



# The Perils of Misspecified Priors and Optional Stopping in Multi-Armed Bandits

Markus Loecher\*

Berlin School of Economics and Law, Berlin, Germany

The connection between optimal stopping times of American Options and multi-armed bandits is the subject of active research. This article investigates the effects of optional stopping in a particular class of multi-armed bandit experiments, which randomly allocates observations to arms proportional to the Bayesian posterior probability that each arm is optimal (*Thompson sampling*). The interplay between optional stopping and prior mismatch is examined. We propose a novel partitioning of regret into peri/post testing. We further show a strong dependence of the parameters of interest on the assumed prior probability density.

**Keywords:** multi-armed bandits, sequential testing, A/B testing, American options, optional stopping

## OPEN ACCESS

### Edited by:

Peter Schwendner,  
Zurich University of Applied Sciences,  
Switzerland

### Reviewed by:

Norbert Hilber,  
ZHAW, Switzerland  
Bertrand Kian Hassani,  
University College London,  
United Kingdom

### \*Correspondence:

Markus Loecher  
markus.loecher@hwr-berlin.de

### Specialty section:

This article was submitted to  
Artificial Intelligence in Finance,  
a section of the journal  
Frontiers in Artificial Intelligence

**Received:** 27 May 2021

**Accepted:** 28 June 2021

**Published:** 09 July 2021

### Citation:

Loecher M (2021) The Perils of  
Misspecified Priors and Optional  
Stopping in Multi-Armed Bandits.  
Front. Artif. Intell. 4:715690.  
doi: 10.3389/frai.2021.715690

## 1 INTRODUCTION

Sequential testing procedures allow an experiment to stop early, once the available/streaming data collected is sufficient to make a conclusion. The benefits of reducing the duration of e.g. Web based A/B tests, clinical trials or the monitoring of adverse events are obviously very attractive in terms of saving costs and lives. Wald (Wald, 2004) proposed a sequential probability ratio test (SPRT) for continuous sequential analyses, where the observation ends if the likelihood ratio exceeds or falls below predetermined bounds. In order to reduce the dependence on the alternate hypothesis, Kulldorff et al. (Kulldorff et al., 2011) introduced the use of a maximized sequential probability ratio test (MaxSPRT), where the alternative hypothesis is composite rather than simple. A detailed comparison of (a modified) MaxSPRT and the O'Brien & Fleming test (O'Brien and Fleming, 1979) can be found in (Kharitonov et al., 2015). Note that all of these methods are typically limited to comparing two treatments at a time and deploy equal allocation, i.e. no attempt is made to maximize "rewards" during the testing phase. Multi-armed bandits (MABs) offer attractive solutions to both of these shortcomings.

One consequence of unplanned optional stopping rules for Null Hypothesis significance testing (NHST) are severely inflated type-I error rates. Bayesian hypothesis testing (BHT) is frequently hailed as a win-win alternative to traditional A/B testing. While there is a growing literature in favor of Bayesian stopping criteria based on the sequential collection of data (Rouder, 2014; Schönbrodt et al., 2017), other authors illustrate the dangers of optional stopping (Erica et al., 2014; Sanborn and Hills, 2014; de Heide and Grünwald, 2017; Hendriksen et al., 2018). This paper expands upon prior work (Loecher, 2017) and specifically focuses on the effects of optional stopping on the multi-armed bandit (MAB) procedure outlined by (Scott, 2010; Scott, 2012). In particular, we use simulations to demonstrate that the choice of the prior distribution has a significant impact on i) the power to detect differences between arms, ii) the average sample size as well as iii) the cumulative and terminal regret. In addition, we examine the precision of the final estimates and the dependence of the sample size on number of arms and effect sizes.

Our work touches upon previous research on American Options (Bank and Föllmer, 2003) that demonstrates a close connection between multi-armed bandits and optimal stopping times in terms of the Snell envelope of the given payoff process.

## 2 MULTI-ARMED BANDITS

As compactly described in (Scott, 2015), the name multi-armed bandit (MAB) describes a hypothetical experiment where one faces  $k > 1$  slot machines (colloquially known as “one-armed bandits”) with potentially different expected payouts. In the typical setup there are  $k$  “arms,”  $a = \{1, \dots, k\}$ , which of course are metaphors for simple Bernoulli experiments with success probability  $p_a$ . Arm  $a$  is associated with an unknown expected reward  $\mu_a$  which for clarity we assume to be proportional to  $p_a$ . We further assume a Beta  $(\alpha_a, \beta_a)$  density as a prior distribution for  $p_a$ . The goal is two-fold: identify the arm with the greatest  $p_a$  (or equivalently  $\mu_a$ ) as soon as possible, and to accumulate the greatest total reward in doing so. This problem is similar to traditional sequential testing from the statistics literature (Wald and Wolfowitz, 1948; Jennison and Turnbull, 2000) but more complex due to the balancing of the so called explore/exploit dilemma: while one wants to find the arm with the highest reward, the total cost of the experiments needs to be minimized at the same time. The fundamental tension is between “exploiting” arms that have performed well in the past and “exploring” new or seemingly inferior arms in case their true performance is even better.

### 2.1 Randomized Probability Matching

Multiple algorithms have been proposed to optimize bandit problems: Upper Confidence Bound (UCB) methods (Lai and Robbins, 1985; Auer et al., 2002) for which strong theoretical guarantees on the regret can be proved, are very popular. The Bayes-optimal approach of Gittins (Gittins et al., 2011) directly maximizes expected cumulative payoffs with respect to a given prior distribution. Randomized probability matching (RPM), also known as Thompson Sampling (TS) (Thompson, 1933; Thompson, 1935), is a particularly appealing heuristic that plays each arm in proportion to its probability of being optimal. Recent results using Thompson sampling seem promising (Graepel et al., 2010; Granmo, 2010; Scott, 2010; May and Leslie, 2011; May et al., 2012), and the theoretical analysis is catching up (e.g. optimal regret guarantees for TS have been proven by (Agrawal and Goyal, 2012)).

In contrast to the fixed length, equal allocation of resources in NHST, RPM assigns the subsequent  $n_{batch}$  samples in proportion to the <sup>1</sup>posterior probability

<sup>1</sup>which can easily be computed either by quadrature or simulation (see Figures 3, 4 in (Scott, 2010)). Note that alternatively (and computationally more efficient) we draw one random sample from each individual arm’s posterior and choose the one with the largest draw.

$$w_{a,t} \equiv P(\mu_a > \mu_{i \neq a} | Y_{a,t}, N_{a,t}) = \int_0^1 \text{Beta}(p_a | Y_{a,t} + \alpha_a, N_{a,t} - Y_{a,t} + \beta_a) \left[ \prod_{j \neq a} \int_{p_j}^1 \text{Beta}(p_j | Y_{j,t} + \alpha_j, N_{j,t} - Y_{j,t} + \beta_j) dp_j \right] dp_a$$

after each batch of recorded samples, where  $Y_{a,t}$  and  $N_{a,t}$  denote the cumulative number of successes and trials observed for arm  $a$  up to time  $t$  (which yield the observed sample proportions  $\hat{p}_{at} = Y_{a,t}/N_{a,t}$ ). Thus, arm  $a \in \{1, \dots, k\}$  obtains  $n_{a,t} = n_{batch} \cdot w_{a,t}$  samples on batch  $t$  respectively. This process is repeated until a set of stopping rules has been satisfied, the discussion of which we will defer to **section 4**. In this paper we do not allow for differences in prior parameters across arms, so from now on we assume  $\alpha_j \equiv \alpha, \beta_j \equiv \beta$ .

### 2.2 Regret

It is worth noting that there are important differences between classical experiments and bandits w.r.t. the assumed optimality criteria. In many situations the “switching costs” between treatment arms with essentially equal rewards is small or even zero, which shifts all the costs to type-II errors. But even the concept of type II errors needs to be modified to properly reflect the loss function of a multi-armed bandit experiment:

1. Instead of “correctly rejecting a hypothesis” we care mainly about identifying the correct superior arm. For two arms, the concept of a *type S* and *type M* error (Gelman and Carlin, 2014) would be more appropriate.
2. For multiple ( $k > 2$ ) arms, there are  $k - 1$  different type-II errors.
3. The magnitude of the difference between the proclaimed and actual superior arm matters greatly.

A popular loss function which addresses all of the above concerns -at least for the testing phase-is regret, defined as the cumulative expected lost reward, relative to playing the optimal arm from the beginning of the experiment (Scott, 2010):

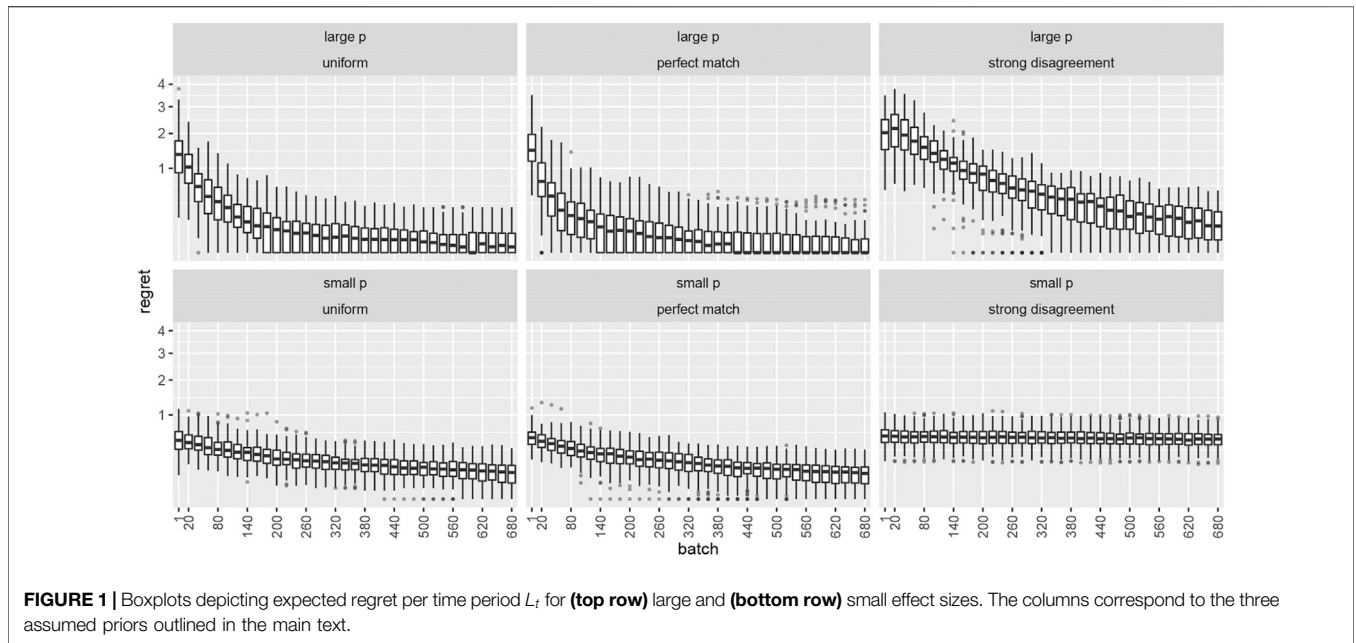
Let  $a_t$  denote the arm of the bandit that was played at time  $t$  and  $\mu_a$  denote the expected reward. Let  $\mu^* = \max_a \{\mu_a\}$  be the expected reward under the truly optimal arm, and let  $n_{a,t}$  denote the number of observations that were allocated to arm  $a$  at time  $t$ . Then the expected cumulative regret is

$$L_T^c = \sum_{t=1}^T L_t = \sum_{t=1}^T \sum_a n_{a,t} (\mu^* - \mu_a)$$

where we refer to the non cumulative  $L_t$  as regret.

### 2.3 Choice of Priors

It is sometimes said that for MABs with Thompson sampling no tuning parameters have to be chosen at the onset of an experiment—unlike in classical testing where one has to fix the smallest difference to be detected at a certain confidence level in advance of the experiment.



**FIGURE 1** | Boxplots depicting expected regret per time period  $L_t$  for (top row) large and (bottom row) small effect sizes. The columns correspond to the three assumed priors outlined in the main text.

We point out that the choice of the prior distribution for the binomial parameter plays a similar role to the parameter settings in NHST. For Gaussian priors (Honda and Takemura, 2014) showed that TS is vulnerable to prior misspecification, and (Bubeck and Liu, 2014) derive prior-free and prior-dependent regret bounds for more general cases.

In Sections 3, 4 we investigate the impact of a misspecified prior on the key metrics of a MAB experiment. Note that the consequences of a mismatch between the assumed and actual prior distribution have recently received a fair amount of attention in the literature (Rouder, 2014; de Heide and Grünwald, 2017; Hendriksen et al., 2018).

In the following simulation study (all simulations are run in R (R Core Team. R, 2016) using the bandit package (Lotze and Loecher, 2014)) we repeated each experiment 500 times with  $k = 10$  arms for a maximum of  $N_{max} = 1000$  periods. Each “batch” consisted of a total of  $n_{batch} = k \times 50$  Bernoulli trials distributed across the  $k$  arms. We need to distinguish between two implementations of the prior distributions (each a beta distr. with parameters  $\alpha, \beta$ ):

1. Sampling the true values for the  $k$  binomial probabilities  $p_i$ . We analyze 2 “effect sizes”:
  - a. “large”:  $p_i$  concentrated in the range  $[0.05, 0.14]$  ( $\alpha = 8.8, \beta = 84$ ).
  - b. “small”:  $p_i$  concentrated in the (4 times narrower) range  $[0.05, 0.14]$  ( $\alpha = 61.5, \beta = 941$ ).
2. The assumed prior in the computation of the posterior. We analyze three types: i) a uniform ( $\alpha = 1, \beta = 1$ ), ii) a perfect match (assumed equal true prior), and iii) “strong disagreement” (a 3 times higher mean of the true prior than the assumed, equal variance).

Hence, we report on a total of  $2 \times 3$  groups of experiments.

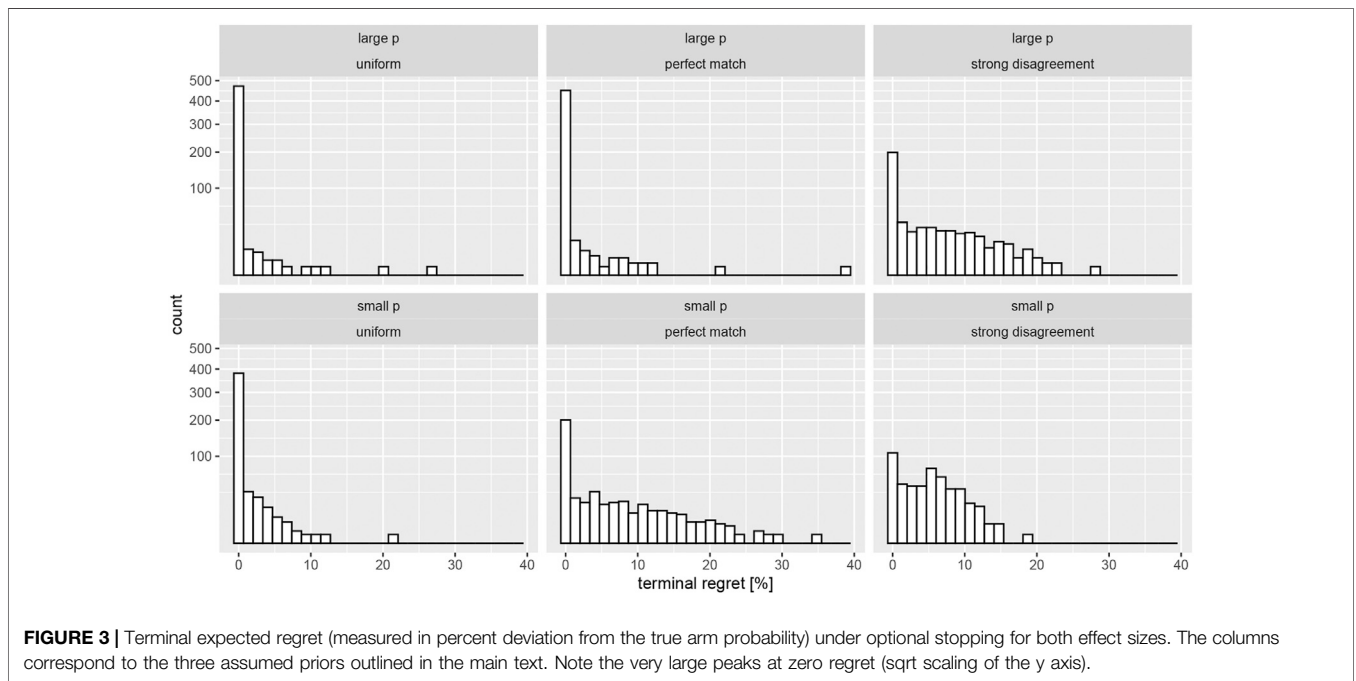
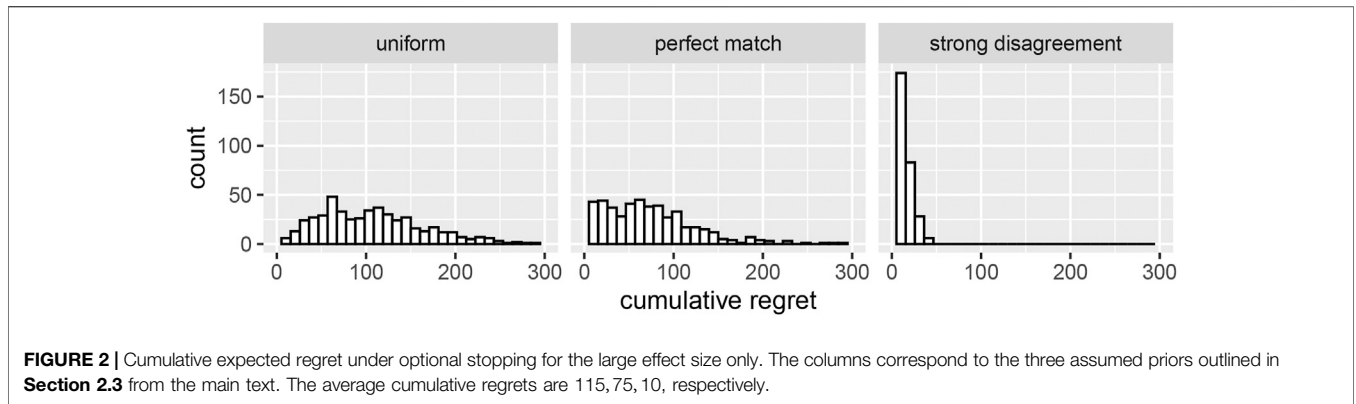
### 3 INFINITE EXPERIMENTS

In principle, bandits can run indefinitely, hence never forfeiting exploration, with the appealing consequence that the superior arms will eventually be correctly identified.

The top row of Figure 1 shows the temporal evolution of the distribution of the expected regret per time period  $L_t$  for the case of relatively large effect sizes. With the exception of persistent outliers, the regret in the first two panels quickly decays to zero as sub-optimal arms are identified and decreasingly likely to receive further explorative samples. The rightmost panel illustrates the effects of a strong disagreement between the assumed and actual prior distribution; while the median regret also decreases, its interquartile range as well as the 95% percentile do not decrease. For the smaller effect size (bottom row), even the median regret does not seem to improve as time passes. Note the similarity of the results for the uniform and the perfectly matched prior which would support the perceived robustness of a non-informative prior. Maybe there is little to be gained (middle panels) but much to be lost (rightmost panels) if one tried to carefully calibrate the prior to the expected actual Bernoulli rates of the particular situation? We will address this question in the next section.

### 4 STOPPED EXPERIMENTS

In reality, infinitely long running experiments are rarely practical or even feasible. The purpose of experimentation is typically to identify superior treatments and eliminate inferior ones. We will discuss and implement two stopping rules both based on the posterior probability  $w_{a,t}$  of arm  $a$  being optimal at time  $t$ . Of course, in practice, a maximum sample size is often enforced as a third termination criterion.



Rule I is to decide in favor of one arm as soon as its posterior probability crosses a threshold which we set to 0.95 and which is related to the power of the test, i.e.  $w_{a,t} > 0.95$ . The second metric being monitored is the potential value remaining in the experiment, which is particularly useful when there are multiple arms. At any instance  $t$ , the arm with the maximum posterior probability,  $a^* = \operatorname{argmax}_a w_{a,t}$ , is the most likely candidate to offer the highest success probability. The value remaining (VR) is the 95th percentile of the distribution of  $(\max(\hat{\mu}_a) - \hat{\mu}_{a^*})/\hat{\mu}_{a^*}$ , where  $\hat{\mu}_a$  are samples from the posteriors of arm(s)  $a$ . The VR attempts to estimate the amount of increased conversion rate<sup>2</sup> one could get by choosing another arm,  $a \neq a^*$ . Note that the VR is a random variable with a distribution that

depends on the observed data and is closely related to the value at risk metric ubiquitous in finance. Rule II (Scott, 2012) is to end the experiment when  $VR < 0.01^3$ . Rule III is to stop the experiments after  $N_{max} = 1000$  batches.

For all three stopping rules, the arm with the largest posterior probability of being optimal is then chosen. Which metric should we use to evaluate the performance of this procedure? While it is tempting to simply take cumulative regret as a measure of loss as done in (Scott, 2010), this would ignore the quality of terminal decisions (It is easy to see that stopping early, i.e. reducing  $T$  would always decrease  $L_T^c$ !). In support of this claim, Figure 2 shows substantial “gains” for the case of the mismatched priors (cumulative regret is just 10). At the same time, we see that the

<sup>2</sup>We use conversion rate synonymously to probability of success in a Bernoulli setting.

<sup>3</sup>i.e. when there is a probability  $\geq 0.95$  that the value remaining is less than 1% of the current estimate of the highest conversion rate  $p_{a^*,t}$ .

**TABLE 1 |** Percent of tests which did not terminate before the maximum test length (= 1000). Shown are results for the two effect sizes and the three priors.

	Uniform	Match	Disagreement
Large	25.4	16.9	0.0
Small	79.1	60.8	0.0

average cumulative regret for the perfectly matched prior is substantially lower (75) than the one under the assumption of a uniform prior (115).

Instead we adapt the train/test framework from machine learning to sequential testing and simultaneously monitor the expected regret after stopping the experiment. We define this terminal regret (TR) as the percent difference between the optimal arm and the one chosen by the bandit through a stopping rule:  $TR = (\mu^* - \mu_a) / \mu^*$ . We mention in passing its close similarity to the so-called simple regret defined in (Bubeck et al., 2011). The total cost of an experiment is a weighted sum of the cumulative and terminal regret, where the weights clearly depend on the ratio of added rewards during (peri) and after (post) testing. **Figure 3** displays the distributions of the relative version of this metric which is most relevant for long running campaigns. As terminal regret is defined conditional upon having triggered the stopping rule, it is helpful to know the proportion of tests that did not terminate before the maximum test length, which is displayed in **Table 1**.

The numbers confirm the pattern gleaned from **Figure 2**; a strong mismatch of priors leads to overly aggressive test terminations, whereas the uniform prior results in a much more cautious decision making. We conclude this section with

a possibly redundant message regarding the precision of the estimates of the superior arm.

**Figure 4** shows the percent difference between the estimated arm probabilities at termination and the true one (for arm 10 which is always chosen to be the one with the largest reward). While there does not appear to be a noteworthy upward bias we again notice a somewhat strong dependence of the precision of the estimates and the assumed prior.

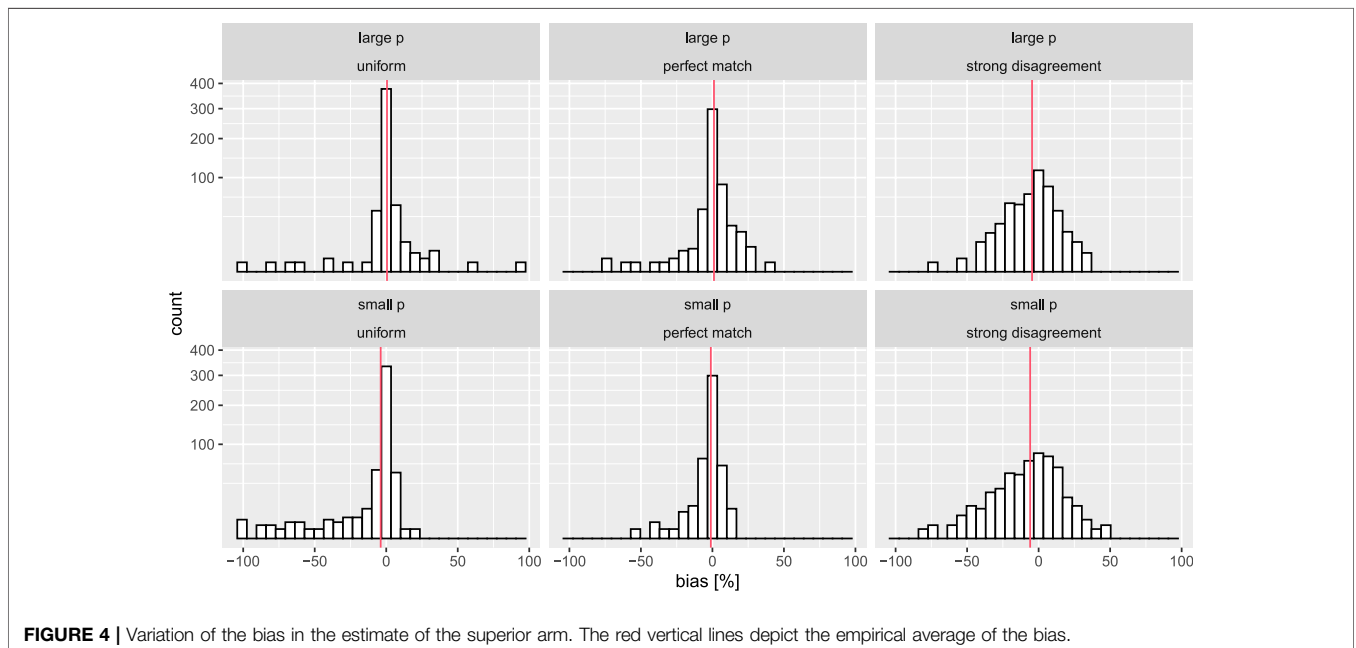
### 4.1 Calibration

While MABs make no guarantees on type-I/II errors, stopping rules I and II suggest precise probabilistic statements which a user might want to rely upon. We now show that the combination of optional stopping and mismatched priors leads to poor calibration of inferential expectations.

**Table 2** displays the percent of correct early stopping decisions for each stopping rule, which should be close to 95% and seems to be correct for the case of agreeing priors (middle column). But the outcomes for the two other priors send a by now familiar message: either one faces overly aggressive test terminations (and hence inflated claims of posterior optimality probabilities) or the MAB is too conservative. The data in **Table 3** complement the conditional nature of **Table 2**.

### 4.2 Accuracy and ASN

All simulations up to this point have been “fully Bayesian” in the sense that the true  $k$  (fixed at  $k = 10$ ) arm probabilities were sampled from the corresponding prior distribution anew for each individual experiment. In this last section, we deviate from that strategy and instead consider equally spaced parameters in the interval  $[0.05 - \Delta; 0.05]$  and vary  $k \in \{2, 5, 10\}$ , where the true effect size is then proportional to  $\Delta$ . We further restrict ourselves to the uniform prior. Our goal is to find a dependence of accuracy



**FIGURE 4 |** Variation of the bias in the estimate of the superior arm. The red vertical lines depict the empirical average of the bias.

**TABLE 2 |** Percent of correct early stopping decisions. Upper table shows results for rule I, lower table for rule II, as explained in the text. These fractions are conditional upon having triggered the stopping rule, the probabilities of which are shown in **Table 3**. For small effect sizes and a perfectly matching prior, a low value remaining is always declared before a significant posterior, which leads to the missing value in the 2nd row.

		Uniform	Match	Disagreement
Posterior >0.95	Large	100.0	95.0	48.7
	Small	80.0		21.3
Value <0.01	Large	98.4	96.4	49.2
	Small	94.1	96.2	27.2

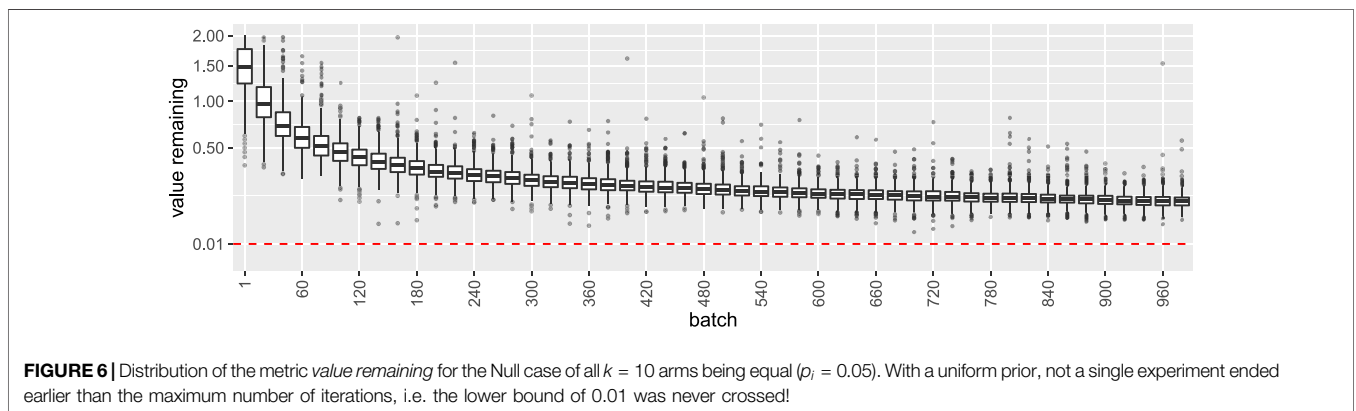
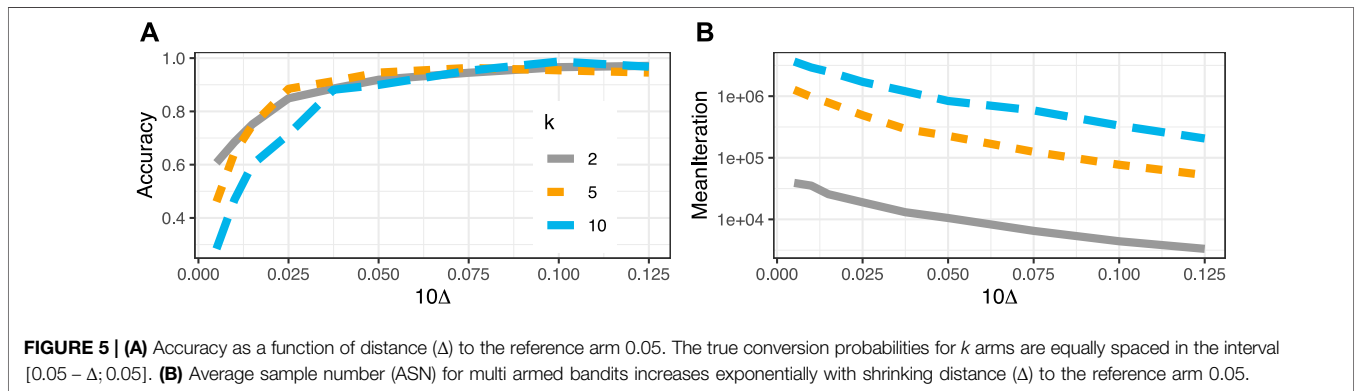
**TABLE 3 |** Percent of triggered early stopping decisions. Upper table shows results for rule I, lower table for rule II, as explained in the text. The column sums for corresponding rows in each table may add up to > 100 as stopping rules are not mutually exclusive.

		Uniform	Match	Disagreement
Posterior >0.95	Large	5.0	16.1	55.0
	Small	1.0	0.0	67.6
Value <0.01	Large	74.6	83.1	100.0
	Small	21.0	39.2	100.0

and average sample number (ASN) on this effect size, which is somewhat more straightforward for this simpler experimental setup.

The left panel of **Figure 5** shows the proportion of choosing the correct arm for a total of 2, 5 and 10 arms respectively as a function of the distance  $\Delta$  to the (optimal) reference arm  $p_0 = 0.05$ . For just two arms, the accuracy drops from the claimed 0.95 to 0.6 as  $\Delta$  decreases from 0.01 to 0.0005 (Since the goal of experimentation is to find the optimal arm, error metrics from classification seem more suited than those from a traditional hypothesis testing framework; we therefore prefer accuracy over e.g. power). For multiple arms the loss of accuracy is much more dramatic and the onset earlier. The right panel of **Figure 5** displays the average number of iterations/samples (ASN) the bandits took to make a decision. The logarithmic y-axis reveals an exponentially increasing effective sample size as the differences between the arms wane.

We treat the special case of  $\Delta = 0$  separately as a “Null simulation.” Surprisingly, with a uniform prior distribution, collecting  $n_{batch} \times k \times N_{max} = 50 \times 10 \times 1000$  samples does not generate sufficient evidence to declare a value remaining to be less than 1%, as **Figure 6** shows.



## 5 CONCLUSION

Our contribution is to demonstrate a heightened sensitivity to prior assumptions in MABs when data-dependent stopping rules are used; a problem which has been known to be true in simple Bayesian inference for a few decades by now (Rosenbaum and Rubin, 1984). We have also shown that i) early high confidence decisions in MABs are much less reliable than claimed, and that ii) the prior parameters have a significant impact on most important metrics used to evaluate MABs. It seems that the burden of specifying tuning parameters has not vanished but is less directly obvious than in competing methods. Clearly, the sensitivity and expected sample size are determined by the beta prior parameters  $\alpha, \beta$ . While the default uniform prior offers reasonable performance, it could be optimized by choosing more appropriate priors, at the risk of greatly worsening the outcome.

We further worry about the inability to easily set priors for the differences in Bernoulli probabilities rather than just the parameters themselves. The additional flexibility would allow to optimize for much more realistic experimental situations.

Of course, it is true that the cost of an experiment can be substantially reduced by deploying RPM based bandits instead of pre-committing to a fixed sample size. Allowing to rapidly detect large differences and not waste resources sticking to an

unnecessarily rigid protocol is obviously a benefit of most sequential testing algorithms. One difference being that optimization attempted by Bayesian bandits is not concerned with a null hypothesis but with minimizing the posterior expected loss. The implicit assumption that Type-I errors incur no additional cost is likely to be wrong in realistic applications: it seems more reasonable that switching campaigns for no good reason should be avoided. These insights complement recent literature on frequentist-oriented bandit algorithms implementing optional stopping (Kharitonov et al., 2015; Johari et al., 2017; Aziz et al., 2018; Kaufmann et al., 2018).

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

ML designed and executed the research.

## REFERENCES

- Agrawal, S., and Goyal, N. (2012). "Analysis of thompson Sampling for the Multi-Armed Bandit Problem," in Conference on learning theory (JMLR Workshop and Conference Proceedings), Edinburgh, Scotland, June 25–27, 2012, 39–41.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learn.* 47, 235–256. doi:10.1023/a:1013689704352
- Aziz, M., Anderton, J., Kaufmann, E., and Aslam, J. (2018). Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence. arXiv preprint arXiv:1803.04665.
- Bank, P., and Föllmer, H. (2003). "American Options, Multi-Armed Bandits, and Optimal Consumption Plans: A Unifying View," in *Paris-Princeton Lectures on Mathematical Finance 2002* (Berlin, Heidelberg: Springer-Verlag), 1–42. doi:10.1007/978-3-540-44859-4\_1
- Bubeck, S., and Liu, C. Y. (2014). "Prior-free and Prior-dependent Regret Bounds for thompson Sampling," in 2014 48th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, Mar 19–21, 2014, (IEEE), 1–9.
- Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure Exploration in Finitely-Armed and Continuous-Armed Bandits. *Theor. Comput. Sci.* 412, 1832–1852. doi:10.1016/j.tcs.2010.12.059
- de Heide, R., and Grünwald, P. D. (2017). Why Optional Stopping Is a Problem for Bayesians. arXiv preprint arXiv:1708.08278.
- Erica, C. Y., Sprenger, A. M., Thomas, R. P., and Dougherty, M. R. (2014). When Decision Heuristics and Science Collide. *Psychon. Bull. Rev.* 21, 268–282. doi:10.3758/s13423-013-0495-z
- Gelman, A., and Carlin, J. (2014). Beyond Power Calculations. *Perspect. Psychol. Sci.* 9, 641–651. doi:10.1177/1745691614551642
- Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed Bandit Allocation Indices*. Hoboken, New Jersey: John Wiley & Sons.
- Graepel, T., Candela, J. Q., Borchert, T., and Herbrich, R. (2010). "Web-scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine," in Proceedings of the 27th International Conference on Machine Learning ICML 2010, Haifa, Israel, June 21–24, (Omnipress).
- Granmo, O. C. (2010). Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *Int. J. Intell. Comput. Cybernetics* 3 (2), 207–234. doi:10.1108/17563781011049179
- Hendriksen, A., de Heide, R., and Grünwald, P. (2018). Optional Stopping with Bayes Factors: a Categorization and Extension of Folklore Results, with an Application to Invariant Situations. arXiv preprint arXiv:1807.09077.
- Honda, J., and Takemura, A. (2014). Optimality of thompson Sampling for Gaussian Bandits Depends on Priors. In *Artificial Intelligence and Statistics* (PMLR), 375–383.
- Jennison, C., and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman-Hall/CRC.
- Johari, R., Kooen, P., Pekelis, L., and Walsh, D. (2017). "Peeking at A/B Tests: Why it Matters, and what to Do about it," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), 1517–1525.
- Kaufmann, E., Koolen, W. M., and Garivier, A. (2018). Sequential Test for the Lowest Mean: From thompson to murphy Sampling. *Adv. Neural Inf. Process. Syst.*, 6332–6342.
- Kharitonov, E., Vorobev, A., Macdonald, C., Serdyukov, P., and Ounis, I. (2015). "Sequential Testing for Early Stopping of Online Experiments," in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM), Santiago, Chile, August 9–13, 2015, 473–482.
- Kulldorff, M., Davis, R. L., Kolczak, M., Lewis, E., Lieu, T., and Platt, R. (2011). A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. *Sequential Anal.* 30, 58–78. doi:10.1080/07474946.2011.539924
- Lai, T. L., and Robbins, H. (1985). Asymptotically Efficient Adaptive Allocation Rules. *Adv. Appl. Math.* 6, 4–22. doi:10.1016/0196-8858(85)90002-8
- Loecher, M. (2017). "Are Multi-Armed Bandits Susceptible to Peeking?," in 2nd Business & Entrepreneurial Economics Conference (BEE 2017), Brijuni, Croatia, May 24–26, 2017.
- Lotze, T., and Loecher, M. (2014). Bandit: Functions for Simple A/B Split Test and Multi-Armed Bandit Analysis. R package version 0.5.0.
- May, B. C., Korda, N., Lee, A., and Leslie, D. S. (2012). Optimistic Bayesian Sampling in Contextual-Bandit Problems. *J. Machine Learn. Res.* 13, 2069–2106.

- May, B. C., and Leslie, D. S. (2011). *Simulation Studies in Optimistic Bayesian Sampling in Contextual-Bandit Problems*, 11. Statistics Group, Department of Mathematics, University of Bristol.
- O'Brien, P. C., and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics* 35, 549–556.
- R Core Team. R (2016). *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rosenbaum, P. R., and Rubin, D. B. (1984). Sensitivity of Bayes Inference with Data-dependent Stopping Rules. *The Am. Statistician* 38, 106–109. doi:10.1080/00031305.1984.1048317610.2307/2683243
- Rouder, J. N. (2014). Optional Stopping: No Problem for Bayesians. *Psychon. Bull. Rev.* 21, 301–308. doi:10.3758/s13423-014-0595-4
- Sanborn, A. N., and Hills, T. T. (2014). The Frequentist Implications of Optional Stopping on Bayesian Hypothesis Tests. *Psychon. Bull. Rev.* 21, 283–300. doi:10.3758/s13423-013-0518-9
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Psychol. Methods* 22, 322–339. doi:10.1037/met0000061
- Scott, S. L. (2012). Multi-armed Bandit Experiments. Available at: <https://analytics.googleblog.com/2013/01/multi-armed-bandit-experiments.html> (Accessed 05 30, 2017).
- Scott, S. L. (2010). A Modern Bayesian Look at the Multi-Armed Bandit. *Appl. Stochastic Models Bus. Ind.* 26, 639–658. doi:10.1002/asmb.874
- Scott, S. L. (2015). Multi-armed Bandit Experiments in the Online Service Economy. *Appl. Stochastic Models Bus. Ind.* 31, 37–45. doi:10.1002/asmb.2104
- Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 285–294. doi:10.1093/biomet/25.3-4.285
- Thompson, W. R. (1935). On the Theory of Apportionment. *Am. J. Maths.* 57, 450–456. doi:10.2307/2371219
- Wald, A. (2004). *Sequential Analysis*. (Courier Corporation).
- Wald, A., and Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *Ann. Math. Statist.* 19, 326–339. doi:10.1214/aoms/1177730197

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Loecher. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.