{i=t-k}^{t-1} X_i = 1) \mathbb{P}(\prod_{i=t-k}^{t-1} X_i = 1 | X_t = 1) \mathbb{P}(X_t = 1)}{\mathbb{P}(\tau_N = t | X_t = 0, \prod_{i=t-k}^{t-1} X_i = 1) \mathbb{P}(\prod_{i=t-k}^{t-1} X_i = 1 | X_t = 0) \mathbb{P}(X_t = 0)} \\
&= \frac{\mathbb{P}(\prod_{i=t-k}^{t-1} X_i = 1 | X_t = 1) \mathbb{P}(X_t = 1)}{\mathbb{P}(\prod_{i=t-k}^{t-1} X_i = 1 | X_t = 0) \mathbb{P}(X_t = 0)} \\
&= \frac{\frac{n_1-1}{n-1} \times \dots \times \frac{n_1-k}{n-k} n_1}{\frac{n_1}{n-1} \times \dots \times \frac{n_1-k+1}{n-k} n_0} \\
&= \frac{n_1 - k}{n_1} \frac{n_1}{n_0} \\
&= \frac{n_1 - k}{n_0}
\end{aligned}$$

Observe that the prior odds in favor of success are attenuated by the likelihood ratio  $\frac{n_1-k}{n_1}$  of producing  $k$  consecutive successes given either hypothetical state of the world:  $X_t = 1$  or  $X_t = 0$ , respectively. That this is a sampling-without-replacement effect can be made most transparent by re-expressing the posterior odds as  $\frac{n_1-k}{n-k} / \frac{n_0}{n-k}$ .<sup>57,58</sup>

In the second case, the probability that  $\tau_I = t$  is drawn from  $I_k(\mathbf{X})$  is completely determined by  $M := |I_k(\mathbf{X})|$ , and equal to  $1/M$ . Upon learning that  $\tau_I = t$  one can infer the following three things: (i)  $I_k(\mathbf{X}) \neq \emptyset$ , i.e.  $M \geq 1$ , which is informative if  $n_1 \leq (k-1)(n-n_1) + k$ , (ii)  $t$  is a member of  $I_k(\mathbf{X})$ , and (iii)  $\prod_{i=t-k}^{t-1} X_i = 1$ , as in sampling-without-replacement. As a result, the

<sup>57</sup>The numerator is the probability of drawing a 1 at random from an urn containing  $n_1$  1's and  $n_0$  0's, once  $k$  1's (and no 0's) have been removed from the urn. The denominator is the probability of drawing a 0 from the same urn.

<sup>58</sup>This effect calls to mind the key behavioral assumption made in Rabin (2002), that believers in the law of small numbers view outcomes from an i.i.d. process as if they were instead generated by random draws without replacement.

posterior odds can be determined via Bayes Rule in the following way:

$$\begin{aligned}
\frac{\mathbb{P}(X_t = 1|\tau_I = t)}{\mathbb{P}(X_t = 0|\tau_I = t)} &= \frac{\mathbb{P}(X_t = 1, \prod_{t-k}^{t-1} X_i = 1, M \geq 1|\tau_I = t)}{\mathbb{P}(X_t = 0, \prod_{t-k}^{t-1} X_i = 1, M \geq 1|\tau_I = t)} \\
&= \frac{\mathbb{P}(\tau_I = t|X_t = 1, \prod_{t-k}^{t-1} X_i = 1, M \geq 1) \mathbb{P}(X_t = 1, \prod_{t-k}^{t-1} X_i = 1, M \geq 1)}{\mathbb{P}(\tau_I = t|X_t = 0, \prod_{t-k}^{t-1} X_i = 1, M \geq 1) \mathbb{P}(X_t = 0, \prod_{t-k}^{t-1} X_i = 1, M \geq 1)} \\
&= \frac{\mathbb{P}(\tau_I = t|X_t = 1, \prod_{t-k}^{t-1} X_i = 1) \mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 1) \mathbb{P}(X_t = 1)}{\mathbb{P}(\tau_I = t|X_t = 0, \prod_{t-k}^{t-1} X_i = 1) \mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 0) \mathbb{P}(X_t = 0)} \tag{17}
\end{aligned}$$

$$\begin{aligned}
&= \frac{E \left[ \frac{1}{M} \mid \prod_{t-k}^{t-1} X_i = 1, X_t = 1 \right] \mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 1) \mathbb{P}(X_t = 1)}{E \left[ \frac{1}{M} \mid \prod_{t-k}^{t-1} X_i = 1, X_t = 0 \right] \mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 0) \mathbb{P}(X_t = 0)} \\
&= \frac{E \left[ \frac{1}{M} \mid \prod_{t-k}^{t-1} X_i = 1, X_t = 1 \right]}{E \left[ \frac{1}{M} \mid \prod_{t-k}^{t-1} X_i = 1, X_t = 0 \right]} \frac{n_1 - k}{n_1} \frac{n_1}{n_0} \tag{18}
\end{aligned}$$

For first term in (17) the event  $M \geq 1$  is dropped from the conditional argument because it is implied by the event  $\prod_{t-k}^{t-1} X_i = 1$ , and the term  $\frac{\mathbb{P}(M \geq 1|X_t=1, \prod_{t-k}^{t-1} X_i=1)}{\mathbb{P}(M \geq 1|X_t=0, \prod_{t-k}^{t-1} X_i=1)}$  does not appear because it is equal to 1.

Equation 18 gives the posterior odds  $\frac{\mathbb{P}(X_t=1|\tau_I=t)}{\mathbb{P}(X_t=0|\tau_I=t)}$  in favor of observing  $X_t = 1$  (relative to  $X_t = 0$ ), for a representative trial  $\tau = t$  drawn at random from  $I_k(\mathbf{X})$ . Observe that the prior odds ratio  $n_1/n_0$  is multiplied by two separate updating factors, which we now discuss.

The first updating factor  $\frac{n_1-k}{n_1}$  is clearly strictly less than one and reflects the restriction that the finite number of available successes places on the procedure for selecting trials into  $I_k(\mathbf{X})$ . In particular, it can be thought of as the information provided upon learning that  $k$  of the  $n_1$  successes are no longer available, which leads to a sampling-without-replacement effect on the prior odds  $n_1/n_0$ . Clearly, the attenuation in the odds due to this factor increases in the streak length  $k$ .

The second updating factor  $\frac{E \left[ \frac{1}{M} \mid \prod_{t-k}^{t-1} X_i=1, X_t=1 \right]}{E \left[ \frac{1}{M} \mid \prod_{t-k}^{t-1} X_i=1, X_t=0 \right]} < 1$ , for  $t < n$ , reflects an additional restriction that the arrangement of successes and failures in the sequence places on the procedure for selecting trials into  $I_k(\mathbf{X})$ . It can be thought of as the additional information provided by learning that the  $k$  successes, which are no longer available, are consecutive and immediately precede  $t$ . To see why the odds are further attenuated in this case, we begin with the random variable  $M$ , which is defined as the number of trials in  $I_k(\mathbf{X})$ . The probability of any particular trial  $t \in I_k(\mathbf{X})$  being selected at random is  $1/M$ . Now, because the expectation in the numerator conditions on  $X_t = 1$ , this means that  $1/M$  is expected to be smaller in the numerator than in the denominator, where the expectation instead conditions on  $X_t = 0$ . The reason why is that for a sequence in which  $X_t = 1$  the streak of 1's continues on, meaning that trial  $t + 1$  must also be in  $I_k(\mathbf{X})$ , and trials  $t + 2$  through  $t + k$  each may also be in  $I_k(\mathbf{X})$ . By contrast, for a sequence in which  $X_t = 0$  the

streak of 1's ends, meaning that trials  $t + 1$  through  $t + k$  cannot possibly be in  $I_k(\mathbf{X})$ , which leads the corresponding  $1/M$  to be smaller in expectation.<sup>59</sup> This last argument provides intuition for why the attenuation of the odds due to this factor increases in  $k$ .

Interestingly, for the special case of  $k = 1$ ,  $\frac{E\left[\frac{1}{M} \mid x_{t-1}=1, x_t=1\right]}{E\left[\frac{1}{M} \mid x_{t-1}=1, x_t=0\right]} = 1 - \frac{1}{(n-1)(n_1-1)} < 1$  when  $t < n$ , and  $\frac{E\left[\frac{1}{M} \mid x_{n-1}=1, x_n=1\right]}{E\left[\frac{1}{M} \mid x_{n-1}=1, x_n=0\right]} = \frac{n_1}{n_1-1} > 1$  when  $t = n$ .<sup>60</sup> These contrasting effects combine to yield the familiar sampling-without-replacement formula:

$$E\left[\hat{P}_1(\mathbf{X}) \mid I_1(\mathbf{X}) \neq \emptyset, N_1(\mathbf{X}) = n_1\right] = \frac{n_1 - 1}{n - 1} \quad (19)$$

as demonstrated in Lemma 1, in Appendix A.3. On the other hand, when  $k > 1$  the bias is substantially stronger than sampling-without-replacement (see Figure 6), though the formula does not appear to admit a simple representation. For further discussion on the relationship between the bias, sampling-without-replacement, and the *overlapping words paradox* (Guibas and Odlyzko 1981) see Web Appendix F.

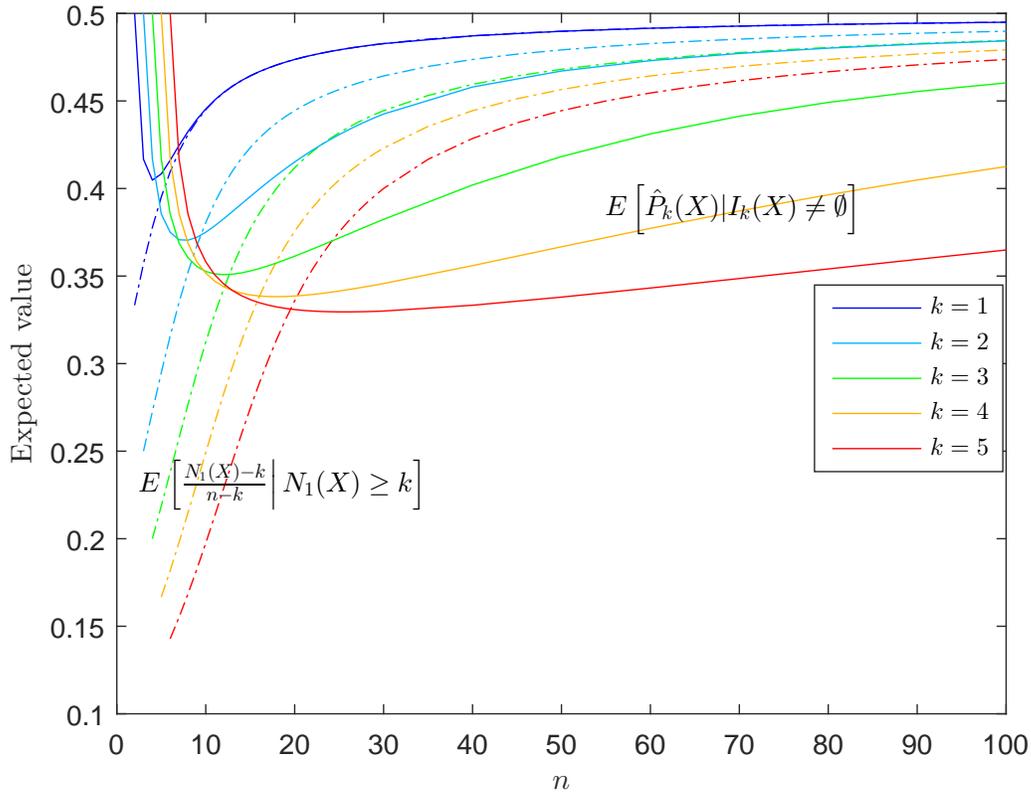
#### *A quantitative comparison with sampling-without-replacement*

For the general case, in which  $\hat{p} = n_1/n$  is unknown, juxtaposing the bias with sampling-without-replacement puts the magnitude of the bias into context. Let the probability of success be given by  $p = \mathbb{P}(X_t = 1)$ . In Figure 6, the expected empirical probability that a randomly drawn trial in  $I_k(\mathbf{X})$  is a success, which is the expected proportion,  $E[\hat{P}_k(\mathbf{X}) \mid I_k(\mathbf{X}) \neq \emptyset]$ , is plotted along with the expected value of the probability that a randomly drawn trial  $t \in \{1, \dots, n\} \setminus T_k$  is a success, given that the  $k$  success trials  $T_k \subseteq \{1, \dots, n\}$  have already been drawn from the sequence (sampling-without-replacement),  $E\left[\frac{N_1(\mathbf{X})-k}{n-k} \mid N_1(\mathbf{X}) \geq k\right]$ . The plot is generated using the combinatorial results discussed in Section 2.1. Note that in the case of  $k = 1$ , the bias is identical to sampling-without-replacement, as shown in Equation 19.<sup>61</sup> Observe that for  $k > 1$ , and  $n$  not too small, the

<sup>59</sup>This is under the assumption that  $t \leq n - k$ . In general, the event  $X_t = 0$  excludes the next  $\min\{k, n - t\}$  trials from  $t + 1$  to  $\min\{t + k, n\}$  from being selected, while the event  $X_t = 1$  leads trial  $t + 1$  to be selected, and does not exclude the next  $\min\{k, n - t\} - 1$  trials from being selected.

<sup>60</sup>The likelihood ratios can be derived following the proof of Lemma 1 in Appendix A.3. In particular, for the equivalent likelihood ratio,  $\frac{\mathbb{P}(\tau=t \mid x_{t-1}=1, x_t=1)}{\mathbb{P}(\tau=t \mid x_{t-1}=1, x_t=0)}$ , the approach used to derive the numerator can also be used to show that the denominator is equal to  $\frac{1}{n-2} \left( \frac{n_0-1}{n_1} + \frac{n_1-1}{n_1-1} \right)$ . Further, in the case of  $t = n$ , it is clear that  $\mathbb{P}(\tau = n \mid x_{n-1} = 1, x_n = 0) = \frac{1}{n_1}$ . Each likelihood ratio then follows from dividing and collecting terms.

<sup>61</sup>This appears to contradict equation 18, i.e. that the bias in the procedure used to select the subset of trials  $I_{1k}(\mathbf{X})$ , is *stronger* than sampling-without-replacement for  $t < n$ , whereas it is non-existent (thus weaker) for  $t = n$ . This disparity is due to the second updating factor, which relates to the arrangement. It turns out that for  $k = 1$ , the determining aspect of the arrangement that influences this updating factor is whether or not the final trial is a success, as this determines the number of successes in the first  $n - 1$  trials, where  $M = n_1 - X_n$ . If one were to instead fix  $M$  rather than  $n_1$ , then sampling-without-replacement relative to the number of successes in the first  $n - 1$  trials would be an accurate description of the mechanism behind the bias, and it induces a negative dependence between any two trials within the first  $n - 1$  trials of the sequence. Therefore, it is sampling-without-replacement with respect to  $M$  that determines the bias when  $k = 1$ .



**Figure 6:** The dotted lines correspond to the bias from sampling-without-replacement. It is the expected probability of a success, given that  $k$  successes are first removed from the sequence (assuming  $p = .5$ ). The solid lines correspond to the expected proportion from Figure 1.

bias in the expected proportion is considerably larger than the corresponding bias from sampling-without-replacement.

## E Web Appendix: The Formula Used to Generate the Sampling Distribution and Calculate Expectations

We describe the formula used to build the exact sampling distribution of the proportion, and difference in proportions, from which we calculate expectations and plot histograms.

### E.1 Proportion

Given  $n$  trials and streaks of length  $k$ , we observe that the proportion of successes on the trials that immediately follow  $k$  consecutive success  $\hat{P}_k(\mathbf{x})$  can be represented simply as the the number of successes on trials that immediately follow a streak of  $k$  consecutive successes divided by the total number of trials—i.e. failures and success—that immediately follow a streak of  $k$  consecutive successes. In particular, for a sequence  $\mathbf{x} \in \{0, 1\}^n$  of successes and failures, we have:

$$\hat{P}_k(\mathbf{x}) = \frac{M^1(\mathbf{x})}{M^0(\mathbf{x}) + M^1(\mathbf{x})}$$

where  $M^0(\mathbf{x}) := |\{i \in \{k + 1, \dots, n\} : (1 - x_i) \prod_{j=i-k}^{i-1} x_j = 1\}|$  is the number of failures that immediately follow  $k$  consecutive successes (suppressing the  $k$  to ease notation). Similarly, the number of successes that immediately follow  $k$  consecutive successes is defined as  $M^1(\mathbf{x}) := |\{i \in \{k + 1, \dots, n\} : x_i \prod_{j=i-k}^{i-1} x_j = 1\}|$ . Finally, the expected value of  $\hat{P}_k(\mathbf{x})$  is uniquely determined by the joint distribution of counts  $\mathbb{P}( (M^0(\mathbf{X}), M^1(\mathbf{X})) = (m^0, m^1) )$ .

The algorithm described below (recursively) constructs the exact joint distribution of counts, by associating each unique count realization, which we call a *key*, with its corresponding probability.<sup>62</sup> In general, for a sequence of length  $n$  and a streak of length  $k$  this joint distribution can be represented as a *dictionary* of (key:probability) pairs  $D := (\mathbf{m} : p_D(\mathbf{m}))_{\mathbf{m} \in D_c}$ , where  $\mathbf{m} := (m^0, m^1)$  is a unique pair,  $D_c$  corresponds to the set of count realizations with non-zero probability, i.e.

$$D_c := \{\mathbf{m} \in \mathbb{N}^2 \mid p_D(\mathbf{m}) > 0\}$$

and  $p_D(\mathbf{m}) := \mathbb{P}( (M^0(\mathbf{X}), M^1(\mathbf{X})) = (m^0, m^1) )$ .

Table 3 reports the distribution over the sample space of sequences, and the corresponding dictionary, for the simple case of  $n = 3$  and  $k = 1$ . From the dictionary one can derive the sampling

<sup>62</sup>This algorithm, which builds upon an algorithm suggested by Michael J. Wiener, replaces an exact formula based on the joint distribution of runs of various lengths that we derived in a previous working paper version of this manuscript. The previous formula, while numerically tractable, was less efficient.

**Table 3:** In the table to the left column one lists the sample space of eight possible sequence realizations from three trials. Column two lists the number of (failures, successes) that immediately follow a success. The third column lists the probability with which the sequence occurs, where  $p$  is the probability of success and  $q$  is the probability of failure. In the table to the right the joint distribution is represented as a dictionary of count-probability pairs. Each unique count  $\mathbf{m} = (m^0, m^1)$  has a unique associated probability equal to the sum of the probabilities of all sequences with the same associated count (see the table on the left).

<i>Sample Space of sequences</i>			<i>Dictionary</i>	
sequence	probability	count	count	probability
000	$q^3$	(0, 0)	$\mathbf{m}$	$p_D(\mathbf{m})$
001	$q^2p$	(0, 0)	(0, 0)	$q^2$
010	$q^2p$	(1, 0)	(1, 0)	$(q + q^2)p$
100	$q^2p$	(1, 0)	(0, 1)	$qp^2$
011	$qp^2$	(0, 1)	(1, 1)	$qp^2$
101	$qp^2$	(1, 0)	(0, 2)	$p^3$
110	$qp^2$	(1, 1)		
111	$p^3$	(0, 2)		

distribution of the proportion and directly compute the expected proportion:

$$E[\hat{P}_k(\mathbf{x}) | I_k(\mathbf{x}) \neq \emptyset] = \sum_{\mathbf{m} \in D_c^*} \frac{m^1}{m^0 + m^1} p_D^*(\mathbf{m})$$

where  $D_c^* = D_c \setminus \{(0, 0)\}$  and  $p_D^*(\mathbf{m}) := p_D(\mathbf{m}) / \sum_{\mathbf{m}' \in D_c^*} p_D(\mathbf{m}')$

Let  $D(\ell, r)$  be the dictionary that represents the count-probability pairs for the remaining  $r$  trials of a sequence that has  $\ell \leq k$  consecutive successes immediately preceding the current trial. For example, if  $k = 2$  then  $D(0, 0) = D(1, 0) = D(2, 0) = ((0, 0) : 1)$ , as when zero trials remain in the sequence the only count possible is (0,0), which occurs with probability 1. Also note that  $D(1, 1) = ((0, 0) : 1)$ ,  $D(2, 1) = ((1, 0) : q, (0, 1) : p)$ , and  $D(2, 2) = ((1, 0) : q, (1, 1) : pq, (0, 2) : p^2)$ , as a trial can only be counted as a fail or success if it is immediately preceded by  $\ell = k = 2$  consecutive successes. The key observation is that given the initial condition  $D(\ell, 0) = ((0, 0) : 1)$  for  $0 \leq \ell \leq k$ , the dictionaries  $D(\ell, r)$  can be defined recursively for  $r > 0$  and  $0 \leq \ell \leq k$ , and take the following form:

$$D(\ell, r) = \begin{cases} D(0, r-1)^{(0,0):q} \uplus D(\ell+1, r-1)^{(0,0):p}, & \text{if } \ell < k \\ D(0, r-1)^{(1,0):q} \uplus D(k, r-1)^{(0,1):p}, & \text{if } \ell = k \end{cases}$$

where: (i) the operation  $D^{\mathbf{m}':p'} := (\mathbf{m} + \mathbf{m}' : p_D(\mathbf{m}) \times p')$  increments each count  $\mathbf{m}$  with the addition of  $\mathbf{m}'$ , and scales its corresponding probability  $p_D(\mathbf{m})$  by the probability  $p'$  of the incre-

ment, and (ii) given the dictionaries  $A$  and  $B$ , the operation  $A \uplus B := (\mathbf{m} : (p_A + p_B)(\mathbf{m}))_{\mathbf{m} \in A_c \cup B_c}$  defines the union of two dictionaries as the union of their counts, where the corresponding probabilities for a key that appears in both dictionaries are summed together (we assume that  $p_A(\mathbf{m}) = 0$  for  $\mathbf{m} \notin A_c$ ; also for  $B$ ). If a trial is immediately preceded by  $\ell < k$  consecutive successes, then with probability  $q$  ( $p$ ) the next trial to its right will be immediately preceded by 0 ( $\ell + 1$ ) consecutive successes; regardless of the outcome of the trial,  $\mathbf{m}' = (0, 0)$  additional failures and successes will be counted as immediately preceded by  $k$  successes and  $r - 1$  trials will remain. If, on the other hand, a trial is immediately preceded by  $\ell = k$  consecutive successes (and there is at least one trial remaining, i.e.  $r > 0$ ), then with probability  $q$  ( $p$ ) the next trial to its right will be immediately preceded by 0 ( $k$ ) consecutive successes and we will count  $\mathbf{m}' = (1, 0)$  ( $(0, 1)$ ) additional failures and successes; regardless of the outcome of the trial,  $r - 1$  trials will remain.

The algorithm that follows describes the complete recursive procedure.

---

**Algorithm 1:** Recursive formula that builds the collection of dictionaries  $D$ . Of interest are the dictionaries  $D(0, n)$  for  $n = k + 1, \dots, N$  which correspond to the joint distribution of the total number of (successes, failures) that immediately follow  $k$  consecutive successes in  $n$  trials.

---

```

1 Function Count_Distribution( $N, k, p$ ):
   | /* For the definition of  $D(\ell, r)$ ,  $A^{\mathbf{m}':p'}$  and  $A \uplus B$  below, see text.          */
2   |  $q \leftarrow 1 - p$ 
3   | for  $n = 0, \dots, N$  do
4   |   |  $L \leftarrow \min\{k, n\}$ 
5   |   | for  $\ell = L, \dots, 0$  do
6   |   |   |  $r \leftarrow n - \ell$ 
7   |   |   | if  $r = 0$  then
8   |   |   |   |  $D(\ell, r) \leftarrow ((0, 0) : 1)$ 
9   |   |   | else if  $r > 0$  then
10  |   |   |   | if  $\ell < k$  then
11  |   |   |   |   |  $D(\ell, r) \leftarrow D(0, r - 1)^{(0,0):q} \uplus D(\ell + 1, r - 1)^{(0,0):p}$ 
12  |   |   |   | else if  $\ell = k$  then
13  |   |   |   |   |  $D(\ell, r) \leftarrow D(0, r - 1)^{(1,0):q} \uplus D(k, r - 1)^{(0,1):p}$ 
14  |   |   | end
15  |   | end
16  | return  $D$ 

```

---

## E.2 Difference in Proportions

The difference in proportions can be computed from a dictionary  $D := (\mathbf{m} : p_D(\mathbf{m}))_{\mathbf{m} \in D_c}$ , where  $D_c$  corresponds to the set of count realizations with non-zero probability i.e.

$$D_c := \{\mathbf{m} \in \mathbb{N}^4 \mid p_D(\mathbf{m}) > 0\}$$

and  $p_D(\mathbf{m}) := \mathbb{P}( (M_0^0(\mathbf{X}), M_0^1(\mathbf{X}), M_1^0(\mathbf{X}), M_1^1(\mathbf{X})) = (m_0^0, m_0^1, m_1^0, m_1^1) )$ . The variables  $M_1^0(\mathbf{X})$  and  $M_1^1(\mathbf{X})$  yield the total number of failures and successes (respectively) on those trials that immediately follow a streak of  $k$  successes, whereas  $M_0^0(\mathbf{X})$  and  $M_0^1(\mathbf{X})$  yield the total number of failures and successes (respectively) on those trials that immediately follow a streak of  $k$  failures.

Let  $D(\ell_0, \ell_1, r)$  be the dictionary that represents the count-probability pairs for the remaining  $r$  trials of a sequence in which there are  $\ell_0 \leq k$  consecutive failures and  $\ell_1 \leq k$  consecutive successes on the immediately preceding trials (so that  $\ell_0 \ell_1 = 0$ ). These dictionaries can be constructed recursively in a way similar to that shown in Web Appendix E.1:

$$D(\ell_0, \ell_1, r) = \begin{cases} D(\ell_0 + 1, 0, r - 1)^{(0,0,0,0):q} \uplus D(0, \ell_1 + 1, r - 1)^{(0,0,0,0):p}, & \text{if } \max\{\ell_0, \ell_1\} < k \\ D(k, 0, r - 1)^{(1,0,0,0):q} \uplus D(0, 1, r - 1)^{(0,1,0,0):p}, & \text{if } \ell_0 = k \\ D(1, 0, r - 1)^{(0,0,1,0):q} \uplus D(0, k, r - 1)^{(0,0,0,1):p}, & \text{if } \ell_1 = k \end{cases}$$

See supplementary materials for the corresponding code.

## F Web Appendix: The relationship between the streak selection bias and known biases and paradoxes

### F.1 Sampling-without-replacement and the bias for streaks of length $k = 1$ .

A brief inspection of Table 1 in Section 1 reveals how the dependence between the first  $n - 1$  flips in the sequence arises. In particular, when the coin is flipped three times, the number of Hs in the first 2 flips determines the number of observations of flips that immediately follow an H. Because TT must be excluded, the first two flips will consist of one of three equally likely sequences: HT, TH or HH. For the two sequences with a single H—HT and TH—if a researcher were to find an H within the first two flips of the sequence and then select the adjacent flip for inspection, the probability of heads on the adjacent flip would be 0, which is strictly less than the overall proportion of heads in the sequence. This can be thought of as a sampling-without-replacement effect. More generally, across the three sequences, HT, TH, and HH, the expected probability of the adjacent flip being a heads is  $(0 + 0 + 1)/3 = 1/3$ . This probability reveals the (negative) sequential dependence that exists between the first two flips of the sequence. Further, the same negative dependence holds for *any two flips* in the first  $n - 1$  flips of a sequence of length  $n$ , *regardless of their positions*. Thus, when  $k = 1$  it is neither time’s arrow nor the arrangement of flips within the sequence that determines the bias.

This same sampling-without-replacement feature also underlies a classic form of selection bias known as Berkson’s bias (aka Berkson’s paradox). Berkson (1946) presented a hypothetical study of the relationship between two diseases that, while not associated in the general population, become negatively associated in the population of hospitalized patients. The cause of the bias is subtle: patients are hospitalized only if they have *at least one* of the two particular diseases. To illustrate, assume that someone from the general population has a given disease (Y=“Yes”) or does not (N=“No”), with equal chances. Just as in the coin flip example, anyone with neither disease (NN) is excluded, while a patient within the hospital population must have one of the three equally likely profiles: YN, NY, or YY. Thus, just as with the coin flips, the probability of a patient having another disease, given that he already has one disease, is  $1/3$ .

The same sampling-without replacement feature again arises in several classic conditional probability paradoxes. For example, in the Monty Hall problem the game show host inspects two doors, which can together be represented as one of three equally likely sequences GC, CG, or GG (G=“Goat,” C=“Car”), then opens one of the G doors from the realized sequence. Thus, the host effectively samples G without replacement (Nalebuff 1987; Selvin 1975; Vos Savant 1990).<sup>63</sup>

<sup>63</sup>The same structure also appears in what is known as the boy-or-girl paradox (Miller and Sanjurjo 2015a). A slight modification of the Monty-Hall problem makes it identical to the coin flip bias presented in Table 1 (see Miller and Sanjurjo [2015a]).

Sampling-without-replacement also underlies a well-known finite sample bias that arises in standard estimates of autocorrelation in time series data (Shaman and Stine 1988; Yule 1926). This interpretation of finite sample bias, which does not appear to have been previously noted, allows one to see how this bias is closely related to those above. To illustrate, let  $\mathbf{x}$  be a randomly generated sequence consisting of  $n$  trials, each of which is an i.i.d. draw from some continuous distribution with finite mean and variance. For a researcher to compute the autocorrelation she must first determine its sample mean  $\bar{x}$  and variance  $\hat{\sigma}^2(\mathbf{x})$ , then calculate the autocorrelation  $\hat{\rho}_{t,t+1}(\mathbf{x}) = \hat{c}ov_{t,t+1}(\mathbf{x})/\hat{\sigma}^2(\mathbf{x})$ , where  $\hat{c}ov_{t,t+1}(\mathbf{x})$  is the autocovariance.<sup>64</sup> The total sum of values  $n\bar{x}$  in a sequence serves as the analogue to the number of Hs (or Gs/Ys) in a sequence in the examples given above. Given  $n\bar{x}$ , the autocovariance can be represented as the expected outcome from a procedure in which one draws (at random) one of the  $n$  trial outcomes  $x_i$ , and then takes the product of its difference from the mean ( $x_i - \bar{x}$ ), and another trial outcome  $j$ 's difference from the mean. Because the outcome's value  $x_i$  is essentially drawn from  $n\bar{x}$ , without replacement, the available sum total ( $n\bar{x} - x_i$ ) is averaged across the remaining  $n - 1$  outcomes, which implies that the expected value of another outcome  $j$ 's ( $j \neq i$ ) difference from the mean is given by  $E[x_j|x_i, \bar{x}] - \bar{x} = (n\bar{x} - x_i)/(n - 1) - \bar{x} = (\bar{x} - x_i)/(n - 1)$ . Therefore, given  $x_i - \bar{x}$ , the expected value of the product  $(x_i - \bar{x})(x_j - \bar{x})$  must equal  $(x_i - \bar{x})(\bar{x} - x_i)/(n - 1) = -(x_i - \bar{x})^2/(n - 1)$ , which is independent of  $j$ . Because  $x_i$  and  $j$  were selected at random, this implies that the expected autocorrelation, given  $\bar{x}$  and  $\hat{\sigma}^2(\mathbf{x})$ , is equal to  $-1/(n - 1)$  for all  $\bar{x}$  and  $\hat{\sigma}^2(\mathbf{x})$ . This result accords with known results on the  $O(1/n)$  bias in discrete-time autoregressive processes (Shaman and Stine 1988; Yule 1926), and happens to be identical to the result in Theorem 4 for the expected difference in proportions (see Appendix A.3). In the context of time series regression this bias is known as the *Hurwicz bias* (Hurwicz 1950), which is exacerbated when one introduces fixed effects into a time series model with a small number of time periods (Nerlove 1967, 1971; Nickell 1981).<sup>65,66</sup>

<sup>64</sup>The autocovariance is given by  $\hat{c}ov_{t,t+1}(\mathbf{x}) := \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - \bar{x})(x_{i+1} - \bar{x})$ .

<sup>65</sup>The bias that is exacerbated by the introduction of exogenous variables is commonly known as the "Nickell bias," which was first explored by simulation by Nerlove (1967, 1971). It is an example of what is known as the *incidental parameter problem* (Lancaster 2000; Neyman and Scott 1948).

<sup>66</sup>In a comment on this paper, Rinott and Bar-Hillel (2015) assert that the work of Bai (1975) (and references therein) demonstrate that the bias in the proportion of successes on the trials that immediately follow a streak of  $k$  or more successes follows directly from known results on the finite sample bias of Maximum Likelihood estimators of transition probabilities in Markov chains, as independent Bernoulli trials can be represented by a Markov chain with each state defined by the sequence of outcomes in the previous  $k$  trials. While it is true that the MLE of the corresponding transition matrix is biased, and correct to note the relationship in this sense, the cited theorems do not indicate the direction of the bias, and in any event do not directly apply in the present case because they require that transition probabilities in different rows of the transition matrix not be functions of each other, and not be equal to zero, a requirement which does not hold in the corresponding transition matrix. Instead, an unbiased estimator of each transition probability will exist, and will be a function of the overall proportion.

## F.2 Pattern overlap and the bias for streaks of length $k > 1$ .

In Figure 6 of Web Appendix D we compare the magnitude of the bias in the (conditional) expected proportion to the pure sampling-without-replacement bias, in a sequence of length  $n$ . As can be seen, the magnitude of the bias in the expected proportion is nearly identical to that of sampling-without-replacement for  $k = 1$ . However, for the bias in the expected proportion, the relatively stronger sampling-without-replacement effect that operates within the first  $n - 1$  terms of the sequence is balanced by the absence of bias for the final term.<sup>67</sup> On the other hand, for  $k > 1$  the bias in the expected proportion is considerably stronger than the pure sampling-without-replacement bias. One intuition for this is provided in the discussion of the updating factor in Section D. Here we discuss another intuition, which has to do with the overlapping nature of the selection criterion when  $k > 1$ , which is related to what is known as the *overlapping words paradox* (Guibas and Odlyzko 1981).

For simplicity, assume that a sequence is generated by  $n = 5$  flips of a fair coin. For the simple case in which streaks have length  $k = 1$ , the number of flips that immediately follow a heads is equal to the number of instances of H in the first  $n - 1 = 4$  flips. For any given number of Hs in the first four flips, say three, if one were to sample an H from the sequence and then examine an adjacent flip (within the first four flips), then because any H could have been sampled, across all sequences with three Hs in the first four flips, any H appearing within the first four flips is given equal weight regardless of the sequence in which it appears. The exchangeability of outcomes across equally weighted sequences with an H in the sampled position (and three Hs overall) therefore implies that for any other flip in the first four flips of the sequence, the probability of an H is equal to  $\frac{3-1}{4-1} = \frac{2}{3}$ , regardless of whether or not it is an adjacent flip. On the other hand, for the case of streaks of length  $k = 2$ , the number of opportunities to observe a flip that immediately follows two consecutive heads is equal to the number of instances of HH in the first 4 flips. Because the pattern HH can overlap with itself, whereas the pattern H cannot, then for a sequence with three Hs, if one were to sample an HH from the sequence and examine an adjacent flip within the first 4 flips, it is not the case that any two of the Hs from the sequence can be sampled. For example, in the sequence HHTH only the first two Hs can be sampled. Because the sequences HHTH and HTHH each generate just one opportunity to sample, this implies that the single instance of HH within each of these sequences is weighted twice as much as any of the two (overlapping) instances of HH within the two sequences HHHT and THHH that each allow two opportunities to sample, despite the fact that each sequence has three heads in the first four flips. This implies that, unlike in the case of  $k = 1$ , when sampling an instance of HH from a sequence with three heads in the first four flips, the remaining outcomes H and T are no longer exchangeable, as the arrangements HHTH

<sup>67</sup>The reason for this is provided in the alternative proof of Lemma 1 in Appendix A.3

and HTHH, in which every adjacent flip within the first four flips is a tails, must be given greater weight than the arrangements HHHT and THHH, in which half of the adjacent flips are heads.

This consequence of pattern overlap is closely related to the *overlapping words paradox*, which states that for a sequence (string) of finite length  $n$ , the probability that a pattern (word) appears, e.g.  $\_HTTHH\_$ , depends not only on the length of the pattern relative to the length of the sequence, but also on how the pattern *overlaps* with itself (Guibas and Odlyzko 1981).<sup>68</sup> For example, while the expected number of (potentially overlapping) occurrences of a particular two flip pattern—TT, HT, TH or HH—in a sequence of four flips of a fair coin does not depend on the pattern, it’s probability of occurrence does.<sup>69</sup> The pattern HH can overlap with itself, so can have up to three occurrences in a single sequence (HHHH), whereas the pattern HT cannot overlap with itself, so can have at most two occurrences (HTHT). Because the expected number of occurrences of each pattern must be equal, this implies that the pattern HT is distributed across more sequences, meaning that any given sequence is more likely to contain this pattern.<sup>70</sup>

---

<sup>68</sup>For a simpler treatment which studies a manifestation of the paradox in the non-transitive game known as “Penney’s” game, see Konold (1995) and Nickerson (2007).

<sup>69</sup>That all fixed length patterns are equally likely ex-ante is straightforward to demonstrate. For a given pattern of heads and tails of length  $\ell$ ,  $(y_1, \dots, y_\ell)$ , the expected number of occurrences of this pattern satisfies  $E[\sum_{i=\ell}^n 1_{[(X_{i-\ell+1}, \dots, X_i)=(y_1, \dots, y_\ell)}]] = \sum_{i=\ell}^n E[1_{[(X_{i-\ell+1}, \dots, X_i)=(y_1, \dots, y_\ell)}]] = \sum_{i=\ell}^n 1/2^\ell = (n - \ell + 1)/2^\ell$ .

<sup>70</sup>Note that the proportion of heads on flips that immediately follow two consecutive heads can be written as the number of (overlapping) HHH instances in  $n$  flips, divided by the number of (overlapping) HH instances in the first  $n - 1$  flips.