Multinomial Thompson Sampling for adaptive experiments with rating scales

Multinomial Thompson Sampling per esperimenti adattivi con scale di rating

Nina Deliu

Abstract Bandit algorithms such as Thompson Sampling (TS) have been put forth for decades as useful for conducting adaptively-randomized experiments. By skewing the allocation ratio towards superior arms, they can substantially improve participants' welfare with respect to particular outcomes of interest. For example, as we illustrate in this work, they may use participants' ratings for understanding and assigning promising text messages for managing mental health issues more often. However, model-based algorithms such as TS, typically assume binary or normal models, which may lead to suboptimal performances in categorical rating scale outcomes. Guided by our field experiment, we extend the application of TS to rating scale data and show its improved performance in a number of synthetic experiments.
Abstract *Gli algoritmi di tipo multi-armed bandits, quali il Thompson Sampling (TS), rappresentano un metodo efficace per la conduzione di esperimenti adattivi. Adattando automaticamente il rapporto di allocazione verso il braccio superiore (quando questo esiste), essi possono migliorare sostanzialmente sia il welfare dei partecipanti–anche semplicemente rilevando le loro preferenze–che l'efficacia e il costo di uno studio. Tuttavia, algoritmi come il TS assumono generalemnte un modello binario o normale per la variabile di risposta, portando a risultati non ottimali in caso di dati con scale di rating. Motivati da un esperimento reale sulla salute mentale, in questo lavoro si propone un'estensione dell'algoritmo TS per scale di valutazione e si studiano la sue performances in una serie di esperimenti sintetici.*

Key words: Adaptive experiments, Thompson Sampling, Multi-armed bandits, Rating Scales, Multinomial Model, Dirichlet Distribution

---

Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF), Sapienza Università di Roma; e-mail: nina.deliu@uniroma1.it

1 Introduction

Adaptively-randomized experiments have the potential to enhance participants' welfare while collecting high-quality data, resulting in a more flexible, efficient, and ethical alternative compared to traditional fixed studies [1]. In such experiments, allocation ratios are skewed towards more efficient or informative arms, with the goal of assigning superior arms to as many participants as possible.

Multi-armed bandit (MAB) algorithms [2] have been argued for decades as useful to adaptively randomize experiments with the aim of optimizing an outcome of interest. In MAB problems, an *agent* chooses at each time t one of the available *arms* for which a *reward* is then provided. The goal is to maximize the cumulative reward, or equivalently minimize regret, under uncertainty on the best arm. As rewards are generally stochastic, and they might vary between arms, the agent must balance *exploring* each arm to gain information with *exploiting* the information gained so far by choosing arms with the higher expected reward. *Thompson Sampling* (TS) is a highly-interpretable randomized MAB strategy that allocates arms in proportion to their posterior mean reward. Due to its empirical and theoretical optimality [3], it has received renewed attention in recent years and have been successfully applied in a wide variety of domains, going from recommendation to education [4].

In many real-world applications, arms are a set of recommendations or explanations, that are allocated according to users' preferences. In this work, for example, we are motivated by mental health experiments we deployed on the Amazon Mechanical Turk [5], in which participants were asked to rate two types of messages in terms of how helpful they would be for managing their mood. Results of a primary experiment (MTurk I), based on a fixed design with an equal allocation ratio for gaining benchmark knowledge, are reported in **Fig. 1**.

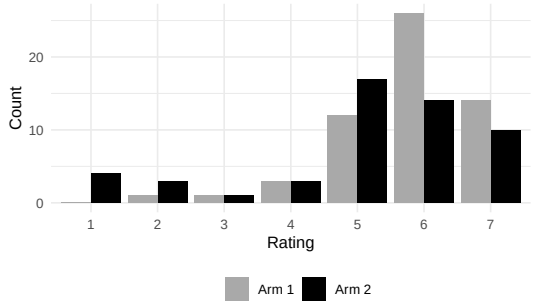

Fig. 1 Arms' ratings distribution in the two-armed *MTurk I* experiment, based on a 7-point Likert scale. A small superiority of Arm 1 vs Arm 2 is shown: sample means $\hat{\mu}_1 = 5.81$ vs $\hat{\mu}_2 = 5.08$, sample standard deviation $\hat{\sigma}_1 = 1.04$ vs $\hat{\sigma}_2 = 1.72$, for a sample size $N = 110$.

Subsequent experiments [6] have then been conducted adaptively with TS and a Gaussian model, in absence of more adequate modeling procedures for rating data within TS. Notably, despite the vast array of MAB and TS variants, most of them assume either binary or Gaussian rewards, and a shortage of MAB methodologies to handle rating scale data persists. The general practices in such cases consist in either dichotomizing the ordinal reward with a threshold (often arbitrary) rule and

then using a binary model, or using a normal model directly on the rating outcomes, as in our MTurk study. Such practices have been long recognized as suboptimal in terms of both reward efficiency [7] and statistical inference [8].

In this work, we extend the applicability of TS to rating scale data, introducing the *Multinomial TS*, which can be easily implemented with the *Dirichlet* conjugate family [9]. Then, guided by our motivating experiment, we evaluate its performance compared to the standard Gaussian TS in a number of synthetic experiments.

2 Problem Setting and Methods

Experimental set-up We consider a two-armed T-time horizon experiment, in which participants are accrued in a fully-sequential way, with a total experiment sample size $N = T$. At accrual, each participant $t = 1, \dots, T$ is assigned one of the two available arms, say $A_t \in \{1,2\}$, and an outcome or reward $Y_t(A_t)$ associated with that arm is subsequently observed before the next time $t+1$. Arms are drawn according to a policy $\boldsymbol{\pi_t} \doteq \{\pi_{t,k}, k = 1,2\}$, where $\pi_{t,k}$ is the allocation probability of arm k at time t. Given the history of selected arms and associated rewards $\mathscr{H}_t \doteq \{A_\tau, Y_\tau(A_\tau), \tau = 1, \dots, t\}$, the goal of an adaptive experiment may be to find an allocation policy so as to maximize the expected cumulative reward over T.

In line with MTurk I, we consider a 7-points rating scale outcome, i.e., $Y_t \in \{1,2,\dots,7\}$ for all t's, with higher values indicating better outcomes. We assume they are drawn independently from a fixed distribution that depends on the arm but not on t (*stochastic stationary bandits*; [2]), with the following conditional mean:

$$\mathbb{E}(Y_{t,i}(A_{t,i})|\mathscr{H}_{t-1}) = \mathbb{E}(Y_t|A_t) = \mu_1 \mathbb{I}(A_t = 1) + \mu_2 \mathbb{I}(A_t = 2), \tag{1}$$

with (β_1, β_2) the unknown arm parameters. In the next sections we will make an additional assumption on rewards distribution, maintaining the parameters unknown.

Thompson Sampling Rooted in a Bayesian framework, TS allocates arms in proportion to their posterior probability of being associated with the maximum expected reward at each round t. In a two-armed setting, denoting by π_{t1}^{TS} TS's allocation probability of arm 1 at step t, and considering Eq. (1), we have that:

$$\pi_{t1}^{\text{TS}} = \mathbb{P}\Big(\mathbb{E}\big(Y_t(A_t = 1)\big) \geq \mathbb{E}\big(Y_t(A_t = 0)\big)|\mathscr{H}_{t-1}\Big) = \mathbb{P}\Big(\mu_{t1} \geq \mu_{t2}|\mathscr{H}_{t-1}\Big), \quad \forall t.$$

The typical way for implementing TS, involves drawing at each round t a sample from the posterior distribution of each of the unknown arms' parameters in Eq. (1), say $\tilde{\mu}_{tk}$, with $k = 1,2$, and then selecting the arm associated with the highest posterior estimated mean reward $\mathbb{E}(\tilde{Y}_t|A_t = k) = \tilde{\mu}_{tk}$, i.e., $\tilde{a}_t \doteq \text{argmax}_{k=1,2}\mathbb{E}(\tilde{Y}_t|A_t = k) = \text{argmax}_{k=1,2}\tilde{\mu}_{tk}$. A conjugate family is generally assumed for the reward variable, with the most common ones being the Binomial and the Gaussian [3].

2.1 Our Proposal: Multinomial Thompson Sampling

The multinomial distribution is the extension of the binomial distribution for categorical outcomes with more than two response categories [10]. Given J mutually exclusive categories of an outcome Y, among which one and only one category is observed at each time t (i.e., $\sum_{j=1}^{J}\mathbb{I}(Y_t=j)=1$ and $\sum_{t=1}^{T}\sum_{j=1}^{J}\mathbb{I}(Y_t=j)=T$), it models the probability of counts $X_{tj}=\sum_{\tau=1}^{t}\mathbb{I}(Y_t=j), \forall j\in[1,J], \forall t\in[1,T]$. At a given time t, the probability mass function of a multinomial, denoted by Multinom$(t;\boldsymbol{p})$, with $\boldsymbol{p}=(p_1,\dots,p_J)=\Big(\mathbb{P}(Y_t=1),\dots,\mathbb{P}(Y_t=J)\Big)$ the unknown parameters, is given by:

$$f(x_{t1},\dots,x_{tJ};t;p_1,\dots,p_J)=\left(\frac{t!}{\prod_{j=1}^{J}x_{tj}!}\right)\prod_{j=1}^{M}p_j^{x_{tj}}, \tag{2}$$

where $t=\sum_{j=1}^{J}x_{tj}$. For a multinomial distribution, we have that $\mathbb{E}(X_{tj})=tp_j$. In case of ordinal, real valued categories (e.g., $j=1,\dots,7$ as in our rating scales), we can compute the mean of the outcome Y_t at each t, as $\mu_t=\mathbb{E}(Y_t)=\sum_{j=1}^{J}jp_j$.

Eq. (2) can be also expressed using the gamma function Γ, directly showing its resemblance to the Dirichlet distribution, which is its conjugate prior. Given $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_J)$, with $\alpha_j>0, \forall j\in[1,J]$, this is denoted by Dir$(\boldsymbol{\alpha})$, and models our beliefs/knowledge on the unknown parameters as:

$$f(p_1,\dots,p_J;\alpha_1,\dots,\alpha_J)=\left(\frac{\prod_{j=1}^{J}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{J}\alpha_j)}\right)\prod_{j=1}^{M}p_j^{\alpha_j-1}, \tag{3}$$

where $\boldsymbol{p}$ belongs to the standard $J-1$ simplex ($\sum_{i=1}^{J}p_j=1$ and $p_j\geq 0, \forall j\in[1,J]$).

We then update our beliefs on the parameters based on the observed outcomes, by updating their posterior distribution as $\alpha_j\leftarrow\alpha_j+c_{tj}$, $\forall t$'s and $\forall j$'s, where $c_{tj}=\mathbb{I}(y_t=j)$ states whether category j was observed at time t or not (see Algorithm 1).

Algorithm 1 Multinomial TS pseudocode

Input: Time horizon T, prior parameters $\boldsymbol{\alpha_k}=(\alpha_{1k},\dots,\alpha_{Jk})$, with J the overall number of categories of the rating scale outcome, and $k=1,2$ the study arm.

for $t=1,2,\dots,T$ **do**
  for $k=1,2$ **do**
    Sample $\tilde{\boldsymbol{p}}_{\boldsymbol{k}}=(\tilde{p}_{1k},\dots,\tilde{p}_{Jk})\sim$ Dir$(\alpha_{1k},\dots,\alpha_{Jk})$
    Compute $\tilde{\mu}_{tk}=\sum_{j=1}^{J}j\tilde{p}_{jk}$
  end for
  Select arm $\tilde{a}_t=\text{argmax}_{k=1,2}\tilde{\mu}_{tk}$ and observe the reward y_t.
  for $j=1,\dots,J$ **do**
    Compute $c_{tj}=\mathbb{I}(y_t=j)$
  end for
  for $k=1,2$ **do**
    Update posteriors: $(\alpha_{1k},\dots,\alpha_{Jk})\leftarrow(\alpha_{1k},\dots,\alpha_{Jk})+(c_{t1},\dots,c_{tJ})$
  end for
end for

3 Synthetic Experiments and Results

For our empirical evaluation, we focus on the set-up introduced in Section 2, with $T = 1000$, and simulation scenarios defined according to the MTurk I motivational study in **Fig. 1**. We evaluate the proposed Multinomial TS with equal Dirichlet priors for the two arms with parameters $\alpha_{jk} = 1$ for $j = 1,\ldots,7$ and $k = 1,2$, and compare it to a Gaussian TS with variances set according to the results of the MTurk I experiment and Normal $N(0,10)$ priors. We also illustrate the behavior of an *Oracle* that always allocates the best arm–when one exists–or both of them with equal probability–when they are identical. We look at standard MAB performance measures.

Regret and Optimal Arm Allocation under a unique optimal arm The unique optimal arm scenario is simulated with $\mu_1 = 5.81 > \mu_2 = 5.08$ ($\sigma_1 = 1.04$, $\sigma_2 = 1.72$), and individual scale frequencies given by the MTurk sample estimates: $p_1 = (0.00, 0.02, 0.02, 0.05, 0.21, 0.45, 0.24)$ and $p_2 = (0.08, 0.06, 0.02, 0.06, 0.33, 0.27, 0.19)$. Results in **Fig. 2** show the increased performance of the proposed Multinomial TS over Normal TS.

Arm Allocation under identical arms In addition to the aforementioned scenario reflecting the MTurk experiment, we now evaluate the behavior of the proposed strategy in an identical arms scenario for understanding how it balances the allocation of one arm over the other one, when no one should be preferred. This scenario is simulated with $\mu_1 = \mu_2 = 4$ ($\sigma_1 = \sigma_2 = 2$) and symmetric scale frequencies: $p_1 = p_2 = (0.02, 0.09, 0.23, 0.31, 0.23, 0.09, 0.02)$. The latter resembles a Gaussian distribution for ensuring higher comparability between the two TS alternatives. Results in **Fig. 2** show that Multinomial TS is more balanced in arms' allocation compared to a Gaussian TS, which can be very extreme–allocating one arm almost the totality of the times–even when this is not superior.

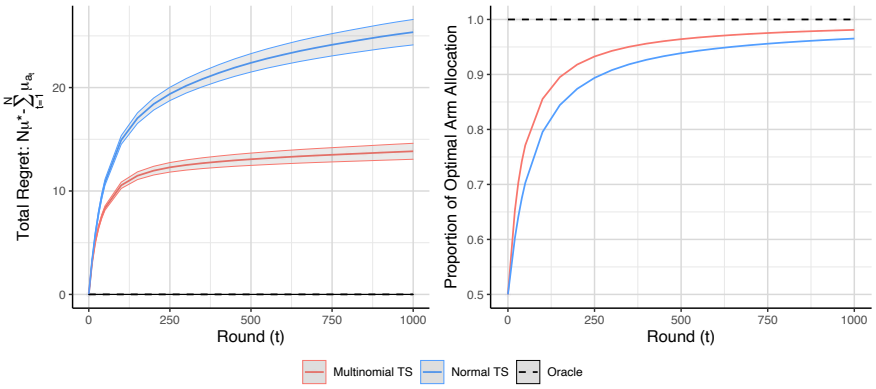

Fig. 2 Regret and proportion of optimal arm allocation in the proposed Multinomial TS vs a Gaussian TS. Values are obtained by averaging across 10^4 independent TS trajectories.

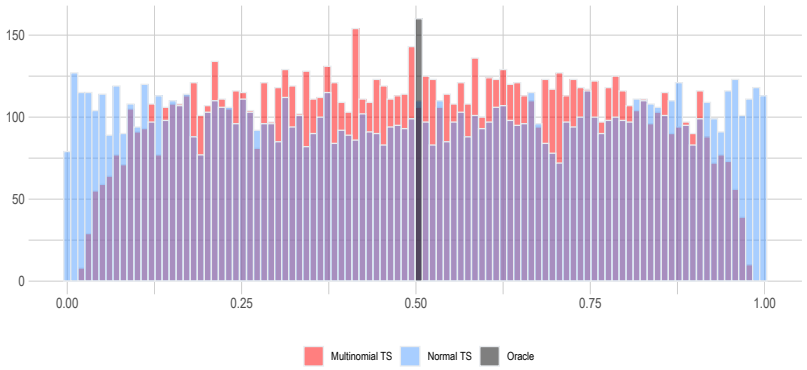

Fig. 3 Empirical allocation of one of the two identical arms (Arm 1). Values are obtained by averaging across 10^4 independent TS trajectories.

4 Conclusion

In this work, motivated by the MTurk field experiment, we extended the applicability of TS to rating scales data, introducing the Multinomial TS. We demonstrated that it can outperform the widely used TS with a Gaussian model in scenarios with a unique optimal arm, and that it is a more balanced solution–that can translate into inferential advantages, due to a lower imbalance in the allocation (see e.g., [6])–when arms are identical. Further work is required to understand how the proposed version would behave in a multi-armed setting or under non-stationarities.

References

1. Bothwell, L. E., Avorn, J., Khan, N. F., & Kesselheim, A. S. (2018). Adaptive design clinical trials: a review of the literature and ClinicalTrials. gov. *BMJ open*, 8(2), e018320.
2. Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.
3. Agrawal, S., & Goyal, N. (2013). Further optimal regret bounds for Thompson Sampling. *In Artificial Intelligence and Statistics* (pp. 99-107). PMLR.
4. Williams, J. J., Rafferty, A., Tingley, D., Ang, A., Lasecki, W. S., & Kim, J. (2018). Enhancing Online Problems Through Instructor-Centered Tools for Randomized Experiments. *CHI2018, 36th Annual ACM Conference on Human Factors in Computing Systems*.
5. Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1), 1-23.
6. Deliu, N., Williams, J. J., & Villar, S. S. (2021). Efficient Inference Without Trading-off Regret in Bandits: An Allocation Probability Test for Thompson Sampling. *arXiv:2111.00137*.
7. Williamson, S. F., & Villar, S. S. (2020). A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, 76(1), 197-209.
8. Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, 332(7549), 1080.
9. Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). *Continuous multivariate distributions, Volume 1: Models and applications* (Vol. 1). John Wiley & Sons.
10. Agresti, A. (2013). *Categorical data analysis. Third edition.* Wiley series in Probability and statistics. Wiley.