Adaptive Sampling for Estimating Probability Distributions

Shubhanshu Shekhar [1] **Tara Javidi** [1] **Mohammad Ghavamzadeh** [2]

Abstract

We consider the problem of allocating a fixed budget of samples to a finite set of discrete distributions to learn them uniformly well (minimizing the maximum error) in terms of four common distance measures: ℓ_2^2, ℓ_1, f-divergence, and separation distance. To present a unified treatment of these distances, we first propose a general *optimistic tracking algorithm* and analyze its sample allocation performance w.r.t. an oracle. We then instantiate this algorithm for the four distance measures and derive bounds on their regret. We also show that the allocation performance of the proposed algorithm cannot, in general, be improved, by deriving lower-bounds on the expected deviation from the oracle allocation for any adaptive scheme. We verify our theoretical findings through some experiments. Finally, we show that the techniques developed in the paper can be easily extended to learn some classes of continuous distributions as well as to the related setting of minimizing the average error (rather than the maximum error) in learning a set of distributions.

1. Introduction

Consider the problem in which a learner must allocate n samples among K discrete distributions to construct *uniformly good* (minimizing the maximum error) estimates of these distributions in terms of a distance measure D. Depending on D, certain distributions may require much fewer samples than the others to be estimated with the same precision. The optimal sampling strategy for a given n requires knowledge of the true distributions. The goal of this paper is to design adaptive allocation strategies that converge to the optimal strategy, oblivious to the true distributions.

The problem described above models several applications which are not captured by existing works. Here, we describe some such applications. **(1) Opinion Polling:** Suppose there are K groups of voters who have different preference distributions over l candidates in an election. While some groups might heavily favour a single candidate, others might be indifferent, resulting in a more uniform distribution over the set of candidates. In this setting, how should the polling agency allocate its sampling budget? Intuitively, more samples should be allocated to the indifferent voter groups to construct uniformly good estimates of their preference distributions. **(2) Compression of text files:** Given a sampling budget of n bytes, consider the problem of designing codes with minimum average length for text files in K different languages. Since different languages may have different symbol frequencies, this can be formulated as learning K distributions uniformly well in terms of certain f-divergences. **(3) Learning MDP model:** In many sequential decision-making problems, the agent's interaction with the environment is modeled as a Markov decision process (MDP). In these problems, it is often important to accurately estimate the dynamics (i.e., the transition structure of the MDP), given a finite exploration budget. Learning the MDP model is equivalent to estimating $S \times A$ distributions, where S and A are the number of states and actions of the (finite) MDP. Therefore, assuming the existence of a known policy that can efficiently transition the MDP between any two states, the problem reduces to finding the optimal allocation of samples to these $S \times A$ distributions. Thus, the framework studied in this paper provides the first step towards solving the general problem of constructing accurate models for MDPs. The requirement of a known policy to transition between states can be relaxed by employing the techniques recently developed for efficient exploration in MDPs (e.g., Tarbouriech & Lazaric 2019; Hazan et al. 2019; Cheung 2019), which we leave for future work.

Antos et al. (2008) were the first to study the problem of learning the *mean* of K distributions uniformly well, and proposed and analyzed an algorithm based on *forced exploration* strategy. Carpentier et al. (2011) proposed and analyzed an alternative approach for the same problem, based on the UCB algorithm (Auer et al., 2002). Carpentier & Munos (2011) analyzed an optimistic policy for the related problem of *stratified-sampling*, where the goal is to learn K distributions in terms of a weighted average distance (instead of max). Soare et al. (2013) extended the optimistic strategy to the case of uniformly estimating K

[1]ECE Department, University of California, San Diego [2]Google Research. Correspondence to: Shubhanshu Shekhar <shshekha@eng.uscd.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

linearly correlated distributions. Riquelme et al. (2017) applied the optimistic strategy to the problem of allocating covariates (drawn in an i.i.d. manner from some distribution) for uniform estimation of K linear models. The prior work mentioned above have focused solely on estimating the means of distributions in squared error sense, and their analytic techniques do not extend to learning entire distributions. In this paper, we generalize the above-mentioned prior work by considering the problem of active sampling to uniformly learn K distributions in terms of pre-specified distance measures on the space probability distributions.

Overview of Results. Intuitively, the optimal allocation should equalize the expected distance between the true distribution and the resulting empirical estimate for all the K distributions. This allocation, however, may have a complex dependence on the true distribution, P_i, for $1 \leq i \leq K$. Our approach in this paper is to first identify an objective function which (i) is a good approximation of the true objective given a distance measure D, (ii) depends on the original distribution P_i through a single real-valued parameter c_i, and (iii) has a decoupled dependence on c_i and T_i. In Sec. 3, we formally define an appropriate function class $\mathcal{F}$ within which the *objective functions* for various distance measures should lie. We then propose a generic optimistic tracking strategy (Alg. 1) which addresses the trade-off in constructing better estimates of the parameter c_i, and using the existing estimates of c_i to drive the allocation towards the optimal. We also obtain a general bound on its deviation from an (approx-) oracle allocation (defined in Sec. 3). In Sec. 4, we first present a road-map for designing adaptive sampling schemes for arbitrary loss functions using the results of Sec. 3, and then specialize this to the case of four widely-used distance measures: ℓ_2^2, ℓ_1, f-divergence, and separation distance. For each distance measure, we obtain bounds on the regret of the proposed sampling scheme w.r.t. an oracle strategy. In Sec. 5, derive matching lower-bounds on the expected deviation from oracle allocation for any algorithm. Experiments with synthetic examples in Sec. 6 validate our theoretical results. Finally, we discuss how our techniques can be extended to learning some classes of continuous distributions as well as to the related problem of minimizing the average error in Sec. 7.

Technical Contributions. The results of this paper require generalizing existing techniques, as well as introducing new methods. More specifically, the proof of Theorem 1 abstracts out the arguments of Carpentier et al. (2011, Thm. 1) to deal with a much larger class of objective functions. Prior work with mean-squared error (Antos et al., 2008; Carpentier et al., 2011) required bounding the first and second moments of random sums that could be achieved by a direct application of Wald's equations (Durrett, 2019, Thm. 4.8.6). Our results on f-divergence (Thm. 7 and Lemma 9 in Appendix F) require analyzing higher moments of random sums for which Wald's equations are not applicable. Deriving the approximate objective function for separation distance involves estimating the expectation of the maximum of some correlated random variables. We obtain upper and lower bounds on this expectation in Lemma 6 by first approximating the maximum with certain sums, and then bounding the sums using a normal approximation result (Ross, 2011, Thm. 3.2).

2. Problem Setup

Consider K discrete distributions, $(P_i)_{i=1}^K$, that belong to the $(l-1)$-dimensional probability simplex Δ_l, and take values in the set $\mathcal{X} = \{x_1, \ldots, x_l\}$. Each distribution P_i is equivalently represented by a vector $P_i = (p_{i1}, \ldots, p_{il})$ with $p_{ij} \geq 0,\ \forall j \in [l]$, and $\sum_{j=1}^l p_{ij} = 1$. For any integer $b > 0$, we denote by $[b]$, the set $\{1, \ldots, b\}$. Given a budget of $n \geq K$ samples, we consider the problem of allocating samples to each of the K distributions in such a way that the maximum (over the K distributions) discrepancy between the empirical distributions (estimated from the samples) and the true distributions is minimized. To formally define this problem, suppose an allocation scheme assigns $(T_i)_{i=1}^K$ samples to the K distributions, such that $T_i \geq 0,\ \forall i \in [K]$, and $\sum_{i=1}^K T_i = n$. Also suppose that $\hat{P}_i$ is the empirical distribution with $\hat{p}_{ij} = T_{ij}/T_i$, where T_{ij} denotes the number of times the output x_j was observed in the T_i draws from P_i, and $D : \Delta_l \times \Delta_l \mapsto [0, \infty)$ is a distribution distance measure. Then, our problem of interest can be defined as finding an allocation scheme $(T_i)_{i=1}^K$ that solves the following constrained optimization problem:

$$\min_{T_1,\ldots,T_K} \max_{i \in [K]} \mathbb{E}\big[D(\hat{P}_i, P_i)\big], \quad \text{s.t.} \sum_{i=1}^K T_i = n. \tag{1}$$

We refer to the (non-integer) solution of (1) with full knowledge of $(P_i)_{i=1}^K$ as the *oracle* allocation $(T_i^*)_{i=1}^K$. It is important to note that $(T_i^*)_{i=1}^K$ ensure that the objective functions $\gamma_i(T_i) \coloneqq \mathbb{E}\big[D(\hat{P}_i, P_i)\big]$ are equal, for all $i \in [K]$. However, in practice, $(P_i)_{i=1}^K$ are not known. In this case, we refer to (1) as a *tracking problem* in which the goal is to design adaptive sampling strategies that approximate the oracle allocation using running estimates of $(P_i)_{i=1}^K$.

Choice of the Distance Measure. It is expected that the optimal allocation will be strongly dependent on the distance measure D. We study four distances: ℓ_2^2, ℓ_1 or total variation (TV), f-divergence, and *separation distance* in this paper. These distances include all those in (Gibbs & Su, 2002) that do not require a metric structure on $\mathcal{X}$. The f-divergence family generalizes the well-known KL-divergence (D_{KL}) and includes a number of other common distances, such as total variation (D_{TV}), Hellinger (D_H), and chi-square (D_{χ^2}). Applications of f-divergence include source and channel coding problems (Csiszár, 1967; 1995), testing goodness-of-fit (Gyorfi et al., 2000), and distribution estimation (Barron et al., 1992). The common f-divergences

mentioned above satisfy the following chain of inequalities: $D_{\text{TV}} \leq D_H \leq \sqrt{D_{\text{KL}}} \leq \sqrt{D_{\chi^2}}$, that define a hierarchy of convergence among these measures (Tsybakov, 2009, Eq. 2.27). The separation distance $D_s(P, Q)$ (defined formally in Sec. 4.5) arises naturally in the study of the convergence of symmetric Markov chains to their stationary distribution. More specifically, if Q is the stationary distribution of a Markov chain and $(P_t)_{t\geq 1}$ is its state distribution at time t, such that $Q = P_T$ at a random time T, then $D_s(P_t, Q) \leq \mathbb{P}(T > t)$ (Aldous & Diaconis, 1987, Sec. 3).

Choice of estimator. In this work, we fix the estimated distribution $\hat{P}_i$ to be the empirical distribution, i.e., $\hat{P}_i = [\hat{p}_{ij}]_{j=1}^{l}$ where $\hat{p}_{ij} = T_{ij}/T_i$. While the empirical estimator is known to be suboptimal in a min-max sense (Kamath et al., 2015), the additional error due to the deviation of $\mathbb{E}[D(\hat{P}_i, P_i)]$ for some of the above distances (ℓ_2^2, ℓ_1 and f-divergence) does not change the final regret obtained. For instance, for the ℓ_2^2 distance, the results of (Kamath et al., 2015) show that $\mathbb{E}[D_{\ell_2^2}(\hat{P}_i, P_i)]$ differs from the min-max value by a $\mathcal{O}\left(n^{-3/2}\right)$ term. Since this term is of the same order as the regret we derive in Theorem 2, we conclude that for this loss the regret cannot be improved by using the min-max optimal estimator estimator. Similar results can be shown for ℓ_1 distance and the $f-$divergence family as well.

Allocation Scheme and Regret. An *adaptive* allocation scheme $\mathcal{A}$ consists of a sequence of mappings $(\pi_t)_{t\geq 1}$, where each mapping $\pi_t : \left(\mathbb{N} \times (\mathcal{X} \times [K])^{t-1}\right) \mapsto [K]$ selects an arm to pull[1] at time t, based on the budget n and the history of pulls and observations up to time t. For an allocation scheme $\mathcal{A}$, a sampling budget n, and a distance measure D, we define the *risk* incurred by $\mathcal{A}$ as

$$\mathcal{L}_n(\mathcal{A}, D) = \max_{i\in[K]} \mathbb{E}\big[D(\hat{P}_i, P_i)\big]. \tag{2}$$

We denote by $\mathcal{A}^*$, the *oracle* allocation rule. The performance of an allocation scheme $\mathcal{A}$ is measured by its suboptimality or *regret* w.r.t. $\mathcal{A}^*$, i.e.,

$$\mathcal{R}_n(\mathcal{A}, D) \coloneqq \mathcal{L}_n(\mathcal{A}, D) - \mathcal{L}_n(\mathcal{A}^*, D). \tag{3}$$

Notations.[2] For $0 < \eta < 1/2$, we define the η-interior of $(l-1)$-dimensional simplex Δ_l, as $\Delta_l^{(\eta)} \coloneqq \{P \in \Delta_l \mid \eta \leq p_j \leq 1-\eta,\ \forall j \in [l]\}$. We use the Bernoulli random variable $Z_{ij}^{(s)}$ to represent the indicator that the s^{th} draw from arm i is equal to $x_j \in \mathcal{X}$. Note that for any draw s, we have $\mathbb{E}[Z_{ij}^{(s)}] = p_{ij}$. For any $t \in [n]$, we define $W_{ij,t} = \sum_{s=1}^{t} \tilde{Z}_{ij}^{(s)}$, where $\tilde{Z}_{ij}^{(s)} \coloneqq Z_{ij}^{(s)} - p_{ij}$ is a centered Bernoulli variable. We also note that several terms such as φ, A, B, and $\tilde{e}_n$ (to be introduced in Sec. 3) are overloaded for different distance measures. For instance, we use φ for both ℓ_1 and KL-divergence, instead of writing $\varphi^{(\ell_1)}$ and $\varphi^{(\text{KL})}$. The meaning should be clear from the local context.

[1] Each distribution can be considered as an arm, and thus, we use the terms sampling from a distribution and pulling an arm interchangeably throughout the paper.

[2] See Table A.1 in App. A for a list of all the notations used.

3. General Allocation via Optimistic Tracking

Before proceeding to the analysis of problem (1) for specific distance measures, we first study an abstract yet more stylized class of problems similar to (1), where the dependency of the objective (loss) functions on the distribution parameter versus number of allocated samples can be explicitly decoupled. In particular, let us consider the problem in which the objective functions satisfy certain regularity conditions that we define next.

Definition 1. *We use $\mathcal{F}$ to denote the class of functions $\varphi : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ satisfying the following properties:* **1)** *$\varphi(\cdot, T)$ is concave and non-decreasing for all $T \in \mathbb{R}$,* **2)** *$\varphi(c, \cdot)$ is convex and non-increasing for all $c \in \mathbb{R}$, and* **3)** *$\varphi(c, \cdot)$ and $\varphi(\cdot, T)$ are differentiable for all $c, T \in (0, \infty)$.*

We now can define an analog of the optimization problem (1) with the objective function belongs to $\mathcal{F}$:

$$\min_{T_1,\ldots,T_K} \max_{i\in[K]} \varphi(c_i, T_i), \quad \text{s.t.} \quad \sum_{i=1}^{K} T_i = n, \tag{4}$$

where the parameters $(c_i)_{i=1}^K$ depend solely on the distance measure D and distributions $(P_i)_{i=1}^K$. Note that in this setting the budget allocation reduces to balancing the value of the objective function by tracking the distribution-dependent parameter $(c_i)_{i=1}^K$ (to be estimated). We refer to the solution of (4) with full knowledge of $(c_i)_{i=1}^K$ as $(\widetilde{T}_i^*)_{i=1}^K$, and to the corresponding allocation scheme as $\widetilde{\mathcal{A}}^*$. Similar to (1), when parameters $(c_i)_{i=1}^K$ are unknown, we refer to (4) as a *general tracking problem*.

Optimistic Tracking Algorithm. We now propose and analyze an adaptive sampling scheme, motivated by the upper-confidence bound (UCB) algorithm (Auer et al., 2002) in multi-armed bandits, for solving the general tracking problem (4). The proposed scheme, whose pseudo-code is shown in Algorithm 1, samples *optimistically* by plugging in high probability upper-bounds of c_i in the objective function φ. Formally, for each arm $i \in [K]$ and time $t \in [n]$, we denote by $T_{i,t}$, the number of times that arm i has been pulled prior to time t. We define the $(1-\delta)$-probability (high probability) event $\mathcal{E} \coloneqq \bigcap_{t\in[n]} \bigcap_{i\in[K]} \left\{|\hat{c}_{i,t} - c_i| \leq e_{i,t})\right\}$, where $\hat{c}_{i,t}$ is the empirical estimate of c_i and $e_{i,t}$ is the radius (half of the length) of its confidence interval at time t computed using $T_{i,t}$ samples. We define the upper-bound of c_i at time t as $u_{i,t} \coloneqq \hat{c}_{i,t} + e_{i,t}$ with the convention that $u_{i,1} = +\infty$. In the rest of the paper, we use $\hat{P}_{i,t}$ and $\hat{p}_{ij,t}$ to represent the estimates of P_i and p_{ij} at time t, computed by $T_{i,t}$ i.i.d. samples.

We now state a theorem that bounds the deviation of the allocation obtained by Algorithm 1 (our optimistic tracking algorithm), $(T_i)_{i=1}^K$, from the allocation $(\widetilde{T}_i^*)_{i=1}^K$, i.e., the solution to (4) when the parameters $(c_i)_{i=1}^K$ are known. Before

Algorithm 1 Optimistic Tracking Algorithm

Input: K, n, δ
Initialize $t \leftarrow 1$;
while $t \leq n$ **do**
  if $t \leq K$ **then**
    $I_t = t$;
  else
    $I_t = \arg\max_{i\in[K]} \ \varphi(u_{i,t}, T_{i,t})$;
  end if
  Observe $X \sim P_{I_t}$; $t \leftarrow t+1$;
  Update $u_{i,t}, \ \forall i \in [K]$;
end while

stating our main theorem, we define $g_i^* := \frac{\partial\varphi(c,\widetilde{T}_i^*)}{\partial c}\big|_{c=c_i}$ and $h_i^* := \frac{\partial\varphi(c_i,T)}{\partial T}\big|_{T=\widetilde{T}_i^*}$.

Theorem 1. *Define* $A := \max_{i\in[K]} g_i^*$, $B := \big|\max_{i\in[K]} h_i^*\big|$, *and* $\tilde{e}_n := \max_{i\in[K]} e_i^*$, *where* e_i^* *is the radius of the confidence interval of arm* i *after* $\widetilde{T}_i^*$ *pulls. Then, under the event* $\mathcal{E}$, *and assuming that* $B > 0$ *and* $\widetilde{T}_i^* > 1, \ \forall i \in [K]$, *we have*

$$-\frac{2A\tilde{e}_n}{B} \leq T_i - \widetilde{T}_i^* \leq \frac{2A(K-1)\tilde{e}_n}{B}, \quad \forall i \in [K].$$

The proof of Theorem 1, given in Appendix C, generalizes the arguments used in Carpentier et al. (2011, Thm. 1) to handle any objective function $\varphi \in \mathcal{F}$.

The idea behind preceding discussion is the following: in cases where the objective function γ_i in (1) lies in $\mathcal{F}$, we can use the result of Theorem 1 to obtain a bound on the deviation of the allocation of the resulting Algorithm 1 from the oracle deviation. In other cases, we can select an appropriate approximation of γ_i within the function class $\mathcal{F}$, and then use Theorem 1 in conjunction with a regret decomposition result (Lemma 1) to obtain the required regret bounds.

4. Adaptive Allocation Algorithms

Algorithm 1 along with the corresponding Theorem 1 provide us with a road-map to design adaptive sampling algorithms for the tracking problem (1) for different choices of distribution distance D.

4.1. Road-map

We proceed in the following steps:

- Step 1: If $\gamma_i := \mathbb{E}[D(\hat{P}_i, P_i)] \notin \mathcal{F}$ (Def. 1), then derive an approximation of $\gamma_i(\cdot)$, denoted by $\varphi(c_i, \cdot)$ lying in $\mathcal{F}$. If $\gamma_i \in \mathcal{F}$, then set $\varphi(c_i, \cdot) = \gamma_i(\cdot)$.
- Step 2: Construct an appropriate UCB for the parameter c_i for $i \in [K]$, to instantiate Algorithm 1, and use Theorem 1 to get a bound on the deviation of the allocation of Algorithm 1 from optimal.
- Step 3: Derive an upper-bound on the regret by employing the decomposition given in Lemma 1 below, along with some distance-specific analysis.

In the sequel, we shall refer to $(\widetilde{T}_i^*)_{i=1}^K$, the optimal solutions to (4), as the *approx-oracle* allocation, and the corresponding (non-adaptive) strategy, $\widetilde{\mathcal{A}}^*$, as the *approx-oracle* allocation rule. We now present the key regret decomposition result that will be used in deriving the regret bounds for the cases where $\gamma_i \notin \mathcal{F}$ and an approximation $\varphi(c_i, \cdot)$ is used in Algorithm 1.

Lemma 1. *For any allocation scheme* $\mathcal{A}$, *a distance measure* D, *and sampling budget* n, *define* $\widetilde{\mathcal{R}}_n(\mathcal{A}, D) = \mathcal{L}_n(\mathcal{A}, D) - \mathcal{L}_n(\widetilde{\mathcal{A}}^*, D)$ *and* $R_i(T) := |\gamma_i(T) - \varphi(c_i, T)|$ *for any* $T > 0$. *Then, assuming* γ_i *is non-increasing for all* $i \in [K]$, *we have*

$$\mathcal{R}_n(\mathcal{A}, D) \leq \widetilde{\mathcal{R}}_n(\mathcal{A}, D) + 3 \max_{i\in[K]} R_i(\widetilde{T}_i^*).$$

This result says that if an approximate objective function $\varphi(c_i, \cdot)$ is used, then the regret $\mathcal{R}_n(\mathcal{A}, D)$ of an allocation scheme $\mathcal{A}$ can be decomposed into its *tracking regret*, $\widetilde{\mathcal{R}}_n(\mathcal{A}, D)$ and the maximum *approximation error* between γ_i and $\varphi(c_i, \cdot)$ computed at $(\widetilde{T}_i^*)_{i=1}^K$. Lemma 1 is proved in Appendix B. The key step in the proof is bounding the quantity $|\varphi(c_i, \widetilde{T}_i^*) - \gamma_i(T_i^*)|$ with $2R_i(\widetilde{T}_i^*)$.

4.2. Adaptive Allocation for ℓ_2^2-Distance

The squared ℓ_2-distance between two distributions P and Q is defined as $D_{\ell_2}(P, Q) := \sum_{j=1}^l (p_j - q_j)^2$. In this case, we can compute the **objective function** of (1) in closed-form as

$$\gamma_i(T_i) = \mathbb{E}\big[D_{\ell_2}(\hat{P}_i, P_i)\big] = \mathbb{E}\big[\sum_{j=1}^l (\hat{p}_{ij} - p_{ij})^2\big]$$
$$= \sum_{j=1}^l \frac{p_{ij}(1-p_{ij})}{T_i} := \frac{c_i^{(\ell_2)}}{T_i} := \varphi(c_i^{(\ell_2)}, T_i).$$

Note that the function $\gamma_i(T_i) = \varphi(c_i^{(\ell_2)}, T_i) = c_i^{(\ell_2)}/T_i$ belongs to $\mathcal{F}$. The **oracle allocation** is obtained by equalizing $c_i^{(\ell_2)}/T_i$, for all $i \in [K]$, and can be written as $T_i^* = \widetilde{T}_i^* = c_i^{(\ell_2)}/(\sum_{k=1}^K c_k^{(\ell_2)}) \times n := \lambda_i^{(\ell_2)} \times n$.

Next, we present a result on the deviation between $c_i^{(\ell_2)}$ and its empirical version $\hat{c}_{i,t}^{(\ell_2)} = 1 - \sum_{j=1}^l \hat{p}_{ij}^2$.

Lemma 2. *Define* $\delta_t := 6\delta/(Kl\pi^2 t^2)$, $e_{i,t}^{(\ell_2)} := \sqrt{(l+2)^2 \log(1/\delta_t)/2T_{i,t}}$ *and the event* $\mathcal{E}_1 = \cap_{t\in[n]} \cap_{i\in[K]} \{|c_i^{(\ell_2)} - \hat{c}_{i,t}^{(\ell_2)}| \leq e_{i,t}^{(\ell_2)}\}$. *Then we have* $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta$.

Using Lemma 2 (proved in Appendix D), we can define the required UCB for $c_i^{(\ell_2)}$ as $u_{i,t}^{(\ell_2)} := \hat{c}_{i,t}^{(\ell_2)} + e_{i,t}^{(\ell_2)}$ to be plugged into Algorithm 1 to obtain an adaptive sampling scheme for D_{ℓ_2}, which we shall refer to as $\mathcal{A}_{\ell_2}$.

We can now state the bound on the regret incurred by the allocation scheme $\mathcal{A}_{\ell_2}$ (proof in Appendix D).

Theorem 2. *If we implement the algorithm* $\mathcal{A}_{\ell_2}$ *with a budget* n *and* $\delta = n^{-5/2}$, *then for* n *large enough, and* $E_{\ell_2} := \max_{1\leq i\leq K} |T_i - \widetilde{T}_i^*|$, *we have*

$$E_{\ell_2} = \mathcal{O}(\sqrt{n}), \text{ and } \mathcal{R}_n(\mathcal{A}_{\ell_2}, D_{\ell_2}) = \tilde{\mathcal{O}}(n^{-3/2}).$$

The precise meaning of the term *'n large enough'*, as well as the exact expression of the hidden constants in the above expressions are provided in Appendix D. Note that the $\tilde{\mathcal{O}}(n^{-3/2})$ convergence rate of Thm. 2 recovers the rate derived in Carpentier et al. (2011) for the special case of Bernoulli $(P_i)_{i=1}^K$. Finally, note that in contrast to the adaptive scheme which achieves a regret of $\tilde{\mathcal{O}}\left(n^{-3/2}\right)$, employing the uniform sampling scheme results in a regret of $\max_i c_i^{(\ell_2)} K/n - \left(\sum_{i=1}^K c_i^{(\ell_2)}\right)/n = \mathcal{O}(1/n)$.

4.3. Adaptive Allocation for ℓ_1-Distance

The ℓ_1-distance between two distributions P and Q is defined as $D_{\ell_1}(P,Q) \coloneqq \sum_{j=1}^l |p_j - q_j|$. Note that the total-variation distance, D_{TV}, is related to D_{ℓ_1} as $D_{\text{TV}} = \frac{1}{2} D_{\ell_1}$. In this case, the objective function γ_i can be obtained in closed-form using the expression for *mean absolute deviation* $\mathbb{E}[|\hat{p}_{ij} - p_{ij}|]$ given in Diaconis & Zabell (1991, Eq. 1.1). However, since this expression does not belong to $\mathcal{F}$, we first obtain an approximation of γ_i in $\mathcal{F}$ as

$$\begin{aligned}\gamma_i(T_i) &= \mathbb{E}\big[D_{\ell_1}(\hat{P}_i, P_i)\big] \coloneqq \mathbb{E}\big[\sum_{j=1}^l |\hat{p}_{ij} - p_{ij}|\big] \\ &\overset{\text{(a)}}{\leq} \sum_{j=1}^l \sqrt{\mathbb{E}\big[(\hat{p}_{ij} - p_{ij})^2\big]} = \frac{1}{\sqrt{T_i}} \sum_{j=1}^l \sqrt{p_{ij}(1-p_{ij})} \\ &\coloneqq \frac{c_i^{(\ell_1)}}{\sqrt{T_i}} \coloneqq \varphi(c_i^{(\ell_1)}, T_i). \end{aligned} \tag{5}$$

(a) follows from the Jensen's inequality and the concavity of the square-root function. We can check that the **approximate objective function** $\varphi(c_i^{(\ell_1)}, T_i) = c_i^{(\ell_1)}/\sqrt{T_i}$ with $c_i^{(\ell_1)} = \sum_{j=1}^l \sqrt{p_{ij}(1-p_{ij})}$ lies in $\mathcal{F}$.

The **approx-oracle allocation** is given by $\widetilde{T}_i^* = (c_i^{(\ell_1)})^2 / C_{\ell_1}^2 \times n \coloneqq \lambda_i^{(\ell_1)} \times n$, where $C_{\ell_1}^2 = \sum_{i=1}^K (c_i^{(\ell_1)})^2$. In order to obtain the **adaptive allocation scheme** for the ℓ_1-distance, which we shall refer to as $\mathcal{A}_{\ell_1}$, we now derive high probability upper-bounds on $(c_i^{(\ell_1)})_{i=1}^K$ and then plug them into Algorithm 1.

Lemma 3. *Define* $\delta_t \coloneqq 3\delta/(Kl\pi^2 t^2)$, $e_{ij,t}^{(\ell_1)} \coloneqq \sqrt{2\log(2/\delta_t)/T_{i,t}}$, *and the event* $\mathcal{E}_2 \coloneqq \bigcap_{t\in[n]} \bigcap_{i\in[K]} \bigcap_{j\in[l]} \big\{ |\sqrt{\hat{p}_{ij,t}(1-\hat{p}_{ij,t})} - \sqrt{p_{ij}(1-p_{ij})}| \leq e_{ij,t}^{(\ell_1)} \big\}$. *Then, we have* $\mathbb{P}(\mathcal{E}_2) \geq 1-\delta$.

The proof (details in Appendix E.1) relies on an application of a concentration inequality of the standard deviation of random variables derived in Maurer & Pontil (2009, Thm. 10), followed by two union bounds. Lemma 3 allows us to define high probability upper-bounds on the parameters $c_i^{(\ell_1)}$ as $u_{i,t}^{(\ell_1)} \coloneqq \hat{c}_{i,t}^{(\ell_1)} + e_{i,t}^{(\ell_1)}$, where $e_{i,t}^{(\ell_1)} = \sum_{j=1}^l e_{ij,t}^{(\ell_1)} = \sqrt{2l^2 \log(2/\delta_t)/T_{i,t}}$.

We now state the regret bound for the adaptive allocation scheme $\mathcal{A}_{\ell_1}$ (proof in Appendix E.2).

Theorem 3. *If we implement the algorithm $\mathcal{A}_{\ell_1}$ with budget n and $\delta = 1/n$, then for n large enough, and $E_{\ell_1} \coloneqq \max_{1\leq i\leq K} |T_i - \widetilde{T}_i^*|$, we have*

$$E_{\ell_1} = \tilde{\mathcal{O}}(\sqrt{n}), \text{ and } \mathcal{R}_n(\mathcal{A}_{\ell_1}, D_{\ell_1}) = \tilde{\mathcal{O}}(n^{-3/4}).$$

The exact expressions for the hidden constants in the above bounds are derived in Appendix E.2.

As a reference, we note that using the uniform allocation for the ℓ_1 loss would result in a regret of $\mathcal{O}\left(n^{-1/2}\right)$ which is larger than the $\tilde{\mathcal{O}}\left(n^{-3/4}\right)$ regret achieved by the adaptive scheme.

4.4. Adaptive Allocation for f-Divergence

For a convex function $f : \mathbb{R} \mapsto \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence between two distributions P and Q is defined as $D_f(P,Q) \coloneqq \sum_{j=1}^l q_j f(p_j/q_j)$. Since we cannot obtain a closed-form expression for the objective function γ_i of f-divergence, we proceed by writing $D_f(\hat{P}_i, P_i) = D_f^{(r)}(\hat{P}_i, P_i) + R_{i,r+1}$, where $D_f^{(r)}(\hat{P}_i, P_i)$ is the r-term Taylor's approximation of $D_f(\hat{P}_i, P_i)$, i.e.,

$$D_f^{(r)}(\hat{P}_i, P_i) \coloneqq \sum_{m=1}^r \frac{f^{(m)}(1)}{m!} \sum_{j=1}^l \frac{1}{p_{ij}^{m-1}} (\hat{p}_{ij} - p_{ij})^m, \tag{6}$$

and $R_{i,r+1} = \sum_{j=1}^l R_{ij,r+1}$ is its remainder term (assuming f is analytic at 1), i.e.,

$$R_{ij,r+1} \coloneqq \sum_{m=r+1}^{\infty} \frac{f^{(m)}(1)}{m!\, p_{ij}^{m-1}} (\hat{p}_{ij} - p_{ij})^m. \tag{7}$$

Note that in (6) and (7), $f^{(m)}(\cdot)$ is the m^{th} derivative of f. We now define the **approximate objective function** for an f-divergence as

$$\begin{aligned}\varphi(c_i, T_i) &\coloneqq \mathbb{E}[D_f^{(r)}(\hat{P}_i, P_i)] \\ &= \sum_{m=1}^r \sum_{j=1}^l \frac{f^{(m)}(1)}{m!\, p_{ij}^{m-1}\, T_i^m} \mathbb{E}\Big[\big(\sum_{s=1}^{T_i} \tilde{Z}_{ij}^{(s)}\big)^m\Big]. \end{aligned} \tag{8}$$

Note that the exact value of the parameter c_i above depends on the values of the terms $f^{(m)}(1)$, for $1 \leq m \leq r$.

Next, we present a general result on the quality of the approximation of γ_i with $\varphi(c_i, T_i)$ under the following two assumptions on f: **(f1)** $f(x)$ is real-analytic at the point $x = 1$ and **(f2)** $f^{(m)}(1)/m! \leq C_1 < \infty$, $\forall m \in \mathbb{N}$. Both these assumptions are satisfied by several commonly used f-divergences, namely KL-divergence with $f(x) = x\log x$, χ^2-divergence with $f(x) = (x-1)^2$, and Hellinger distance with $f(x) = 2(1-\sqrt{x})$.

Lemma 4. *Assume that f satisfies **(f1)** and **(f2)**. Then, there exists a constant $C_{f,r+1} < \infty$, whose exact definition is given by Eqs. 23, 24, and 28 in Appendix F.2, such that the following holds:*

$$\mathbb{E}\left[R_{ij,r+1}\right] \leq C_{f,r+1} (p_{ij} T_i)^{-(r+1)/2}.$$

To proceed further according to the roadmap of Section 4.1, we need to construct an upper-bound of the parameter c_i that depends on the specific choice of function f. We carry out these derivations for $f(x) = x \log x$ (i.e., KL-divergence) in Section. 4.4.1 below.

Remark 1. *An alternative approach is to proceed by assuming that there exist $\tau_{0,i}$ and $\tau_{1,i}$ such that with probability at least $1-\delta$, we have $\tau_{0,i} \leq T_i \leq \tau_{1,i}$ for $i \in [K]$ (analogous to the statement of Thm. 1). Under this assumption, we can obtain a very general regret decomposition for arbitrary f-divergence distance measures satisfying **(f1)** and **(f2)**. The details of this approach are given in Appendix F.1, and in particular the formal statement of regret decomposition is in Thm. 7 in Appendix F.1.*

4.4.1. Adaptive Allocation for KL-Divergence

The KL-divergence between distributions P and Q is defined as $D_{\mathrm{KL}}(P,Q) \coloneqq \sum_{j=1}^{l} p_j \log(p_j/q_j)$. We begin by deriving its r-term approximation with $r = 5$, i.e.,

$$\mathbb{E}\big[D_{\mathrm{KL}}(\hat{P}_i, P_i)\big] = \mathbb{E}\big[\sum_{j=1}^{l} \hat{p}_{ij} \log(\hat{p}_{ij}/p_{ij})\big]$$

$$\overset{\text{(a)}}{=} \frac{l-1}{2T_i} + \frac{1}{12T_i^2} \sum_{j=1}^{l} \big(\frac{1}{p_{ij}} - 1\big) + \mathcal{O}(1/T_i^3), \quad (9)$$

where **(a)** is by calculating the 5^{th} order Taylor's approximation of the mapping $x \mapsto x\log(x)$. The calculations involved in this derivation are described in Harris (1975, Sec. 2). The choice of $r = 5$ is sufficient as it is the smallest r for which the approximation error, which is of $\mathcal{O}(n^{-3})$ according to Lemma 4, is smaller than the tracking regret, that as we will show in the proof of Thm. 4, is of $\mathcal{O}(n^{-5/2})$.

Eq. (9) gives us the **approximate objective function** $\varphi(c_i^{(\mathrm{KL})}, T_i) \coloneqq \frac{l-1}{2T_i} + \frac{c_i^{(\mathrm{KL})}}{T_i^2}$, with $c_i^{(\mathrm{KL})} \coloneqq \big(\sum_{j=1}^{l} 1/p_{ij} - 1\big)/12$. Note that this $\varphi(c_i^{(\mathrm{KL})}, T_i)$ belongs to the class of functions $\mathcal{F}$ introduced in Definition 1. Deriving the **approx-oracle allocation** $(\widetilde{T}_i^*)_{i=1}^K$ requires solving a cubic equation. Instead of computing the exact form of $\widetilde{T}_i^*$, we show in Lemma 5 that the deviation of $\widetilde{T}_i^*$ from the uniform allocation is bounded by a problem-dependent constant, implying that the uniform allocation is near-optimal. This is not surprising as the first order approximation of φ (the first term on the RHS of Eq. 9) does not change with P_i.

Lemma 5. *For $(P_i)_{i=1}^K \in \Delta_l^{(\eta)}$ and $(\widetilde{T}_i^*)_{i=1}^K$ denoting the approx-oracle allocation, we have*

$$\Big|\widetilde{T}_i^* - T_0\Big| \leq K \frac{c_{\max}^{(KL)} - c_{\min}^{(KL)}}{l-1}, \quad \forall i \in [K],$$

where $c_{\min}^{(KL)}$ and $c_{\max}^{(KL)}$ denote the minimum and maximum values of $c_i^{(KL)}$, respectively.

Next, with $e_{ij,t} \coloneqq \sqrt{2\log(2/\delta_t)/T_{i,t}}$, we define the following upper-bound for the parameters $c_i^{(\mathrm{KL})}$ $\forall i \in [K]$:

$$u_{i,t}^{(KL)} = \begin{cases} \big(\sum_{j=1}^{l} \frac{1}{\hat{p}_{ij,t} - e_{ij,t}} - 1\big)/12 & \text{if } \hat{p}_{ij} \geq \frac{7e_{ij,t}}{2} \\ +\infty & \text{otherwise.} \end{cases}$$

The deviation of this upper-bound from the true parameters $c_i^{(\mathrm{KL})}$ can be computed by exploiting the convexity of the mapping $x \mapsto 1/x$ and the exact expression of the length of the confidence interval reported in Lemma 10 in Appendix G. These upper-bounds can then be plugged into Algorithm 1 to obtain an **adaptive allocation scheme** for KL-divergence, denoted by $\mathcal{A}_{\mathrm{KL}}$. Finally, we state the regret bound for $\mathcal{A}_{\mathrm{KL}}$ in the following theorem (proof in Appendix G):

Theorem 4. *Let $(P_i)_{i=1}^K \in \Delta_l^{(\eta)}$ and the adaptive scheme $\mathcal{A}_{KL}$ is implemented with $\delta = (3K/n)^6$. Then for large enough n and with $E_{KL} \coloneqq \max_{i\in[K]} |T_i - \widetilde{T}_i^*|$, we have*

$$E_{KL} = \tilde{\mathcal{O}}(n^{-1/2}), \;\; \text{and} \;\; \mathcal{R}_n(\mathcal{A}_{KL}, D_{KL}) = \tilde{\mathcal{O}}(n^{-5/2}).$$

As we showed in Lemma 5, the approx-oracle allocation for D_{KL} is close, although not identical to, the uniform allocation. This is due to the fact that the first order term in the approximation given in (9) only depends on the support size of the distributions, which is assumed to be the same for all K distributions in our setting. Thus, the uniform allocation is the approx-oracle allocation for the first order allocation for D_{KL} and it achieves an upper bound on the regret of $\mathcal{O}\left(n^{-2}\right)$. However, as shown in Theorem 4 above, consideration of the higher order terms allows us to achieve a $\tilde{\mathcal{O}}\left(n^{-5/2}\right)$ regret (see Eq. 45 in Appendix G for exact expression)

4.5. Adaptive Allocation for Separation Distance

The separation distance (Gibbs & Su, 2002) between distributions P and Q is defined as $D_s(P,Q) \coloneqq \max_{j\in[l]}(1 - p_j/q_j)$. We start by introducing new notations. Given a probability distribution $P_i \in \Delta_l$ and a *non-empty* set $S \subset [l]$, we define $p_{i,S} \coloneqq \sum_{j\in S} p_{ij}$. We also define the functions $\rho_1(p) \coloneqq \sqrt{(1-p)/p}$ and $\rho_2(p) \coloneqq \rho_1(p) + \rho_1(1-p)$, and introduce the terms $c_i^{(s)} \coloneqq \sum_{j=1}^{l} \rho_1(p_{ij})$ and $\tilde{c}_i^{(s)} \coloneqq \max_{S\subset[l]}\{\rho_2(p_{i,S})\}$. Note that $\tilde{c}_i^{(s)} = c_i^{(s)}$ for $l = 2$. Because of the $\max$ operation in the definition of D_s, in general, we cannot obtain a closed-form expression for the objective function $\gamma_i(T_i) = \mathbb{E}[D_s(\hat{P}_i, P_i)]$. We now state a key lemma (proof in Appendix H) that provides an approximation of $\mathbb{E}[D_s(\hat{P}_i, P_i)]$.

Lemma 6. *For a distribution $P_i \in \Delta_l$, let $\hat{P}_i = (\hat{p}_{ij})_{j=1}^l$ be the empirical distribution constructed from T_i i.i.d. draws from P_i. Then, we have*

$$\tilde{c}_i^{(s)}\sqrt{\frac{1}{2\pi T_i}} - \frac{\tilde{C}_i^{(s)}}{T_i} \leq \mathbb{E}\big[D_s(\hat{P}_i, P_i)\big] \leq c_i^{(s)}\sqrt{\frac{1}{2\pi T_i}} + \frac{C_i^{(s)}}{T_i},$$

where $C_i^{(s)}$ and $\tilde{C}_i^{(s)}$ are P_i-dependent constants defined by (47) *and* (50) *in Appendix H.*

The proof of the upper bound in Lemma 6 proceeds by first upper bounding the term inside the expectation with a normalized sum of random variables, and then using a non-asymptotic version of the Central Limit Theorem (Ross, 2011, Thm. 3.2). For deriving the lower bound, we first show (Lemma 12 in Appendix H) that 'bunching together' probability masses can only reduce the separation distance between two distributions, and then proceed by another application of (Ross, 2011, Thm. 3.2).

Lemma 6 gives us an interval that contains the true objective function we aim to track. To implement the adaptive scheme, we employ the **approximate objective function** $\varphi(c_i^{(s)}, T_i) \coloneqq c_i^{(s)}\sqrt{1/2\pi T_i}$. In order to instantiate Algorithm 1 for D_s, we require to derive high probability confidence intervals for the terms $\sqrt{(1-p_{ij})/p_{ij}}$ in the definition $(c_i^{(s)})_{i=1}^K$. We use the event $\mathcal{E}_1$ defined in Lemma 2 and prove the following result:

Lemma 7. *Let $P_i \in \Delta_l^{(\eta)}$, and the event $\mathcal{E}_1$ and the terms δ_t and $e_{ij,t}$ defined as in Lemma 2. Define the terms $a_{i,t} \coloneqq \big(8\log(2/\delta_t)/T_{i,t}\big)^{1/4}$ and $b_{i,t} \coloneqq \big(\frac{l\, a_{i,t}}{\eta}\big)\max\big\{1, \frac{a_{i,t}}{2\eta^{3/2}}\big\}$.*

Then, under the high probability event $\mathcal{E}_1$, we have

$$\sum_{j=1}^{l}\sqrt{\frac{1}{p_{ij}}-1} \le \sum_{j=1}^{l}\sqrt{\frac{1}{\hat{p}_{ij,t}-e_{ij,t}}-1} \le \sum_{j=1}^{l}\sqrt{\frac{1}{p_{ij}}-1}+b_{i,t}.$$

Using the concentration result of Lemma 7, we can now implement Algorithm 1 with the upper-bound $u_{i,t}^{(s)} = \big(\sum_{j=1}^{l}\sqrt{\frac{1}{\hat{p}_{ij,t}-e_{ij,t}}-1}\big)/(\sqrt{2\pi})$, if $\hat{p}_{ij,t} \ge 7e_{ij,t}/2$, and $u_{i,t}^{(s)} = +\infty$, otherwise. This will give us an adaptive allocation scheme for the separation distance, which we shall refer to it as $\mathcal{A}_s$. Finally, we prove the following regret bound for $\mathcal{A}_s$ (proof in Appendix H.3).

Theorem 5. *Let $P_i \in \Delta_l^{(\eta)}$ and the adaptive scheme $\mathcal{A}_s$ is implemented with $\delta = \eta/n$. Then, for large enough n and with $E_s \coloneqq \max_{1\le i\le K}|T_i - \widetilde{T}_i^*|$, we have*

$$E_s = \tilde{\mathcal{O}}(\sqrt{n}), \quad \textit{and}$$

$$\mathcal{R}_n(\mathcal{A}_s, D_s) = \tilde{\mathcal{O}}\big(\frac{\max_{i\in[K]}\big(c_i^{(s)} - \tilde{c}_i^{(s)}\big)}{\sqrt{n}} + \frac{\sqrt{E_s}}{n}\big). \quad (10)$$

The exact condition for n, and the expressions for E_s and the higher-order terms in (10) *are given in Appendix H.3.*

Remark 2. *Note that the second term on the RHS of* (10) *is the approximation error term in the regret decomposition introduced in Lemma 1, while the second term is the tracking regret w.r.t. the approx-oracle allocation scheme. In general, the approximation error, which is $\tilde{\mathcal{O}}(n^{-1/2})$, dominates the tracking regret term, which is $\tilde{\mathcal{O}}(n^{-3/4})$. However, for the special case of $l = 2$, the approximation error term becomes $\tilde{\mathcal{O}}(1/n)$ using the fact that $\tilde{c}_i^{(s)} = c_i^{(s)}$ in Lemma 6, and we achieve an overall regret of $\tilde{\mathcal{O}}(n^{-3/4})$.*

5. Lower Bound

Lemma 1 provided a general high probability bound on the deviation of the adaptive allocation $(T_i)_{i=1}^K$ from the approx-oracle allocation $(\widetilde{T}_i^*)_{i=1}^K$. In Sec. 4, we observed that when specialized to the objective functions corresponding to D_{ℓ_2}, D_{ℓ_1} and D_s, we have $|T_i - \widetilde{T}_i^*| = \tilde{\mathcal{O}}(\sqrt{n})$. A natural question to ask is whether there exists any other adaptive scheme that can achieve a smaller deviation from the approx-oracle allocation. We now show that this is not the case by deriving a lower-bound on the expected deviation of any allocation scheme $\mathcal{A}$.

To derive the lower-bound, we consider a specific class of problems with two arms, $K = 2$, Bernoulli distributions, $l = 2$, and objective functions of the form $\varphi(c_i, T_i) = c_i/T_i^\alpha$, for some $\alpha > 0$. For some $p_0 \in (1/2, 1)$ and $\epsilon > 0$, we define two Bernoulli distributions $P_1 \sim \text{Ber}\,(p_0)$ and $P_2 \sim \text{Ber}\,(p_0 - \epsilon)$. We consider two problem instances $\mathcal{P}_1$ and $\mathcal{P}_2$ with $K = 2$ and distributions P_1 and P_2, but with orders swapped, i.e., $\mathcal{P}_1 = (P_1, P_2)$ and $\mathcal{P}_2 = (P_2, P_1)$. Finally, we introduce the notation $\kappa(p)$ to represent the distribution dependent constant in the objective function φ corresponding to a $\text{Ber}\,(p)$ distribution. We now state the lower bound result.

Theorem 6. *For some $p_0 \in (1/2, 3/4]$ and $0 < \epsilon < p_0 - 1/2$, consider two tracking problems $\mathcal{P}_1 = (P_1, P_2)$ and $\mathcal{P}_2 = (P_2, P_1)$, with $P_1 \sim$ Ber (p_0) and $P_2 \sim$ Ber $(p_0 - \epsilon)$ and objective function $\varphi(c, T) = c/T^\alpha$ for $\alpha > 0$ where the constant $c = \kappa(p)$ for Ber (p) distributions. Finally, introduce the notation $\tau = (n/2)(|\kappa(p_0)^{1/\alpha} - \kappa(p_0 - \epsilon)^{1/\alpha}|)/(|\kappa(p_0)^{1/\alpha} + \kappa(p_0 - \epsilon)^{1/\alpha}|)$. If $(T_i)_{i=1}^2$ denotes the allocation of any allocation scheme $\mathcal{A}$, we have*

$$\max_{\mathcal{P}_1,\mathcal{P}_2}\max_{i=1,2}\mathbb{E}\left[|T_i - \widetilde{T}_i^*|\right] \ge \sup_{0<\epsilon<p_0-1/2}\Gamma_\epsilon(\kappa, p_0),$$

$$\textit{where}\ \ \Gamma_\epsilon(\kappa, p_0) = \frac{\tau}{2}\left(1 - \epsilon\sqrt{n/(1-p_0)}\right).$$

As an immediate corollary of Theorem 6, we can observe that the deviation of the optimistic tracking scheme from the approx-oracle for D_{ℓ_2}, D_{ℓ_1} and D_s cannot be improved upon by any adaptive scheme.

Corollary 1. *For $p_0 = 3/4$, $\epsilon = 1/(4\sqrt{n})$ and the κ arising in the study of D_{ℓ_2}, D_{ℓ_1} and D_s, we have $\Gamma_\epsilon(\kappa, p_0) = \Omega(\sqrt{n})$.*

The proofs of Theorem 6 and Corollary 1 are provided in Appendix I.

Remark 3. *Note that in Theorem 6, we present an algorithm independent lower bound on the allocation and not on the regret. The main reason is that our problem does not admit a straightforward regret decomposition as in the case of multi-armed bandit problems (Lattimore & Szepesvári, 2018, Lemma 4.5). Nevertheless, Theorem 6 establishes the optimality of our proposed algorithm in terms of the deviation from the optimal allocation for ℓ_2, ℓ_1 and separation distances. Furthermore, it also establishes a similar sense*

of optimality of the algorithms of (Antos et al., 2008) and (Carpentier et al., 2011) for the problem of learning the mean of K distributions in squared error sense.

6. Experiments

Setup. We study the performance of the proposed adaptive schemes on a problem with $K = 2$, and $l = 10$. We set P_1 as the uniform distribution in Δ_l and $P_2 = P_\epsilon$ for $\epsilon \in \{0.1, 0.2, \ldots, 0.9\}$, where $P_\epsilon = (p_j)_{j=1}^l$ with $p_1 = \epsilon$ and $p_j = (1-\epsilon)/(l-1)$ for $1 < j \leq l$.

To compare the performance of the adaptive schemes, we used three baseline schemes:

(i) *Uniform allocation*, in which each arm is allocated n/K samples. Note that the uniform allocation is the oracle scheme for D_{χ^2} (see Appendix G.4),

(ii) *Greedy allocation*, in which the arms are pulled by plugging in the current empirical estimate $(\hat{c}_{i,t})_{i=1}^K$ of $(c_i)_{i=1}^K$ in the objective function, and

(iii) *Forced Exploration*, in which the arms are pulled according to the greedy scheme, while also ensuring that at any time t, each arm is pulled at least $\sqrt{t}$ times. This scheme is motivated by the strategy of Antos et al. (2008).

For every value of ϵ, we ran 500 trials of all the allocation schemes with the budget $n = 5000$. We focus our experiments on the ℓ_2^2, ℓ_1 and separation distances, since we observed no statistically significant difference in the performance of the different schemes for KL-divergence. To compare the performance of the allocation schemes, we plot the term $\varphi(c_i, T_i) - \varphi(c_i, \widetilde{T}_i^*)$.

Observations. We plot the $\varphi(c_i, T_i) - \varphi(c_i, \widetilde{T}_i^*)$ values for the different allocation schemes and loss functions in Figs. 1, 2 and 3. As we can see from Fig. 1, the adaptive scheme outperforms the uniform allocation for the three distance metrics for both $\epsilon = 0.5$ and $\epsilon = 0.9$. Note that as ϵ increases, the optimal allocation get more skewed, and hence the gap in performance between uniform and adaptive also increases. The greedy and forced exploration schemes, both perform comparably to our proposed adaptive scheme for $\epsilon = 0.5$, although their resulting allocations have higher variability especially for ℓ_2^2 and ℓ_1 distances. For the case of $\epsilon = 0.9$ however, the adaptive scheme performs significantly better than both greedy and forced exploration methods for ℓ_2^2 and ℓ_1 distances, and result in a lower variance solution for separation distance.

7. Extensions

We now discuss two extensions of the results of the previous sections.

Continuous Distributions. The results presented in this paper can be extended to some classes of continuous distributions and some distance measures. For instance, assume that $(P_i)_{i=1}^K$ are continuous distributions on $[0, 1]$ which admit density functions $(\nu_i)_{i=1}^K$ which can be expanded in terms of a finite number of orthonormal basis functions

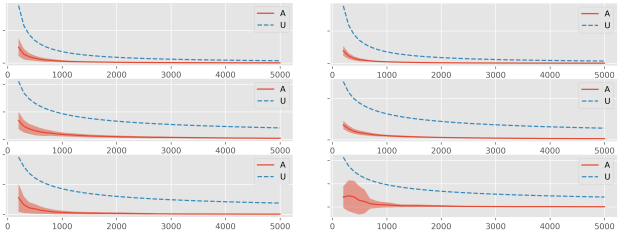

Figure 1: Comparison of Algorithm 1 with Uniform Allocation for ℓ_2^2 (top), ℓ_1 (middle) and separation distance (bottom) for $\epsilon = 0.5$ (left) and $\epsilon = 0.9$ (right).

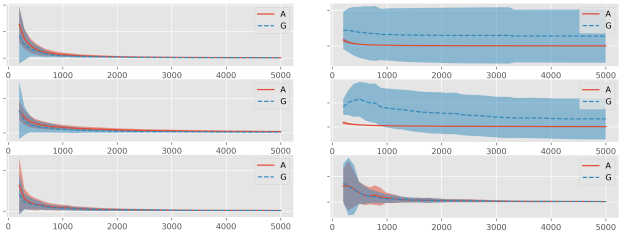

Figure 2: Comparison of Algorithm 1 with Greedy Allocation for ℓ_2^2 (top), ℓ_1 (middle) and separation distance (bottom) for $\epsilon = 0.5$ (left) and $\epsilon = 0.9$ (right).

$(\psi_j)_{j=1}^l$, i.e., $\nu_i = \sum_{j=1}^l a_{ij}\psi_j$. By using appropriate basis functions, such as *Fourier Basis* and *wavelet basis*, a large class of density functions can be modeled under this assumption. For constructing an estimate of ν_i, denoted by $\hat{\nu}_i$, we can employ the *projection estimator* (Tsybakov, 2009, § 1.7) which estimates the coefficients of the basis expansion using the observations. By exploiting the orthonormality of $(\psi_j)_{j=1}^l$, we can show that the expectation of integrated mean-squared error, i.e., $\mathbb{E}[\int(\hat{\nu}_i - \nu_i)^2 dx]$, belongs to $\mathcal{F}$. With this objective function available, we can instantiate the optimistic tracking algorithm and derive bounds on the regret similar to the discrete case. The details about the estimator construction and the objective function derivation are provided in Appendix J.

Minimizing Average Discrepancy. In this paper, our focus has been on minimizing the maximum distance between the estimate and true distributions, i.e., optimization problems (1) and (4). An important alternative formulation that has been studied in the bandit literature involves minimizing the average discrepancy (Carpentier & Munos, 2011; Riquelme et al., 2017). Our results, in particular our general tracking scheme, can be extended to this case and we are able to provide adaptive allocation strategies to minimize the average distance between distributions, for all distances studied in this paper. Consider the following tracking/optimization problem, which is the equivalent of (4) for the average case:

$$\min_{T_1,\ldots,T_K} \frac{1}{K}\sum_{i=1}^K \varphi_i(c_i, T_i) \quad \text{s.t.} \sum_{i=1}^K T_i = n. \tag{11}$$

If φ_i's are convex in T_i, then the optimal solution must satisfy $\frac{1}{K}\frac{\partial \varphi_i(c_i,T_i)}{\partial T_i} - \lambda = 0$, for all $i \in [K]$ and for some

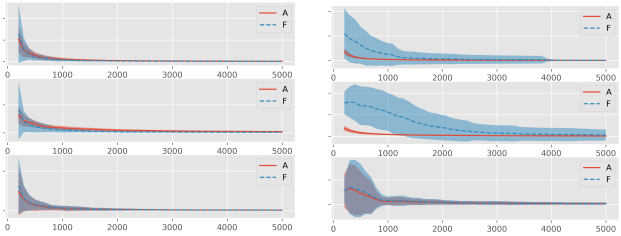

Figure 3: Comparison of Algorithm 1 with Forced Allocation for ℓ_2^2 (top), ℓ_1 (middle) and separation distance (bottom) for $\epsilon = 0.5$ (left) and $\epsilon = 0.9$ (right).

$\lambda \in \mathbb{R}$. Thus, if the T_i-derivatives of φ_i are in the function class $\mathcal{F}$ (Definition 1), then (11) can be solved using the tools developed in Section 3. It is easy to show that the distances studied in this paper (i.e., ℓ_2, ℓ_1, KL, and separation) satisfy this condition.

Wasserstein Distances. An important class of distance metrics between two probability measures μ, ν is the Wasserstein family of distances, denoted by $W_p(\mu, \nu)$ for $p \geq 1$. These distance metrics have recently been used in several problems in machine learning and statistics. Some existing results, such as (Weed et al., 2019, Proposition 1), derive sharp relations between $W_p(\mu, \nu)$ and the ℓ_1 distance between the discrete distributions obtained by restricting μ and ν to a given partition $\mathcal{Q}$ of the input space. Thus for a fixed partition $\mathcal{Q}$, our analysis of ℓ_1 distance can be used to get an upper bound on the regret of learning two distributions in W_p distance. Such an upper bound would depend on the properties of the partition $\mathcal{Q}$ (such as cardinality and diameter of its elements) in addition to the sampling budget n. An interesting question for future work is designing an adaptive method of constructing the partition $\mathcal{Q}$ given n in order to achieve the tightest bounds on the regret for learning distributions in terms of W_p.

8. Conclusion

We studied the problem of allocating a fixed budget of samples to learn K discrete distributions uniformly well in terms of four distance measures: ℓ_2^2, ℓ_1, f-divergence, and separation. We proposed a general optimistic tracking strategy for problems with concave-convex and differentiable objective functions and then showed that this class of functions is rich enough to either contain or well approximate the true objective functions of all the considered distances. We then derived regret bounds for the proposed algorithm for all four distances. We showed that the allocation performance of the proposed scheme cannot in general be improved, by deriving lower-bounds. We also empirically verified our theoretical findings through numerical experiments. Finally, we ended with a discussion on extending our results to certain classes of continuous distributions and to a related setting of average error minimization.

Following the style of results presented in the related works of (Antos et al., 2008; Carpentier et al., 2011), we derived upper-bounds on the regret in terms of the budget n and in the large n regime, with l and K fixed. However, there are several interesting directions not considered in this paper, which can be explored in future work, such as **1)** improving the performance of the adaptive algorithms and the hidden constants in the regret-bounds by employing stronger concentration results, **2)** handling the large l case by using appropriate estimators such as the estimator of Santhanam et al. (2007), and the large K case by imposing some additional similarity assumptions among the different arms similar to Bubeck et al. (2011), and **3)** extending the results of the paper to the general problem of learning the dynamics (model) of a finite MDP, as discussed in Section 1.

References

Aldous, D. and Diaconis, P. Strong uniform times and finite random walks. *Advances in Applied Mathematics*, 8(1): 69–97, 1987.

Antos, A., Grover, V., and Szepesvári, C. Active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 287–302, 2008.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Barron, A., Gyorfi, L., and van der Meulen, E. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE transactions on Information Theory*, 38(5):1437–1454, 1992.

Blyth, C. Expected absolute error of the usual estimator of the binomial parameter. *The American Statistician*, 34(3): 155–157, 1980.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

Carpentier, A. and Munos, R. Finite time analysis of stratified sampling for monte carlo. In *Advances in Neural Information Processing Systems 24*, pp. 1278–1286, 2011.

Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R., and Auer, P. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 189–203, 2011.

Cheung, W. C. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. In *Advances in Neural Information Processing Systems*, pp. 724–734, 2019.

Csiszár, I. Two remarks to noiseless coding. *Information and Control*, 11(3):317–322, 1967.

Csiszár, I. Generalized cutoff rates and rényi's information measures. *IEEE Transactions on information theory*, 41 (1):26–34, 1995.

Diaconis, P. and Zabell, S. Closed form summation for classical distributions: variations on a theme of de moivre. *Statistical Science*, pp. 284–302, 1991.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Gibbs, A. and Su, F. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Gyorfi, L., Morvai, G., and Vajda, I. Information-theoretic methods in testing the goodness of fit. In *2000 IEEE International Symposium on Information Theory (Cat. No. 00CH37060)*, pp. 28. IEEE, 2000.

Harris, B. The statistical estimation of entropy in the non-parametric case. Technical report, University of Wisconsin-Madison Mathematics Research Center, 1975.

Hazan, E., Kakade, S., Singh, K., and Soest, A. V. Provably efficient maximum entropy exploration. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. On learning distributions from their samples. In *Conference on Learning Theory*, pp. 1066–1100, 2015.

Kaufmann, E. and Garivier, A. Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and Surveys*, 60:114–131, 2017.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *preprint*, 2018.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Riquelme, C., Ghavamzadeh, M., and Lazaric, A. Active learning for accurate estimation of linear models. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2931–2939, 2017.

Rivasplata, O. Subgaussian random variables: An expository note. *Technical Report*, 2012.

Ross, N. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

Santhanam, N., Orlitsky, A., and Viswanathan, K. New tricks for old dogs: Large alphabet probability estimation. In *2007 IEEE Information Theory Workshop*, pp. 638–643. IEEE, 2007.

Soare, M., Lazaric, A., and Munos, R. Active learning in linear stochastic bandits. In *Bayesian Optimization in Theory and Practice*, 2013.

Tarbouriech, J. and Lazaric, A. Active exploration in markov decision processes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2009.

Weed, J., Bach, F., et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.