From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits

Anonymous Author(s)
Affiliation
Address
email

Abstract

The stochastic multi-arm bandit problem has been extensively studied under standard assumptions on the arm's distribution (e.g bounded with known support, exponential family, etc). These assumptions are suitable for many real-world problems but sometimes they require knowledge (on tails for instance) that may not be precisely accessible to the practitioner, raising the question of the robustness of bandit algorithms to model misspecification. In this paper we study a generic *Dirichlet Sampling* (DS) algorithm, based on pairwise comparisons of empirical indices computed with *re-sampling* of the arms' observations and a data-dependent *exploration bonus*. We show that different variants of this strategy achieve provably optimal regret guarantees when the distributions are bounded and logarithmic regret for semi-bounded distributions with a mild quantile condition. We also show that a simple tuning achieve robustness with respect to a large class of unbounded distributions, at the cost of slightly worse than logarithmic asymptotic regret. We finally provide numerical experiments showing the merits of DS in a decision-making problem on synthetic agriculture data.

1 Introduction

The K-armed stochastic bandit model is a decision-making problem in which a learner sequentially picks an action among K alternatives, called arms, and collects a random reward. In this setting, all rewards drawn from an arm are independent and identically distributed. Hence, we can formally associate each arm $k \in \{1, \dots, K\}$ with its reward distribution ν_k, with mean μ_k. The objective of the learner is to adapt her strategy $(A_t)_{t\in[T]}$ in order to maximize the expected sum of rewards obtained after T selections (where T is the horizon, unknown to the learner). This is equivalent to minimizing the *regret*, defined as the difference between the expected total reward of an oracle strategy always selecting an arm with largest mean and that of the algorithm, which is equal to

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu^\star - \mu_{A_t}\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}\left[N_k(T)\right] \ . \tag{1}$$

Here, $N_k(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = k)$ denotes the number of selections of arm k after T time steps, $\mu^\star = \max_{j\in\{1,\dots,K\}} \mu_j$ and $\Delta_k = \mu^\star - \mu_k$ is called the *gap* between arm k and the largest mean. To assess the performance of a bandit algorithm, one naturally studies the best guarantees achievable by a uniformly efficient algorithm, i.e with sub-linear regret on any instance of a given class of problems. This guarantee was first provided by Lai and Robbins (1985) for parametric families of distributions, and then extended by Burnetas and Katehakis (1996) for more general families. It states that any

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

algorithm that is uniformly efficient[1] on a family of distributions $\mathcal{F}$ must satisfy

$$\liminf_{T\to\infty} \frac{\mathcal{R}_T}{\log(T)} \geq \sum_{k:\Delta_k>0} \frac{\Delta_k}{\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k,\mu^\star)} \ , \quad \mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k,\mu^\star) = \inf_{G\in\mathcal{F}} \{\mathrm{KL}(\nu_k,G) : \mathbb{E}_G(X) > \mu^\star\} \ . \quad (2)$$

A bandit algorithm is then called *asymptotically optimal* for a family of distributions $\mathcal{F}$ when its regret matches this lower bound. When $\mathcal{F}$ is a *Single-Parameter Exponential Family* (SPEF), $\mathcal{K}_{\inf}^{\mathcal{F}}$ is simply the Kullback-Leibler divergence between the distribution of mean μ_k and that of mean μ in $\mathcal{F}$, making for a theoretically appealing setting. The quantity $\mathcal{K}_{\inf}^{B}$, corresponding to the family $\mathcal{F}_{[-\infty,B]}$ of distributions supported in $(-\infty, B]$ is also often considered in the literature, see e.g (Honda and Takemura, 2010, 2015; Cappé et al., 2013).

Overview of existing strategies An efficient strategy faces the classical exploration/exploitation dilemma: it needs to obtain enough information from arms that have not been sampled a lot (exploration), but also to sample arms that are well-performing sufficiently often (exploitation). Many algorithms have been proposed for the multi-armed bandits problem (see Lattimore and Szepesvári (2020) for a survey), and we propose in the following a non-exhaustive list of such methods. A first category contains the deterministic index policies, built on the concept of *Optimism in Face of Uncertainty*, the most celebrated of which being the Upper Confidence Bound (UCB) algorithms (Agrawal, 1995; Auer et al., 2002). These algorithms can obtain a logarithmic regret under classical hypothesis on the distributions (e.g bounded, sub-gaussian, sub-exponential, ...), and the strongest guarantees have been achieved by kl-UCB Cappé et al. (2013), DMED (Honda and Takemura, 2010), and IMED (Honda and Takemura, 2015), which share a common pattern of solving a convex optimization problem at each round. To be asymptotically optimal, these algorithms require either 1) the knowledge of a specific SPEF for each arm, or 2) a known upper bound on the support of each arm. A second general category is that of *randomized* bandit algorithms, which has been formulated for instance in (Kveton et al., 2019b) as *General Randomized Exploration* (GRE). The common feature of these methods is that, at each time step and for each arm, the algorithm draws an index from a distribution that depends on 1) the rewards observed from the arm, and 2) some knowledge on the arms distributions and chooses the arm with the largest index. Thompson Sampling (TS) (Thompson, 1933; Agrawal and Goyal, 2012) belongs to this category, and a proper choice of Bayesian prior/posterior ensures optimality of TS in SPEF (Korda et al., 2013). Different algorithms using Bootstrapping schemes have also been proposed (Osband and Roy, 2015; Kveton et al., 2019a,b; Wang et al., 2020; Riou and Honda, 2020): they share the idea of computing a noisy mean for empirical samples, enhanced by some exploration aid appropriately tuned to the family of distributions they consider. A last category contains the methods based on *sub-sampling* Baransi et al. (2014); Chan (2020); Baudry et al. (2020), that achieve asymptotic optimality in SPEF *without knowing which family*, when all arms share the same. However the proofs heavily rely on properties of the tails of SPEF so the results seem difficult to generalize outside these families.

Motivations While many algorithms achieve optimal regret for bounded distributions with the sole knowledge of the upper bound, the assumptions needed for algorithms working with unbounded distributions (e.g SPEF, sub-Gaussian, sub-exponential) generally assume a known parametric model for the tails. While such assumption entails convenient properties on the theoretical side, the practitioner may have some difficulty to determine which setting/parameters correspond to her problem. Furthermore, this uncertainty raises the question of robustness with respect to these hypotheses. Several works have considered this question: Hadiji and Stoltz (2020) shows that adapting to an unknown bounded range requires a tradeoff between instance-dependent and worst-case regret, and recently (Agrawal et al., 2020; Ashutosh et al., 2021) proved the impossibility of an instance-dependent logarithmic regret for light-tailed distributions without explicit control on the tail parameters. The root cause for this is the lack of compactness of such families $\mathcal{F}$, which allows mass to "leak" at infinity so that maximally confusing distributions with mean μ^* exist arbitrarily close to ν_k, meaning $\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k,\mu^*) = 0$. The latter work also introduces a robust variant of UCB, that trades off logarithmic regret for $\mathcal{O}\left(f(T)\log(T)\right)$, where f essentially tracks the possible mass leakage at infinity. These results puts into question the usual hypotheses under which bandit algorithms are designed: considering a *parametric* control of the tails is indeed sensitive to model mis-specification, but on the other hand the examples chosen to prove infeasability results seem a bit extreme for the practitioner. In this paper, we propose simple alternative setups allowing unspecified tail shapes

[1] That is, for each bandit on $\mathcal{F}$, for each arm k with $\Delta_k > 0$, then $\mathbb{E}[N_k(T)] = o(T^\alpha)$ for all $\alpha \in (0,1]$.

but avoiding "mass leakage" to infinity, for instance with mild conditions linking the quantiles and the means of the distributions. We consider in this paper *light-tailed* distributions (see definition in Appendix A.1). This problem is already non-trivial, so we let possible extensions for heavy-tail distributions for future work (e.g with tools like median-of-means, see (Bubeck et al., 2013)).

Outline In the novel settings we consider, we want algorithms that require the smallest level of knowledge on the tails of distributions. To this extent, the *Non-Parametric Thompson Sampling* (NPTS, Riou and Honda (2020)) algorithm is a good candidate, considering how little knowledge it requires to reach asymptotic optimality for bounded distributions with known bounds. We provide a generalization of NPTS that we call *Dirichlet Sampling* (DS): we combine the core elements of NPTS and a duel-based framework inspired by (Chan, 2020), introducing data-dependent exploration bonuses. We present the resulting algorithm and detail the technical motivations of the duel-based approach in Section 2. We then introduce in Section 3 a first regret decomposition of DS algorithms under general assumptions, and the technical results that allow to fine-tune the algorithm for different families (see Section 3.1). We provide three instances of DS algorithms, their settings and regret guarantees in Section 3.2: *Bounded Dirichlet Sampling* (BDS) tackles bounded distributions with possibly unknown upper bounds, *Robust Dirichlet Sampling* (RDS) has a slightly larger than logarithmic regret for any unspecified *light-tailed* unbounded distributions, making it a competitor of the Robust-UCB algorithm of Ashutosh et al. (2021). Last, under a mild quantile condition, we prove a logarithmic regret for unbounded distributions using the *Quantile Dirichlet Sampling* algorithm (QDS). Finally, we study in Section 4 a use-case in agriculture using the DSSAT simulator (see Hoogenboom et al. (2019)), which naturally faces all the questions (robustness, model specification) that motivate this work and shows the merit of DS over state-of-the-art methods for this problem.

2 Dirichlet Sampling Algorithms

In this section we introduce Dirichlet Sampling, a strategy that aims at generalizing the Non-Parametric Thompson Sampling algorithm of Riou and Honda (2020) outside the scope of bounded distributions with a known support upper bound. For this purpose, we build an adaptive strategy in a duel-based framework, already used in sub-sampling based algorithms like SSMC (Chan, 2020).

Background NPTS is itself the adaptation of the Beta-Bernoulli Thompson sampling for distributions with bounded support $[b, B]$, where B is known. The authors first introduce Multinomial Thompson Sampling: the Beta-Bernoulli prior/posterior is naturally replaced by Dirichlet distributions when distributions are multinomial with known support. They then depart from the Bayesian framework to propose NPTS, an index strategy where the index of each arm is a *random re-weighting* of their observations, augmented by an *exploration bonus*. The weights are drawn from the Dirichlet distribution $\mathcal{D}_n = \text{Dir}((1, \dots, 1))$ for n data, which is the uniform distribution on the simplex $\mathcal{P}^n = \{w \in [0,1]^n : w^t 1 = 1\}$ and matches the Bayesian posterior for multinomial arms. The exploration bonus is simply the known upper bound of the support, and avoids under-exploration of potentially "unlucky" good arm. We provide further explanations on the Dirichlet distribution and NPTS respectively in Appendix C.1 and A.3.

The simplicity of NPTS and its strong theoretical guarantees are appealing for further generalization. As we fully depart from the Bayesian approach, considering other exploration bonuses, we derive a new family of algorithms under the name of *Dirichlet Sampling*. We keep the two principles of re-weighting the observations using a Dirichlet distribution and the exploration aid, and explore how to apply them to more general (e.g unbounded) distributions. In particular, we allow in DS some pre-processing of the observations before re-weighting (see section 3.1 and 3.2) and motivate in Section 3.1 the use of a *data-dependent* bonus, that use information from several arms. The complexity introduced by such bonus in the analysis requires a change of algorithm structure, dropping the index policy for a *leader vs challenger* approach (Chan, 2020).

Round-based algorithm We define a round as a step of the algorithm at the end of which a set of (possibly several) arms are selected to be pulled. Let $\mathcal{A}_r \subset \{1, \dots, K\}$ be the subset of the arms pulled at the beginning of a round r, we call T-round regret the quantity

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{r=1}^{T}\sum_{k=1}^{K} \Delta_k \mathbb{1}(k \in \mathcal{A}_r)\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)] \ , \tag{3}$$

where we slightly change the definition of N_k (compared with 1) to $N_k(T) = \sum_{r=1}^{T} \mathbb{1}(k \in \mathcal{A}_r)$. We consider the T-round regret for simplicity, as it is a simple upper bound of the regret after T pulls. At the beginning of each round we define a reference arm (leader), and then organize pairwise comparisons called *duels* between this arm and the other arms (challengers). The leader is chosen as the arm with largest sample size,

$$\ell^r \in \operatorname*{argmax}_{k \in \{1,\dots,K\}} N_k(r) \ ,$$

where ties are broken first in favor of the best arm, then with a random choice. A major motivation of this choice is that the leader will have a sample size that is *linear* in the number of rounds, as at least one arm is chosen at each round. This ensures strong statistical properties that we will exploit to design the exploration bonus of DS strategies. Randomizing the index of the leader is also unnecessary: it competes against each challenger with its *empirical mean*. We also dismiss all the arms k that satisfy $N_k(r) = N_\ell(r)$ with the same argument. These choices have a practical interest as they avoid the computation time of drawing the largest weight vectors. We believe this can be an alternative of independent interest to computationally intensive index policies.

Challenger's index We fix an index that is not dependent on the round, but only on the history of the challenger and the leader available at this round, that we denote respectively $\mathcal{X} = (X_1, \dots, X_n)$, $\mathcal{Y} = (Y_1, \dots, Y_N)$ for simplicity of notations. We denote $\mu : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ the function that computes the average of a set of observations. The duel can includes two steps, and the challenger wins if

1. $\mu(\mathcal{X}) \geq \mu(\mathcal{Y})$ (first attempt with empirical mean), or
2. $\widetilde{\mu}(\mathcal{X}, \mathcal{Y}) \geq \mu(\mathcal{Y})$, where $\widetilde{\mu} : \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ denotes the chosen DS index.

We summarize in Algorithm 1 the steps of the DS algorithm for a generic index $\widetilde{\mu}$, that we completely detail in Appendix A.2. The algorithm is part of the DS family if this index includes Dirichlet re-weighting of the observations augmented by an exploration bonus. However, the algorithm structure in Algorithm 1 could be combined with any randomized index, which is of independent interest as we will see in Section 3. Before this, we show a canonical example of Dirichlet Sampling index as a direct generalization of the NPTS index with a data-depend (instead of fixed) bonus $B(\mathcal{X}, \mathcal{Y})$,

$$\widetilde{\mu}(\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{n} w_i X_i + w_{n+1} B(\mathcal{X}, \mathcal{Y}) \ \ , \ w = (w_1, \dots, w_{n+1}) \sim \mathcal{D}_{n+1} \, .$$

In the next section we study the theoretical properties of Dirichlet Sampling algorithms, and discuss the choice of the index $\widetilde{\mu}$ for different families of distributions.

Algorithm 1 Generic Dirichlet Sampling Bandit Algorithm

Input: K arms, horizon T, DS index $\widetilde{\mu}$
Init.: $t = 1, r = 1, \forall k \in \{1, ..., K\}$: $\mathcal{X}_k = \{X_1^k\}$, $N_k = 1$; ▷ `Draw each arm once`
while $t < T$ **do**
  $\mathcal{A} = \{\}$; ▷ `Arm(s) to pull at the end of the round`
  $\ell = \text{Leader}((\mathcal{X}_1, N_1), \dots, (\mathcal{X}_k, N_k))$; ▷ `Choose a Leader`
  for $k \in \{1, \dots, K\} : N_k < N_\ell$ **do**
    if $\max(\mu(\mathcal{X}_k), \widetilde{\mu}(\mathcal{X}_k, \mathcal{X}_\ell)) \geq \mu(\mathcal{X}_\ell)$ **then**
      $\mathcal{A} = \mathcal{A} \cup \{k\}$; ▷ `Play the duels`

Draw arms from $|\mathcal{A}|$ if $\mathcal{A}$ is non-empty, else draw arm ℓ.
Update $t, r, (N_k)_{k \in \{1,\dots,K\}}, (\mathcal{X}_k)_{k \in \{1,\dots,K\}}$. ; ▷ `Collect Reward(s) and update data`

3 Regret Analysis and Technical Results

In this section, we analyze the regret of DS algorithms. We first derive a general regret decomposition for any index $\widetilde{\mu}$ that holds thanks to the duel-based structure. We then introduce several properties of Dirichlet sampling, that theoretically guide proper tuning of a DS index. We finally instantiate DS for three different problems and provide regret bounds in these settings. Starting with the regret

decomposition, we exhibit general conditions to ensure guarantees that are independent on the index and the run of the bandit algorithm. Allowing a different family of distribution $\mathcal{F}_k$ for each arm k, the first one concerns the concentration of the mean of each distribution.

Condition 1 (C1) [Concentration] For all $\nu_k \in \mathcal{F}_k$, there exists a good rate function I_k satisfying $I_k(x) > 0$ for $x \neq \mu_k$ and for all $x > \mu_k$, $y < \mu_k$, and any i.i.d sequence $X_1, \dots, X_n$ drawn from ν_k

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq x\right) \leq e^{-nI_k(x)} \ , \text{ and } \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq y\right) \leq e^{-nI_k(y)} \ . \tag{4}$$

This hypothesis is standard in the bandit literature, and is for instance satisfied by any *light-tailed* distributions. We refer to (Dembo and Zeitouni, 2010) for techniques to derive such functions.

We now provide an upper bound on the round-regret presented in Section 2 for Algorithm 1. This upper bound exhibits two non-constant terms, that represent 1) the expected number of pulls of sub-optimal arms when the best arm is the leader, and 2) the cost of under-exploration of the best arm. These terms are, to some extent, present in the regret bounds of most randomized algorithms. Our result is in this sense similar to Theorem 1 of Kveton et al. (2019b). Theorem 3.1 generalizes it for an index that can depend on the history of two arms instead of one, which is a result of independent interest (indeed, the index is not specified at this step of the analysis). To simplify the notations we consider that there is only one optimal arm and, without loss of generality, that $\forall k > 1, \mu_k < \mu_1$. Furthermore, for simplicity we write the following theorem for an index $\widetilde{\mu}(\mathcal{X}, \mu)$, that only uses the mean of the leader. The same result holds for any index using statistics on the leader's history that have concentration properties similar to (C1) (e.g possibly quantiles, variance, etc) with slight adaptations of the proof.

Theorem 3.1 (Generic regret decomposition of DS strategies)**.** *Consider a bandit model $\nu = (\nu_1, \dots, \nu_K)$, where all distributions in ν satisfy (C1). Then for any DS index the expected number of pulls of each arm $k \in \{2, \dots, K\}$ is upper bounded for each $\varepsilon \in [0, \Delta_k)$ by*

$$\mathbb{E}\left[N_k(T)\right] \leq n_k(T) + B^k_{T,\varepsilon} + C^k_{\nu,\varepsilon} \ ,$$

where $n_k(T) = \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \ell^r = 1)\right]$, $C^k_{\nu,\varepsilon}$ is independent on T and, denoting $\mathcal{X}_n$ the set of n first observations of arm 1,

$$B^k_{T,\varepsilon} = \sum_{j=2}^{K} \sum_{n=1}^{\lceil 2\log(T)/I_1(\mu_k+\varepsilon)\rceil} \sup_{\mu \in [\mu_j - \varepsilon, \mu_j + \varepsilon]} \mathbb{E}_{\mathcal{X}_n}\left[\frac{\mathbb{1}\left(\mu(\mathcal{X}_n) \leq \mu\right)}{\mathbb{P}(\widetilde{\mu}(\mathcal{X}_n, \mu) \geq \mu)}\right] \ .$$

The details of the proof of this result are to be found in Appendix B. The proof follows the general outline of Chan (2020), and makes all the components of $C^k_{\nu,\varepsilon}$ explicit. In particular, they do not depend on the choice of the index but only on the rate functions and the means of each arm. The term $B^k_{T,\varepsilon}$ corresponds to the case where arm 1 is not leader and under-sampled.

Theorem 3.1 is formulated to be as general as possible, and we show in Corollary 3.1.1 that we will look for the first-order term of the regret in $n_k(T)$.

Condition 2 (C2) For any $\mu < \mu_1$, and any $n_1(T) = o(\log T)$ it holds that

$$\sum_{n=1}^{n_1(T)} \mathbb{E}_{\mathcal{X}_n \sim \nu_1^n}\left[\frac{\mathbb{1}(\mu(\mathcal{X}_n) \leq \mu)}{\mathbb{P}_{w \sim \mathcal{D}_{n+1}}\left(\widetilde{\mu}(\mathcal{X}, \mu) \geq \mu\right)}\right] = o(\log T) \ .$$

Corollary 3.1.1 (Conditions for controlled regret)**.** *If condition (C1) and (C2) holds for the DS index on the families of distribution $(\mathcal{F}_k)_{k \in \{1,\dots,K\}}$, the regret of the DS algorithm satisfies*

$$\mathcal{R}_T \leq \sum_{k=2}^{K} \Delta_k n_k(T) + o(\log T) \ .$$

Up to this point this result is quite abstract, but this standardized analysis allows us to instantiate the Dirichlet Sampling algorithm on different class of problems and calibrate it in order to ensure condition (C2) holds and to make $n_k(T)$ explicit. In particular if $n_k(T) = \mathcal{O}(\log T)$, we recover the logarithmic regret. In the next section, we present technical results to justify calibrations of the DS index for several kind of families.

3.1 Technical tools: boundary crossing probability of a DS index

In this section, we highlight some key properties of a sum of random variables re-weighted by a Dirichlet weight vector that help us suggest a sound tuning of the bonus $\mathcal{B}(\mathcal{X}, \mathcal{Y})$ for different kind of families. We then detail such tuning.

Boundary crossing probability (BCP) We consider a set of $n+1$ observation points $\mathcal{X} = (X_1, \dots, X_{n+1}) \subset \mathbb{R}^{n+1}$. (Intuitively, n points are samples from a challenger arm, and one point corresponds to the added bonus). Then, for any $\mu \in \mathbb{R}$, we introduce the following "Boundary Crossing Probability" (BCP) term, conditional on $\mathcal{X}$

$$[\text{BCP}] := \mathbb{P}_{w \sim \mathcal{D}_{n+1}} \left(\sum_{i=1}^{n+1} w_i X_i \geq \mu \right) ,$$

where we recall that $\mathcal{D}_{n+1}$ is the Dirichlet distribution with parameter $(1, \dots, 1)$ of size $n+1$, i.e the uniform distribution on the $(n+1)$-simplex. We emphasize that here $\mathcal{X}$ is considered fixed, and the only source of randomness comes from the weights w. When all observations are distinct this expression has a closed form, which is unfortunately untractable in the proof, as discussed in Appendix C.2. This quantity is of much interest as both the growth of $n_k(T)$ and (C2) can be derived from respectively upper and lower bounds for the BCP. Lemma 14 and Lemma 15 in (Riou and Honda, 2020) provide such bounds, resorting to classical concentration results and properties of the Dirichlet distribution that we recall in Appendix C.1 and C.2, and complete with additional technical results. The lower bounds suggest non-trivial tuning of the bonus. We first exhibit a necessary condition when the bonus is not allowed to depend on the set of observations $\mathcal{X}$.

Lemma 3.2 (Necessary condition with a data-independent bonus)**.** *Consider a fixed bonus $B(\mathcal{X}, \mu) = B(\mu)$, and a distribution F (with CDF also denoted F). If Condition (C2) holds then*

$$B(\mu) > \mu + \frac{1}{1 - F(\mu)} \mathbb{E}_F \left[(\mu - X)_+ \right] .$$

This result is obtained using a "worst-case" scenario when all observations are below the threshold μ. Hence, it does not cover all possible trajectories, yet it suggests to investigate the properties of bonuses with a similar form. Since the right-hand side of the inequality requires a knowledge on the arms distributions that we would like to avoid, we use an empirical estimator for the expectation. This suggests to introduce some parameter ρ and data-dependent bonuses of the form

$$B(\mathcal{X}, \mu, \rho) = \mu + \rho \times \frac{1}{n} \sum_{i=1}^{n} (\mu - X_i)^+ , \tag{5}$$

We then tune ρ assuming an hypothesis on some upper quantile of the arm distribution, which is much less constraining than assuming knowledge of the shape of the entire tail. In all DS algorithms we propose (see next section), we use Equation 5 as the basis for defining the appropriate bonus. Finally, we provide in lemma 3.3 a novel lower bound on the BCP that reveals that in the general case of unbounded distributions, without further processing of the data, DS cannot catch a logarithmic regret when the maximum of the data tends to $+\infty$ at some too fast rate $g(n)$.

Lemma 3.3 (Lower bound for the BCP)**.** *Consider a set $\mathcal{X} = (X_1, \dots, X_{n+1}) \in \mathbb{R}^{n+1}$, and assume that $\overline{\mathcal{X}} = \max_{i \in \{1, \dots, n+1\}} X_i \geq g(n)$ for some function g. Denoting $\bar{\Delta}_n^+ = \frac{1}{n} \sum_{i=1, X_i < \overline{\mathcal{X}}}^{n+1} (\mu - X_i)^+$ the empirical positive gap, it holds that*

$$\mathbb{P}_{w \sim \mathcal{D}_{n+1}} \left(\sum_{i=1}^{n+1} w_i X_i \geq \mu \right) \geq \exp \left(-n \frac{\bar{\Delta}_n^+}{g(n) - \mu} \right) .$$

In particular, we see in this expression that $g(n)$ may hinder the exponential rate in n. In the next section we discuss three examples of DS algorithms and their theoretical guarantees.

3.2 Theoretical guarantees for Dirichlet Sampling algorithms

Building on the results from previous the section, we now instantiate the DS algorithms for three bandits problems, namely bounded distributions, general light-tailed distributions, and light-tailed

distributions under an additional quantile condition. In this section we denote $B(\mathcal{X}, \mu, \rho)$ the bonus defined in Equation 5 for a set $\mathcal{X}$, a mean μ and some parameter ρ. We detail each algorithm and their components in Appendix A.2 and the proofs of the three theorems in Appendix D. In all cases, the proof consists in showing that (C1) and (C2) hold for each proposed algorithms in the settings they tackle and deriving an expression for $n_k(T)$.

Optimality for bounded distributions Let $\mathcal{F}_{[b,B]}$ be the set of distributions supported in $[b, B]$, and $\mathcal{F}_{[b,B]}^{\gamma,p} = \{\nu \in \mathcal{F}_{[b,B]} : \mathbb{P}_\nu([B-\gamma, B]) \geq p\}$. Let $\overline{\mathcal{X}}$ denote the maximum of $\mathcal{X}$. We denote *Bounded Dirichlet Sampling* (BDS) the DS algorithm with bonus defined as follows for some $\rho > 0$

$$B(\mathcal{X}, \mu) := \begin{cases} B & \text{if } \nu \in \mathcal{F}_{[b,B]} \text{ with } \textit{known } B, \quad \text{(B1)} \\ \max\{\overline{\mathcal{X}} + \gamma, B(\mathcal{X}, \mu, \rho)\} & \text{if } \nu \in \mathcal{F}_{[b,B]}^{\gamma,p} \text{ with } B \textit{ unknown} \text{ but known } p, \gamma. \quad \text{(B2)} \end{cases}$$

Theorem 3.4 (Optimality of BDS)**.** *BDS is asymptotically optimal under case (B1), and under case (B2) it satisfies for any $\rho \geq -1/\log(1-p)$ and any $k \in \{2, \dots, K\}$*

$$\mathbb{E}[N_k(T)] \leq \frac{\log(T)}{\mathcal{K}_{\inf}^{B_{\rho,\gamma}}(\nu_k, \mu_1)} + O(1) \ ,$$

where $B_{\rho,\gamma} = \max(B + \gamma, \mu_1 + \rho \mathbb{E}_{\nu_k}[(\mu_1 - \mu_k)_+])$.

Hence, we show that the knowledge of the upper bound can be replaced by an alternative hypothesis on the mass of the distribution in a fixed neighborhood of this bound. We provide in Appendix E examples of applications of this setting.

Robust regret for light-tailed distributions We call *Robust Dirichlet Sampling* (RDS) the algorithm with bonus $B(\mathcal{X}, \mu, \rho_n)$, where ρ_n depends on the sample size n. We prove that RDS achieves a robust sub-linear regret bound when each arm comes from any *unknown* light-tailed distribution, that we define as the family $\mathcal{F}_\ell = \{\nu \in \mathcal{F}_{(-\infty,+\infty)} : \exists \lambda_\nu > 0, \forall \lambda \in [-\lambda_\nu, \lambda_\nu], \mathbb{E}_\nu[\exp(\lambda X)] < +\infty\}$.

Theorem 3.5 (Robust regret bound for RDS)**.** *Let $\nu = (\nu_1, \dots, \nu_K)$ a bandit model satisfying $\nu_k \in \mathcal{F}_\ell$ for all k. Consider an increasing sequence $(\rho_n)_{n\in\mathbb{N}}$ with $\rho_n \to +\infty$, $\rho_n = o(n)$. Then, for T large enough the expected number of pull of any sub-optimal arm k in RDS is upper bounded by*

$$\mathbb{E}[N_k(T)] \leq n_k^{\eta,\varepsilon_0}(T) + \mathcal{O}(1) \ ,$$

where for any $\eta \in (0, 1], \varepsilon_0 > 0$, $n_k^{\eta,\varepsilon_0}(T)$ is the sequence satisfying

$$n_k^{\eta,\varepsilon_0}(T) = \frac{\log T}{\eta(\Delta_k - \varepsilon_0)}(M_{k,n_k(T)} - \mu) \ \ , \textit{ with } M_{k,n} = \max\left\{F_k^{-1}\left(\exp\left(-\frac{1}{n^2(\log n)^2}\right)\right), \rho_n\right\}.$$

The sequence $M_{k,n}$ is in fact a large probability upper bound of the maximum of n observations from F_k, that we discuss in Appendix D. For many usual distributions this formula simply provides bounds in $\mathcal{O}(\log(T) g_k(\log T))$, where $g_k(n)$ is typically the expectation of the maximum of n data from ν_k in n. For instance, $g_k(n) = \mathcal{O}(\sqrt{\log n})$ for sub-gaussian tails and $g_k(n) = \mathcal{O}(\log n)$ for exponential tails. We see that in these two cases the regret is in $\mathcal{O}(\log(T) \log\log(T))$, which is not much more than $\mathcal{O}(\log T)$. This larger-than-logarithmic rate is a direct consequence of Lemma 3.3. Interestingly, as ρ_n grows, the algorithm is adapting to the true nature of the tails even though it does not use any information on it. It is in that sense that this algorithm is *robust*.

Logarithmic regret for unbounded distributions Theorem 3.5 shows that we cannot expect a logarithmic regret for unbounded distributions with the vanilla re-weighting scheme. It is necessary to find a way to "truncate" the tails, which should be feasible if some summary tail statistic can be "representative" of the distribution outside a given range. Considering a base family $\mathcal{F} \subset \mathcal{F}_{[b,+\infty)}$, we consider the subset of $\mathcal{F}$ satisfying

$$\mathcal{F}_{[b,+\infty)}^{\alpha} = \{\nu \in \mathcal{F} : \forall \mu > \mathbb{E}_\nu(X), \mathcal{K}_{\inf}^{\mathcal{F}}(\nu, \mu) \geq \mathcal{K}_{\inf}^{\mathfrak{M}_k}(\mathcal{T}_\alpha(\nu), \mu)\} \ ,$$

where $q_{1-\alpha}(\nu)$ denotes the quantile $1-\alpha$ of ν, $\mathfrak{M}_k = \max\{q_{1-\alpha}(\nu), \mu + \rho \mathbb{E}_\nu[(\mu - X)^+]$ and $\mathcal{T}_\alpha : \mathcal{F} \to \mathcal{F}_{[-\infty, M_{\nu,\mu,\rho}]}$ is the truncation satisfying $\mathcal{T}_\alpha(\nu)(A) = A$ for any $A \subset [b, q_{1-\alpha}(\nu)]$ and $\mathcal{T}_\alpha(\nu)(\{x\}) = \alpha \mathbb{1}(x = \mathbb{E}_\nu[X | X > q_{1-\alpha}(\nu)])$ for any $x > q_{1-\alpha}(\nu)$. In short, the motivation to

consider this family is that finding a distribution of mean μ that fits a constrained tail can be arbitrarily costly in terms of $\mathcal{K}_{\inf}$ due to the mass leakage phenomenon at infinity, much more so than fitting a summary statistic of said tail like the average above a given quantile, also known as *conditional Value-at-Risk*. Following these intuitions, we propose the *Quantile Dirichlet Sampling* (QDS) for the family $\mathcal{F}^{\alpha}_{[b,+\infty)}$. Given an empirical distribution $\widehat{\nu}$, the algorithm runs the Dirichlet re-weighting on $\mathcal{T}_{\alpha'}(\widehat{\nu})$ with bonus $B(\mathcal{X}_n, \mu, \rho)$ for some $\rho > 0$ and $\alpha' < \alpha$ (larger quantile than in the definition of the family). We detail these steps in A.2. Note that the parameter of the Dirichlet distribution becomes $(1, \dots, 1, n_{\alpha'})$, where $n_{\alpha'}$ is the exact number of observations above the quantile $1 - \alpha'$ (asymptotically $n_{\alpha'} \sim \alpha' n$). This way, the QDS index remains unbiased w.r.t the true empirical distribution. We show in Theorem 3.6 that the regret of QDS is logarithmic on $\mathcal{F}^{\mathcal{F},\alpha}_{[b,+\infty]}$ for a finite (unknown) b and a known α.

Theorem 3.6 (Logarithmic Regret of QDS)**.** *Consider a bandit model $\nu = (\nu_1, \dots, \nu_K)$ satisfying $\forall k$, $\nu_k \in \mathcal{F}^{\mathcal{F},\alpha}_{[b,+\infty)}$ for some base family $\mathcal{F}$, $b > -\infty$ (lower-bounded support) and a known $\alpha > 0$. Then, QDS with any parameters $\alpha' < \alpha$ and $\rho \geq (1 + \alpha')/\alpha'^2$ satisfies*

$$\mathbb{E}[N_k(T)] \leq \frac{\log T}{\mathcal{K}^{\mathfrak{M}_k}_{\inf}(\mathcal{T}_\alpha(\nu_k), \mu_1)} + \mathcal{O}(1) \ ,$$

where all variables are defined in the introducing paragraph.

This result is of particular interest as it captures the continuum between bounded and light-tailed distributions. In our opinion, it sheds new light on the interpretation of infeasability results of e.g Ashutosh et al. (2021): logarithmic regret can be achieved *without specifying the tail with precise parameters*, but a simple quantile condition is required to avoid pathological distributions that makes little sense in practice (e.g very small mass at a very large value). We further discuss this condition in D.5 and provide examples of families for which it holds (e.g exponential distributions).

Remark 3.7. *Our techniques of proof does not allow $b = -\infty$, even though b does not need to be known by the algorithm. We believe that different theoretical tools could allow to prove a logarithmic regret of QDS in this case, possibly with a symmetric treatment of the two tails.*

4 Application in a crop-farming environment

We consider a practical decision-making problem using the DSSAT[2] simulator (Hoogenboom et al., 2019). Harnessing more than 30 years of expert knowledge, this simulator is calibrated on historical field data (soil measurements, genetics, planting date...) and generates realistic crop yields. Such simulations are used to explore crop management policies *in silico* before implementing them in the real world, where their actual effect may take months or years to manifest themselves. More specifically, we model the problem of selecting a planting date for maize grains among 7 possible options, all else being equal, as a 7-armed bandit. The resulting distributions incorporate historical variability as well as exogenous randomness coming from a stochastic meteorologic model. We illustrate this in Figure 1 with the histogram of four of these distributions, computed on 10^6 samples. They are typically right-skewed, multimodal and exhibit a peak at zero corresponding to years of poor harvest, hence they hardly fit to a convenient parametric model (e.g SPEF/sub-Gaussian. . .).

A natural choice for the learner would be to use algorithms adapted for bounded distributions with known support. Indeed, one could argue that crop yields are fundamentally bounded by a very large value, that can be provided with some expert knowledge. However this method can have limits when the upper bound cannot be estimated accurately (few data, new environment, . . .), as a conservative bound can have a cost on the regret. For this reason, we believe that the novel Dirichlet Sampling algorithms we introduce in Section 3.2 are a good alternative choice for this problem. In particular, the three algorithms we propose in this paper are relevant in this setting: BDS keeps the bounded-support hypothesis but introduces the possible uncertainty on the bound, while the light-tailed hypothesis of RDS and the quantile condition of QDS look reasonable (regarding that they are the only realistic alternative to bounded support). In Figure 2 we compare DS algorithms to empirical IMED (Honda and Takemura, 2015) and NPTS (Riou and Honda, 2020), with two upper bounds: 1) the "exact"

[2] *Decision Support System for Agrotechnology Transfer* is an open-source project maintained by the DSSAT Foundation, see https://dssat.net/

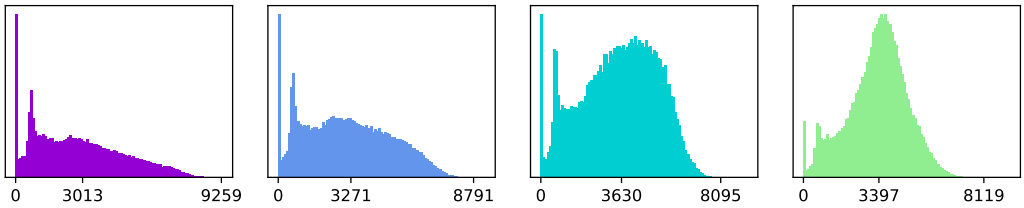

Figure 1: Distribution of simulated dry grain yield (kg/ha) for four out of seven different planting dates. Reported on the x-axis are the distribution minimum, mean and maximum values.

upper bound is provided looking at the maximum of all historical data collected (left figure), and 2) the algorithms use a conservative estimate with a value 1.5 times larger than the previous one (right figure). We report the performance of IMED and NPTS as they were strongest baselines on this problem but report figures with other competitors in Appendix E. For BDS we choose the parameters $\rho = 4, \gamma = 3500$, corresponding to $p \approx 20\%$ in the hypothesis of Theorem 3.4, which is conservative in our example. For RDS, we choose $\rho_n = \sqrt{\log n}$, which enters into the theoretical framework of Theorem 3.5. Finally, for QDS we set $\rho = 4$ to be able to compare with BDS and a quantile 95%. Our results show that Dirichlet Sampling algorithms achieve very similar regret to their competitors when the latter are allowed to use the "exact" upper bound, and compare favorably when they use a conservative estimate (1.5 times larger, right). Our code is available with the paper[3]. To further

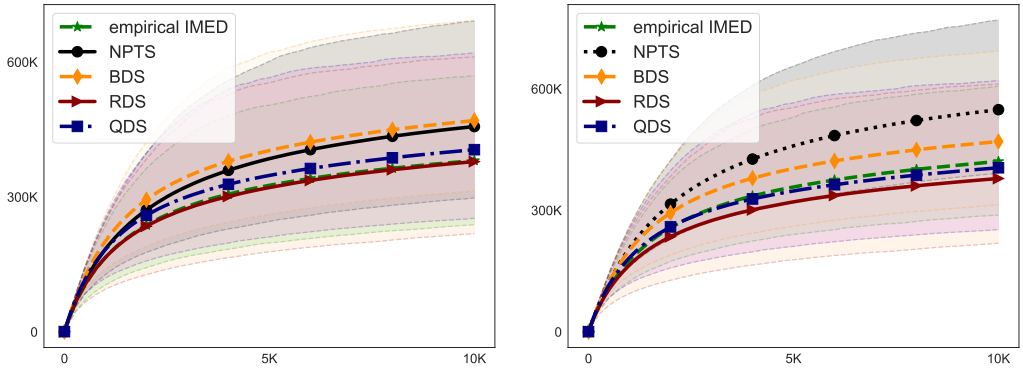

Figure 2: Average regret on 5000 simulations and horizon $T = 10^4$. Dashed lines correspond to 5%-95% regret quantiles. Empirical IMED and NPTS are run with exact upper bounds around 1.5×10^4 kg/ha (left) and the conservative upper bound 1.5×10^4 kg/ha (right).

illustrate the properties of DS algorithms, we perform additional experiments on synthetic examples. Due to space limits, we present our results in Appendix E. First, we test the *sensitivity* of DS w.r.t its parameters, and their impact on the performance of the algorithms is very smooth. Then, we show the merits of DS algorithms in case of *model misspecification* in various examples where other competitors perform poorly. Finally, we consider the case of *Gaussian mixtures*, which are common distributions in practice that fit the scope of DS but not that of usual bandit algorithms.

Perspectives In this paper we proposed the Dirichlet Sampling algorithms, that generalizes an optimal Thompson Sampling strategy for bounded distributions to *light-tailed* families. In our opinion, these new algorithms are appealing for the practitioner because, even if the analysis is quite technical: 1) our theoretical results show strong guarantees under different settings, 2) DS algorithms are simple to implement and achieve strong practical performance, and 3) they provide alternative robust ways to tackle unbounded distributions in bandit problems, outside well-parametrized tails. Interesting future directions include for instance considering DS algorithms for *heavy tail* distributions, and finding new theoretical results on the *Boundary Crossing Probability* of Section 3.1 to design sharper bonuses for general families of distributions motivated by real use-cases.

[3]a link will be provided in the final version of the paper

References

R. Agrawal. Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

S. Agrawal, W. M. Koolen, and S. Juneja. Optimal best-arm identification methods for tail-risk measures. *arXiv preprint arXiv:2008.07606*, 2020.

K. Ashutosh, J. Nair, A. Kagrecha, and K. Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2021.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

A. Baransi, O.-A. Maillard, and S. Mannor. Sub-sampling for multi-armed bandits. In *Proceedings of the European Conference on Machine Learning*, page 13, 2014.

D. Baudry, E. Kaufmann, and O.-A. Maillard. Sub-sampling for efficient non-parametric bandit exploration. In *Advances in Neural Information Processing Systems*, 2020.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

H. P. Chan. The multi-armed bandit problem: An efficient nonparametric solution. *The Annals of Statistics*, 48(1):346–373, 2020.

Y. Cho and E. Cho. The volume of simplices clipped by a half space. *Applied mathematics letters*, 14 (6):731–735, 2001.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 95. 2010.

H. Hadiji and G. Stoltz. Adaptation to the range in k-armed bandits. *arXiv preprint arXiv:2006.03378*, 2020.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79, 2010.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756, 2015.

G. Hoogenboom, C. Porter, K. Boote, V. Shelia, P. Wilkens, U. Singh, J. White, S. Asseng, J. Lizaso, L. Moreno, et al. The dssat crop modeling ecosystem. *Advances in crop modelling for a sustainable agriculture*, pages 173–216, 2019.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for one-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, 2013.

B. Kveton, C. Szepesvári, M. Ghavamzadeh, and C. Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, 2019a.

B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, T. Lattimore, and M. Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610. PMLR, 2019b.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

I. Osband and B. V. Roy. Bootstrapped Thompson sampling and deep exploration. *CoRR*, abs/1507.00300, 2015.

C. Riou and J. Honda. Bandit algorithms based on Thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pages 777–826. PMLR, 2020.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

C.-H. Wang, Y. Yu, B. Hao, and G. Cheng. Residual bootstrap exploration for bandit algorithms. *arXiv preprint arXiv:2002.08436*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We present in this work a bandit algorithm and show an application in a decision-making problem in agriculture.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See for instance Section 3.2 where we introduce our main result.
   (b) Did you include complete proofs of all theoretical results? [Yes] To respect the page limit all the proofs are presented in Appendix C.2 (results on the BCP), B (proof of Theorem 3.1), D (proof of Theorem 3.4, 3.5, 3.6).
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify the parameters of the distributions and the parameters of the bandit algorithms in the code, and explain their choice in section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Confidence intervals are represented in all our figures.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Our code does not require powerful computing ressources, experiments are replicable with a laptop.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We provided references for the DSSAT environment.
   (b) Did you mention the license of the assets? [Yes] DSSAT is open-source.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]