# 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: $\ell_2^2$, $\ell_1$, $f$-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 37th International Conference on Machine Learning*, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

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: $\ell_2^2$, $\ell_1$, $f$-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 $\Delta_l$, 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:

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

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 $\gamma_i(T_i) := \mathbb{E}\big[D(\hat{P}_i, P_i)\big]$ 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: $\ell_2^2$, $\ell_1$ 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 ($D_{\text{KL}}$) and includes a number of other common distances, such as total variation ($D_{\text{TV}}$), Hellinger ($D_H$), and chi-square ($D_{\chi^2}$). 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: $D_{\mathrm{TV}} \leq D_H \leq \sqrt{D_{\mathrm{KL}}} \leq \sqrt{D_{\chi^2}}$, 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 ($\ell_2^2$, $\ell_1$ and $f$-divergence) does not change the final regret obtained. For instance, for the $\ell_2^2$ 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 $\ell_1$ 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 $\eta$-interior of $(l-1)$-dimensional simplex $\Delta_l$, 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 $\varphi$, $A$, $B$, and $\tilde{e}_n$ (to be introduced in Sec. 3) are overloaded for different distance measures. For instance, we use $\varphi$ for both $\ell_1$ and KL-divergence, instead of writing $\varphi^{(\ell_1)}$ and $\varphi^{(\mathrm{KL})}$. The meaning should be clear from the local context.

## 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 $\mathcal{F}$:

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

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 $\varphi$. 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

[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.

**Algorithm 1** Optimistic Tracking Algorithm

```
**Input:** $K, n, \delta$
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* $A := \max_{i \in [K]} g_i^*$, $B := \left|\max_{i \in [K]} h_i^*\right|$, *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 $\gamma_i$ 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 $\gamma_i$ 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* $\gamma_i$ *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 $\gamma_i$ 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 $\ell_2^2$-Distance

The squared $\ell_2$-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 $D_{\ell_2}$, 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*

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

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 $\ell_1$-Distance

The $\ell_1$-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, $D_{\text{TV}}$, is related to $D_{\ell_1}$ as $D_{\text{TV}} = \frac{1}{2} D_{\ell_1}$. In this case, the objective function $\gamma_i$ 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 $\gamma_i$ 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 $\ell_1$-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* $\mathcal{E}_2 \coloneqq \bigcap_{t\in[n]} \bigcap_{i\in[K]} \bigcap_{j\in[l]} \{|\sqrt{\hat{p}_{ij,t}(1 - \hat{p}_{ij,t})} - \sqrt{p_{ij}(1 - p_{ij})}| \leq e_{ij,t}^{(\ell_1)}\}$. *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*

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

*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 $\ell_1$ 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 $\gamma_i$ 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 $\gamma_i$ 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 $f(x) = x \log x$, $\chi^2$-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 $D_{\text{KL}}(P, Q) := \sum_{j=1}^{l} p_j \log(p_j/q_j)$. We begin by deriving its $r$-term approximation with $r = 5$, i.e.,

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

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

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** $\varphi(c_i^{(\text{KL})}, T_i) := \frac{l-1}{2T_i} + \frac{c_i^{(\text{KL})}}{T_i^2}$, with $c_i^{(\text{KL})} := \big(\sum_{j=1}^{l} 1/p_{ij} - 1\big)/12$. Note that this $\varphi(c_i^{(\text{KL})}, T_i)$ 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 $\varphi$ (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*

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

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

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

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

The deviation of this upper-bound from the true parameters $c_i^{(\text{KL})}$ 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 $\mathcal{A}_{\text{KL}}$. Finally, we state the regret bound for $\mathcal{A}_{\text{KL}}$ 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 $E_{KL} := \max_{i \in [K]} |T_i - \widetilde{T}_i^*|$, we have*

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

As we showed in Lemma 5, the approx-oracle allocation for $D_{\text{KL}}$ 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 $D_{\text{KL}}$ 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 $D_s(P, Q) := \max_{j \in [l]}(1 - p_j/q_j)$. 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 $p_{i,S} := \sum_{j \in S} p_{ij}$. We also define the functions $\rho_1(p) := \sqrt{(1-p)/p}$ and $\rho_2(p) := \rho_1(p) + \rho_1(1-p)$, and introduce the terms $c_i^{(s)} := \sum_{j=1}^{l} \rho_1(p_{ij})$ and $\tilde{c}_i^{(s)} := \max_{S \subset [l]}\{\rho_2(p_{i,S})\}$. 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** $\varphi(c_i^{(s)}, T_i) := c_i^{(s)}\sqrt{1/2\pi T_i}$. 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 $\delta_t$ and $e_{ij,t}$ defined as in Lemma 2. Define the terms $a_{i,t} := \big(8\log(2/\delta_t)/T_{i,t}\big)^{1/4}$ and $b_{i,t} := \big(\frac{l\, a_{i,t}}{\eta}\big)\max\big\{1, \frac{a_{i,t}}{2\eta^{3/2}}\big\}$.*

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

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

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 $\hat{p}_{ij,t} \geq 7e_{ij,t}/2$, 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 $E_s := \max_{1\leq i\leq K}|T_i - \widetilde{T}_i^*|$, we have*

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

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

*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 $D_{\ell_2}$, $D_{\ell_1}$ and $D_{\text{s}}$, 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 $\varphi$ corresponding to a 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*

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

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

As an immediate corollary of Theorem 6, we can observe that the deviation of the optimistic tracking scheme from the approx-oracle for $D_{\ell_2}$, $D_{\ell_1}$ and $D_{\text{s}}$ cannot be improved upon by any adaptive scheme.

**Corollary 1.** *For $p_0 = 3/4$, $\epsilon = 1/(4\sqrt{n})$ and the $\kappa$ arising in the study of $D_{\ell_2}$, $D_{\ell_1}$ and $D_{\text{s}}$, 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 $\ell_2$, $\ell_1$ 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 $\Delta_l$ 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 $D_{\chi^2}$ (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 $\epsilon$, we ran 500 trials of all the allocation schemes with the budget $n = 5000$. We focus our experiments on the $\ell_2^2$, $\ell_1$ 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 $\epsilon$ 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 $\ell_2^2$ and $\ell_1$ distances. For the case of $\epsilon = 0.9$ however, the adaptive scheme performs significantly better than both greedy and forced exploration methods for $\ell_2^2$ and $\ell_1$ 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 $(\psi_j)_{j=1}^l$, 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 $\nu_i$, 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.

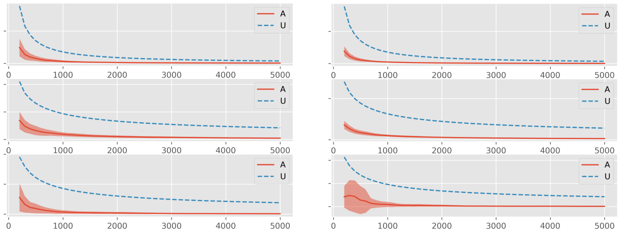

Figure 1: Comparison of Algorithm 1 with Uniform Allocation for $\ell_2^2$ (top), $\ell_1$ (middle) and separation distance (bottom) for $\epsilon = 0.5$ (left) and $\epsilon = 0.9$ (right).

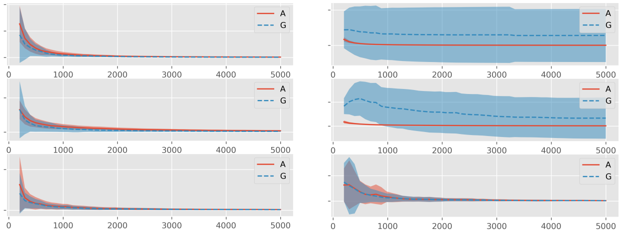

Figure 2: Comparison of Algorithm 1 with Greedy Allocation for $\ell_2^2$ (top), $\ell_1$ (middle) and separation distance (bottom) for $\epsilon = 0.5$ (left) and $\epsilon = 0.9$ (right).

**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 $\varphi_i$'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

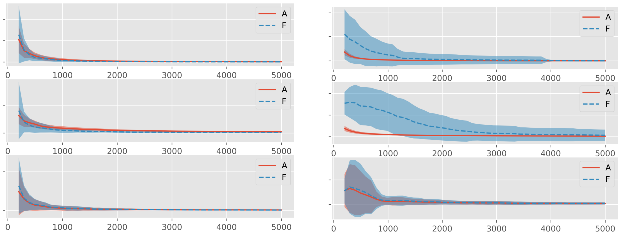

Figure 3: Comparison of Algorithm 1 with Forced Allocation for $\ell_2^2$ (top), $\ell_1$ (middle) and separation distance (bottom) for $\epsilon = 0.5$ (left) and $\epsilon = 0.9$ (right).

$\lambda \in \mathbb{R}$. Thus, if the $T_i$-derivatives of $\varphi_i$ 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., $\ell_2$, $\ell_1$, KL, and separation) satisfy this condition.

**Wasserstein Distances.** An important class of distance metrics between two probability measures $\mu, \nu$ 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 $\ell_1$ distance between the discrete distributions obtained by restricting $\mu$ and $\nu$ to a given partition $\mathcal{Q}$ of the input space. Thus for a fixed partition $\mathcal{Q}$, our analysis of $\ell_1$ 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: $\ell_2^2$, $\ell_1$, $f$-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.

# A. Table of Symbols

| Symbols | Description |
|---:|---|
| | **Preliminaries** |
| $K$ | number of probability distributions. |
| $(P_i)_{i=1}^K$ | the $K$ probability distributions. |
| $\mathcal{X}$ | $\mathcal{X} = \{x_1, x_2, \dots, x_l\}$, the support of $(P_i)_{i=1}^K$ of size $l$. |
| $\Delta_l$, $\Delta_l^{(\eta)}$ | the $(l-1)$ dimensional probability simplex, and its $\eta$-interior. |
| $D$ | a general distance measure $D : \Delta_l \times \Delta_l \mapsto [0, \infty)$. |
| $D_{\ell_2}, D_{\ell_1}, D_f, D_{KL}, D_s$ | $\ell_2^2$ distance, $\ell_1$ distance, $f$-divergence, KL-divergence and separation distance. |
| $\gamma_i(T_i)$ | The exact objective function in (1), i.e., $\mathbb{E}\left[D\left(\hat{P}_i, P_i\right)\right]$ with $\hat{P}_i$ constructed from $T_i$ i.i.d. samples. |
| $\mathcal{A}^*$, $(T_i^*)_{i=1}^K$ | The oracle allocation scheme and the oracle allocation. Solution to (1) assuming full knowledge of $(P_i)_{i=1}^K$. |
| $T_{i,t}$ | Number of times distribution $i$ has been sampled by an algorithm $\mathcal{A}$ before time $t$. |
| $T_i$ | Total number of times distribution $i$ is sampled by an algorithm till the budget is exhausted. Note that $T_i = T_{i,n+1}$. |
| $Z_{ij}^{(s)}$ | Indicator that the $s^{th}$ sample from arm $i$ is $x_j$. (Bernoulli with probability $p_{ij}$) |
| $\tilde{Z}_{ij}^{(s)}$ | $Z_{ij}^{(s)} - p_{ij}$. centered version of $Z_{ij}^{(s)}$. |
| $W_{ij,t}$ | $\sum_{s=1}^t \tilde{Z}_{ij}^{(s)}$ |
| $\hat{p}_{ij,t}$ | empirical estimate of $p_{ij}$ at time $t$, $\hat{p}_{ij,t} = W_{ij,t}/T_i$ (assuming $T_i > 0$). |
| $\hat{p}_{ij}$ | The empirical estimate of $p_{ij}$ constructed by an algorithm after the sampling budget is exhausted. Note that $\hat{p}_{ij} = \hat{p}_{ij,n+1}$. |
| $\mathcal{L}_n(\mathcal{A}, D)$ | Risk incurred by algorithm $\mathcal{A}$, for distance $D$ and sampling budget $n$. (Defined in (2)) |
| $\mathcal{R}_n(\mathcal{A}, D)$ | Regret of algorithm $\mathcal{A}$. (Defined in (3)) |
| $\varphi(c_i, T_i)$ | a *regular* function. (see Definition 1). |
| $\mathcal{F}$ | the class of all regular functions. |
| $g_i*$ and $h_i^*$ | $g_i^* := \frac{\partial \varphi(c, \widetilde{T}_i^*)}{\partial c}\Big\vert_{c=c_i}$ and $h_i^* := \frac{\partial \varphi(c_i, T)}{\partial T}\Big\vert_{T=\widetilde{T}_i^*}$. |
| $\widetilde{\mathcal{A}}^*$, $(\widetilde{T}_i^*)_{i=1}^K$ | Solution to (4) assuming full knowledge of parameters $(c_i)_{i=1}^K$. When the objective function $\varphi$ of (4) is an approximation of the objective (1), then we shall refer to $\widetilde{\mathcal{A}}^*$ and $(\widetilde{T}_i^*)_{i=1}^K$ as the *approx-oracle* allocation rule and the *approx-oracle* allocation respectively. |
| $A$, $B$, and $C$ | Terms defined in Lemma 1 for a general adaptive algorithm described in Alg. 1. |
| | **Adaptive scheme for $\ell_2$** |
| $\mathcal{E}_1$, $\delta_t$ | event defined in Lemma 2, and $\delta_t = (6\delta)/(Kl\pi^2 t^2)$ |
| $e_{ij,t}^{(\ell_2)}$, and $e_{i,t}^{(\ell_2)}$ | $\sqrt{3p_{ij}\log(2/\delta_t)/T_{i,t}}$, and $\sqrt{27\log(1/\delta_t)/T_{i,t}}$ resp. |
| $c_i^{(\ell_2)}$, $\hat{c}_{i,t}^{(\ell_2)}$ and $u_{i,t}^{(\ell_2)}$ | $1 - \sum_{j=1}^l p_{ij}^2$ and $1 - \sum_{j=1}^l \hat{p}_{ij,t}^2$ and $\hat{c}_{i,t}^{(\ell_2)} + e_{i,t}^{(\ell_2)}$ |
| $\lambda_i^{(\ell_2)}$, $\lambda_{\min}^{(\ell_2)}$, $\lambda_{\max}^{(\ell_2)}$ | $c_i^{(\ell_2)}/(\sum_k c_k^{(\ell_2)})$, min and max of $\lambda_i^{(\ell_2)}$. |

| | |
|---|---|
| $M_{\ell_2}$ | $\frac{\lambda_{\max}^{(\ell_2)}\sqrt{2\log(1/\delta_n)}}{\lambda_{\min}^{(\ell_2)}\sum_{i=1}^{K} c_i^{(\ell_2)}}$ |

**Adaptive scheme for $\ell_1$**

| | |
|---|---|
| $\mathcal{E}_2$, $\delta_t$ | event defined in Lemma 3, and $\delta_t = (3\delta)/(Kl\pi^2t^2)$ |
| $e_{ij,t}^{(\ell_1)}$, and $e_{i,t}^{(\ell_1)}$ | $\sqrt{2\log(2/\delta_t)/T_{i,t}}$, and $\sqrt{2l^2\log(2/\delta_t)/T_{i,t}}$ resp. |
| $c_i^{(\ell_1)}$, $\hat{c}_{i,t}^{(\ell_1)}$ and $u_{i,t}^{(\ell_1)}$ | $\sum_{j=1}^{l}\sqrt{p_{ij}(1-p_{ij})}$ and $\sum_j\sqrt{\hat{p}_{ij,t}(1-\hat{p}_{ij,t})}$ and $\hat{c}_{i,t}^{(\ell_1)} + e_{i,t}^{(\ell_1)}$ |
| $\lambda_i^{(\ell_1)}$, $\lambda_{\min}^{(\ell_1)}$, $\lambda_{\max}^{(\ell_1)}$ | $(c_i^{(\ell_1)})^2/(\sum_k(c_k^{(\ell_1)})^2)$, min and max of $\lambda_i^{(\ell_2)}$. |
| $M_{\ell_1}$ | $\frac{2l\lambda_{\max}^{(\ell_1)}\sqrt{\log(1/\delta_n)}}{C_{\ell_1}\lambda_{\min}^{(\ell_1)}}$ |

**Adaptive scheme for $D_f$ and $D_{\text{KL}}$**

| | |
|---|---|
| $D_f^{(r)}(\hat{P}_i, P_i)$, $R_{i,r+1}$ | The $r$-term Taylor's approximation of $D_f(\hat{P}_i, P_i)$, and the remainder term. Furthermore, we have $R_{i,r+1} = \sum_{j=1}^{l} R_{ij,r+1}$ where $R_{ij,r+1}$ is defined in (7). |
| $\varphi_i$ | the approximate objective function obtained by taking the expectation of $D_f^{(r)}$. We do not use $\varphi(c_i, \cdot)$ to represent it, the derivation of $c_i$ requires the knowledge of $f^{(m)}(1)$ values for $1 \leq m \leq r$. |
| $C_1$ | constant in assumption **(f2)** in Sec. 4.4 |
| $C_{f,r+1}$ | Constant defined in Eqs. 23 24 and 28 in Appendix F.2. |
| $\mathcal{E}_\delta$ | The $(1-\delta)$ probability event used to characterize a general adaptive scheme, in which we have $\tau_{0,i} \leq T_i \leq \tau_{1,i}$ for non-negative constants $\tau_{0,i}$ and $\tau_{1,i}$ |
| $\beta_m^{(ij)}(\tau_{0,i}, \tau_{1,i})$ | Defined in Lemma 9 |
| $\Psi_i$, $\psi_{1,i}, \psi_{2,i}, \psi_{3,i}$ and $\psi_4$ | Defined in Theorem 7. |
| $c_i^{(\text{KL})}$ | $(1/12)(\sum_{j=1}^{l} 1/p_{ij} - 1)$ |
| $e_{ij,t}$ and $e_{i,t}$ | $\sqrt{2\log(2/\delta_t)/T_{i,t}}$ and $\sqrt{(32l^2\log(2/\delta_t))/T_{i,t}}$. |
| $M_{\text{KL}}$ | $\sqrt{96K\log(1/\delta_n)}\eta^3$. |

**Adaptive scheme for $D_{\text{s}}$**

| | |
|---|---|
| $p_{i,S}$ for $S \subset [l]$, $S \neq \emptyset$ | $\sum_{j \in S} p_{ij}$ |
| $c_i^{(s)}$ | $\sum_{j=1}^{l}\rho_1(p_{ij})$ where $\rho_1(p) = \sqrt{(1-p)/p}$ |
| $\tilde{c}_i^{(s)}$ | $\max_{S \subset [l]}\rho_2(p_{i,S})$ where $\rho_2(p) = \rho_1(p) + \rho_1(1-p)$ |
| $a_{i,t}$ and $b_{i,t}$ | $\big(32\log(2/\delta_t)/T_{i,t}\big)^{1/4}$ and $\big(\frac{l\,a_{i,t}}{\eta}\big)\max\big\{1, \frac{a_{i,t}}{2\eta^{3/2}}\big\}$ |
| $E$ | $A\tilde{e}_n/B$ |
| $N_0$ | $\min\{n \geq 1 \;:\; (n/\log(2/\delta_n)) \geq 2/(\lambda_{\min}^{(s)}\eta^6)\}$ |

Table A.1: Table of symbols used in the paper.

## B. Regret with approximate objective function

In this section, we restate and prove Lemma 1 introduced in Section 4.1, which provides a decomposition of the regret when using an approximate objective function $\varphi_i(T_i)$ in place of $\mathbb{E}[D(\hat{P}_i, P_i)]$, where $\hat{P}_i$ is the empirical estimate of $P_i$ constructed using $T_i$ i.i.d. samples.

**Note:** In this section, we use the term $\varphi_i$ to represent any approximation ,and not just the approximations $\varphi(c_i, \cdot)$ introduced in Sec. 3, of the true objective function $\gamma_i$ in (1). This is because the stated result does not require the regularity assumptions to be satisfied by the approximation.

**Lemma 8.** *Suppose $\mathcal{A}$ is any allocation scheme for a loss function $D$ and sampling budget $n$. Let $\varphi_i(T_i)$ be an approximation for $\gamma_i(T_i) \coloneqq \mathbb{E}[D(\hat{P}_i, P_i)]$, where $\hat{P}_i$ is the empirical estimate of $P_i$ constructed using $T_i$ i.i.d. samples. Define the remainder term $R_i(T_i) \coloneqq \gamma_i(T_i) - \varphi_i(T_i)$. Let $\tilde{\mathcal{A}}^*$ and $\mathcal{A}^*$ represent the oracle and the approx-oracle allocation schemes, and let $(T_i^*)_{i=1}^K$ and $(\widetilde{T}_i^*)_{i=1}^K$ denote their allocations respectively. Then, assuming that $(\gamma_i(\cdot))_{i=1}^K$ are non-increasing functions, we have the following:*

$$\mathcal{R}_n(\mathcal{A}, D) \coloneqq \mathcal{L}_n(\mathcal{A}, D) - \mathcal{L}_n(\tilde{\mathcal{A}}^*, D) \le \mathcal{L}_n(\mathcal{A}, D) - \mathcal{L}_n(\mathcal{A}^*, D) + 3 \max_{i \in [K]} |R_i(\widetilde{T}_i^*)| \quad \textit{and} \tag{12}$$

$$\mathcal{R}_n(\mathcal{A}, D) \le \mathcal{L}_n(\mathcal{A}, D) - \varphi_i(\widetilde{T}_i^*) + 2 \max_{i \in [K]} |R_i(\widetilde{T}_i^*)| \tag{13}$$

*Proof.* Note that we have $\gamma_i(T_i) = \varphi_i(T_i) + R_i(T_i)$. By definition of regret, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{A}, D) &= \max_{1 \le i \le K} \left( \mathbb{E}[D(\hat{P}_i, P_i)] - \gamma_i(T_i^*) \right) \\
&\overset{(a)}{\le} \max_{1 \le i \le K} \left( \mathbb{E}[D(\hat{P}_i, P_i)] - \varphi_i(\widetilde{T}_i^*) \right) + \max_{1 \le i \le K} \left| \varphi_i(\widetilde{T}_i^*) - \gamma_i(T_i^*) \right| \\
&\overset{(b)}{=} \mathcal{L}_n(\mathcal{A}, D) - \varphi_i(\widetilde{T}_i^*) + \max_{1 \le i \le K} \left| \varphi_i(\widetilde{T}_i^*) - \gamma_i(T_i^*) \right| . \\
&\overset{(c)}{\le} \mathcal{L}_n(\mathcal{A}, D) - \mathcal{L}_n(\widetilde{\mathcal{A}}^*, D) + \max_{1 \le i \le K} |R_i(\widetilde{T}_i^*)| + \max_{1 \le i \le K} \left| \varphi_i(\widetilde{T}_i^*) - \gamma_i(T_i^*) \right|
\end{aligned}$$

In the above display,
**(a)** follows from an application of triangle inequality,
**(b)** uses the fact that by definition of the approx-oracle rule, we must have $\varphi_i(\widetilde{T}_i^*) = \varphi_j(\widetilde{T}_j^*)$ for all $i, j \in [K]$.
**(c)** follows from another application of triangle inequality.

To complete the proof, it suffices to show that the term $\max_{1 \le i \le K} \left| \varphi_i(\widetilde{T}_i^*) - \gamma_i(T_i^*) \right|$ is less than $2 \max_{1 \le i \le K} |R_i(\widetilde{T}_i^*)|$. Note that since $\widetilde{\mathcal{A}}^*$ is the approx-oracle sampling scheme, there must exist a $\lambda \in [0, \infty)$ such that $\varphi_i(\widetilde{T}_i^*) = \lambda$ for all $i \in [K]$. Introduce the notation $\Delta T_i^* = \widetilde{T}_i^* - T_i^*$. Then we consider two cases.

**Case 1: $\boldsymbol{\Delta T_i^* = 0}$ for all $\boldsymbol{i}$.** In this case, we have $\widetilde{T}_i^* = T_i^*$ for all $i$, which implies that $|\varphi_i(\widetilde{T}_i^*) - \gamma_i(T_i^*)| \le |R_i(\widetilde{T}_i^*)|$ for all $i$, thus proving that $\max_{i \in [K]} |\varphi_i(\widetilde{T}_i^*) - \gamma_i(T_i^*)| \le \max_{i \in [K]} |R_i(\widetilde{T}_i^*)|$.

**Case 2: $\boldsymbol{\Delta T_k^* \neq 0}$ for some $k \in [K]$.** Since we have $\sum_{i=1}^K T_i^* = \sum_{i=1}^K \widetilde{T}_i^*$, it means that $\sum_{i=1}^K \Delta T_i^* = 0$. Thus there must exist $i, j \in [K]$ such that $\Delta T_i^* > 0$ and $\Delta T_j^* < 0$. Define $i_0 = \arg\min_{i \in [K]} \gamma_i(\widetilde{T}_i^*)$ and $i_1 = \arg\max_{i \in [K]} \gamma_i(\widetilde{T}_i^*)$. By monotonicity assumption on the functions $\gamma_i$, we must have the following $\gamma_{i_0}(\widetilde{T}_{i_0}^*) < \lambda < \gamma_{i_1}(\widetilde{T}_{i_1}^*)$. We claim that this implies that

$$\gamma_{i_0}\left(\widetilde{T}_{i_0}^*\right) \le \gamma_i(T_i^*) \le \gamma_{i_1}(\widetilde{T}_{i_1}^*). \tag{14}$$

If (14) is true, then the result of the proposition immediately follows from the observation that $\lambda - \max_{k \in [K]} |R_k(\widetilde{T}_k^*)| \le \gamma_i(\widetilde{T}_i^*) \ \le \lambda + \max_{k \in [K]} |R_k(\widetilde{T}_k^*)|$, which coupled with (14) implies that $\lambda - \gamma_i(T_i^*) \le 2 \max_{k \in K} |R_k\left(\widetilde{T}_k^*\right)|$.

Finally, it remains to show that (14) is true. We prove this by contradiction. Introduce the notation $\mu = \gamma_i(T_i^*)$ for all $i \in [K]$ (by the definition of the oracle allocation rule). Next, suppose that $\mu \notin [\lambda - \max_k |R_k(\widetilde{T}_k^*)|, \ \lambda + \max_k |R_k(\widetilde{T}_k^*)|]$.

Without loss of generality, assume that $\mu < \lambda - \max_k |R_k(\widetilde{T}_k^*)|$ (the other case can be argued similarly). This means that $\mu = \gamma_i(T_i^*) < \gamma_i(\widetilde{T}_i^*)$ for all $i$. By the monotonicity of $\gamma_i$ this implies that all the $\Delta T_i^* = \widetilde{T}_i^* - T_i^*$ have the same sign. However, since $\sum_i \Delta T_i^* = 0$, this can only happen if $\Delta T_i^* = 0$ for all $i$. This contradicts the defining assumption of **Case 2** above.

□

## C. Proof of General Tracking Result (Theorem 1)

Assume that arm $k$ was played at least once during the period $K \leq t \leq n-1$. Let $t_k$ be the time at which arm $k$ was played for the last time. Recall that for any $t \geq 1$, we use $T_{i,t}$ to denote the number of times the arm $i$ is played before the start of round $t$, and $T_i$ denotes the total number of times arm $i$ is played until the total sampling budget is exhausted, i.e., $T_i = T_{i,n+1}$ Then we have the following for any $1 \leq j \leq K$:

$$\varphi\left(u_{j,t_k}, T_{j,t_k}\right) \overset{(a)}{\leq} \varphi\left(u_{k,t_k}, T_{k,t_k}\right) \overset{(b)}{\leq} \varphi\left(c_k + 2e_{k,t_k}, T_{k,t_k}\right) \tag{15}$$

In the above display, **(a)** follows from the arm selection rule for $t \geq K$ and **(b)** follows from the fact that $u_{k,t_k} - c_k \leq 2e_{k,t_k}$ and that $\varphi$ is non-decreasing in its first argument.

Next, we define $i = \arg\min_{1\leq k\leq K} \varphi(c_k, T_k)$. Using the fact that $\varphi$ is non-increasing in its second argument, we have $T_i \geq \widetilde{T}_i^*$. We now consider two cases:

**Case 1:** $T_i = \widetilde{T}_i^*$ . In this case, we must have $T_j = \widetilde{T}_j^*$ for all $1 \leq j \leq K$. This is because if $\min_k \varphi(c_k, T_k) = \varphi(c_k, \widetilde{T}_k^*)$, then $\varphi(c_k, T_k) = \varphi(c_k, \widetilde{T}_k^*)$ for all $1 \leq k \leq K$, and thus $T_k = \widetilde{T}_k^*$ for all $1 \leq k \leq K$. In this case the result of Lemma 1 holds trivially.

**Case 2:** $T_i > \widetilde{T}_i^*$**.** In this case, the arm $i$ must have been played at least once during the time interval $K \leq t \leq n-1$. Denote that time by $t_i$. Defining $g_i^* := \frac{\partial \varphi(c_i,T)}{\partial T}\Big|_{T=\widetilde{T}_i^*}$, we have the following sequence of inequalities for any $1 \leq j \leq K$:

$$\begin{aligned}\varphi(c_j, T_j) &\overset{(a)}{\leq} \varphi(c_i + 2e_{i,t_i}, T_i - 1) \overset{(b)}{\leq} \varphi\left(c_i + 2e_i^*, \widetilde{T}_i^*\right) \overset{(c)}{\leq} \varphi\left(c_i, \widetilde{T}_i^*\right) + 2g_i^* e_i^* \\ &\overset{(d)}{\leq} \varphi(c_i, \widetilde{T}_i^*) + 2Ae_i^* \overset{(e)}{\leq} \varphi(c_j, \widetilde{T}_j^*) + 2A\tilde{e}_n.\end{aligned} \tag{16}$$

In the above display,
**(a)** follows from the fact that since arm $i$ was played at time $t_i$ we must have $\varphi(u_{j,t_i}, T_{j,t_i}) \leq \varphi(u_{i,t_i}, T_{i,t_i})$. Furthermore, since $u_{j,t_i} \geq c_j$ under the event $\mathcal{E}$ by definition, and the fact that $\varphi$ is non-decreasing in its first argument implies that $\varphi(c_j, T_{j,t_i}) \leq \varphi(u_{j,t_i}, T_{j,t_i})$,
**(b)** follows from the fact that $T_i - 1 \geq \widetilde{T}_i^*$ (defining assumption of Case 2), that $\varphi$ is non-increasing in its second argument, and that $e_i(\cdot)$ is a non-increasing in its argument and $\varphi$ is non-decreasing in its first argument,
**(c)** follows from the fact that $\varphi$ is a concave function in its first argument and thus is majorized by its linear approximation,
**(d)** follows from the fact that $A = \max_{1\leq j\leq K}\ g_j^* \geq g_i^*$ by definition, and
**(e)** uses the fact that by definition of $\widetilde{T}_j^*$ we have $\varphi(c_j, \widetilde{T}_j^*) = \varphi(c_i, \widetilde{T}_i^*)$ and that $e_i^* \leq \max_{1\leq k\leq K} e_k^* := \tilde{e}_n$.

Next, by defining $h_j^* = \frac{\partial \varphi\left(c, \widetilde{T}_j^*\right)}{\partial c}\Big|_{c=c_j}$ and from the relation between the first and last terms of (16) we observe the following:

$$\varphi(c_j, \widetilde{T}_j^*) + 2A\tilde{e}_n \geq \varphi(c_j, T_j) \overset{(a)}{\geq} \varphi(c_j, \widetilde{T}_j^*) + h_j^*\left(T_j - \widetilde{T}_j^*\right). \tag{17}$$

In the above display, **(a)** follows from the assumption that $\varphi$ is a convex function in its second argument. (17) implies that

$$h_j^*(T_j - \widetilde{T}_j^*) \leq 2A\tilde{e}_n \quad \Rightarrow T_j - \widetilde{T}_j^* \overset{(a)}{\geq} -\frac{2A\tilde{e}_n}{|h_j^*|} \overset{(b)}{\geq} -\frac{2A\tilde{e}_n}{B} \quad \Rightarrow T_j \geq \widetilde{T}_j^* - \frac{2A\tilde{e}_n}{B}.$$

In the above display, **(a)** follows from the assumption that $\varphi$ is non-increasing function of its second argument, and **(b)** follows from the definition of $B = |\max_{1 \leq j \leq K} \; h_j^*|$ and the assumption that $B > 0$. This completes the proof for the lower bound on $T_j$ for $1 \leq j \leq K$.

Next, we obtain the upper bound on $T_j$ as follows:

$$T_j = n - \sum_{i \neq j} T_i \leq n - \sum_{i \neq j} \widetilde{T}_i^* + 2(K-1)A\tilde{e}_n/B = \widetilde{T}_j^* + 2(K-1)A\tilde{e}_n/B.$$

## D. Analysis for $\ell_2^2$ loss

### D.1. Proof of Lemma 2

Write $\mathcal{E}_1 = \cap_{t,i,j} \mathcal{E}_1^{(t,i,j)}$ where the events $\mathcal{E}_1^{(t,i,j)} := \{|\hat{p}_{ij,t} - p_{ij}| \leq e_{ij,t}\}$ and $e_{ij,t} = \sqrt{\log(2/\delta_t)/(2T_{i,t})}$. Then by union bound we have that $Pr\left((\mathcal{E}_1)^c\right) \leq \sum_{t=1}^{\infty} \sum_{i=1}^{K} \sum_{j=1}^{l} Pr\left(\left(\mathcal{E}_1^{(t,i,j)}\right)^c\right)$. To show that $Pr(\mathcal{E}_1^c) \leq \delta$, it suffices to show that $Pr\left(\left(\mathcal{E}_1^{(t,i,j)}\right)^c\right) \leq \delta_t = \frac{6\delta}{Klt^2\pi^2}$. We proceed by using Hoeffding's inequality for bounded random variables.

$$\begin{aligned} Pr\left(|\hat{p}_{ij,t} - p_{ij}| > e_{ij,t}^{(\ell_2)}\right) &= \sum_{s=0}^{t} Pr\left(|\hat{p}_{ij,t} - p_{ij}| > eij, t^{(\ell_2)}, \, T_{i,t} = s\right) \\ &\leq \sum_{s=0}^{t} 2Pr\left(T_{i,t} = s\right) \exp\left(\frac{-2s\log(2/\delta_t)}{2s}\right) = \delta_t. \end{aligned}$$

With the above result, we can construct a confidence interval for the parameter $c_i^{(\ell_2)}$. Under the event $\mathcal{E}_1$, we have

$$\begin{aligned} &\sum_{j=1}^{l} \hat{p}_{ij,t}^2 \geq \sum_{j=1}^{l} \left(p_{ij}^2 + (e_{ij,t}^{(\ell_2)})^2 - 2e_{ij,t}^{(\ell_2)} p_{ij}\right) \\ \Rightarrow &\sum_{j=1}^{l} \hat{p}_{ij,t}^2 \overset{(a)}{\geq} \sum_{j=1}^{l} p_{ij}^2 - 2\sqrt{\frac{\log(2/\delta_t)}{2T_{i,t}}} \left(\sum_{j=1}^{l} p_{ij}\right) = \sum_{j=1}^{l} p_{ij}^2 - 2\sqrt{\frac{\log(2/\delta_t)}{2T_{i,t}}}. \end{aligned}$$

where $(a)$ uses the fact that $e_{ij,t}^{(\ell_2)} \geq 0$.

Similarly, by using the fact that $e_{ij,t}^{(\ell_2)} \leq 1$, we can obtain the following:

$$\sum_{j=1}^{l} \hat{p}_{ij,t}^2 \leq \sum_{j=1}^{l} \left(p_{ij}^2 + (e_{ij,t}^{(\ell_2)})^2 + 2e_{ij,t}^{(\ell_2)} p_{ij}\right) \leq \sum_{j=1}^{l} p_{ij}^2 + \sum_{j=1}^{l} \left((1 + 2p_{ij})e_{ij,t}^{(\ell_2)}\right) = \sum_{j=1}^{l} p_{ij}^2 + (l+2)\sqrt{\frac{\log(2/\delta_t)}{2T_{i,t}}}.$$

Combining the inequalities from the previous two displays, we get the required confidence interval for the term $c_i^{(\ell_2)}$ around its empirical counterpart $\hat{c}_{i,t}^{(\ell_2)} := 1 - \sum_{j=1}^{l} \hat{p}_{ij,t}^2$, as follows:

$$|c_i^{(\ell_2)} - \hat{c}_{i,t}^{(\ell_2)}| \leq (l+2)\sqrt{\frac{\log(2/\delta_t)}{2T_{i,t}}} := e_{i,t}^{(\ell_2)}.$$

### D.2. Proof of Theorem 2

*Proof.* We begin by obtaining the constants $A$, $\tilde{e}_n$ and $B$ from Lemma 1. We introduce the notation $\lambda_i^{(\ell_2)} := \frac{c_i^{(\ell_2)}}{\sum_{k=1}^K c_k^{(\ell_2)}}$, $C_{\ell_2} := \sum_{i=1}^K c_i^{(\ell_2)}$, $\lambda_{\max}^{(\ell_2)} = \max_{1\le i\le K} \lambda_i^{(\ell_2)}$, and $\lambda_{\min}^{(\ell_2)} = \min_{1\le i\le K} \lambda_i^{(\ell_2)}$.

$$\begin{aligned}
A &= \max_{1\le i\le K} \frac{1}{\widetilde{T}_i^*} = \frac{1}{\lambda_{\min}^{(\ell_2)} n} = \mathcal{O}\left(\frac{1}{n}\right), \\
B &= \min_{1\le i\le K} \frac{c_i^{(\ell_2)}}{\left(\widetilde{T}_i^*\right)^2} = \frac{C_{\ell_2}}{\lambda_{\max}^{(\ell_2)} n^2} = \mathcal{O}\left(\frac{1}{n^2}\right), \\
\tilde{e}_n &= \max_{1\le i\le K} e_i^* \le \sqrt{\frac{(l+2)^2 \log(2/\delta_n)}{2\lambda_{\min}^{(\ell_2)} n}} = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right).
\end{aligned}$$

The above calculations imply that we have

$$\frac{A\tilde{e}_n}{B} = \frac{\lambda_{\max}^{(\ell_2)} \sqrt{(l+2)^2 \log(2/\delta_n)/2}}{\lambda_{\min}^{(\ell_2)} C_{\ell_2}} \sqrt{n} := M_{\ell_2} \sqrt{n}. \tag{18}$$

Before proceeding, recall that we drop the $n+1$ from the subscript when referring to the final probability mass estimates, i.e., we use $\hat{p}_{ij}$ instead of $\hat{p}_{ij,n+1}$ to refer to the estimate of $p_{ij}$ after the end of $n$ rounds. Next, consider any arm $i$, and decompose the expected $\ell_2^2$ distance as follows:

$$\mathbb{E}\left[D_{\ell_2}\left(\hat{P}_i, P_i\right)\right] = \sum_{j=1}^{l} \mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^2\right] = \sum_{j=1}^{l} \mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^2 \left(\mathbb{1}_{\{\mathcal{E}_1\}} + \mathbb{1}_{\{\mathcal{E}_1^c\}}\right)\right].$$

We now consider each term of the summation above separately.

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^2 \mathbb{1}_{\{\mathcal{E}_1^c\}}\right] &\le 1 \times Pr\left(\mathcal{E}_1^c\right) \overset{(a)}{\le} \delta \\
\Rightarrow \sum_{j=1}^{l} \mathbb{E}\left[(\hat{p}_{ij} - p_{ij})^2 \mathbb{1}_{\{\mathcal{E}_1^c\}}\right] &\le l\delta.
\end{aligned}$$

In the above display, the inequality **(a)** follows from the statement of Lemma 2.

Next, we proceed as follows:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^2 \mathbb{1}_{\{\mathcal{E}_1\}}\right] &= \mathbb{E}\left[\frac{1}{T_i^2}\left(\sum_{s=1}^{T_i} Z_{ij}^{(s)} - p_{ij}\right)^2 \mathbb{1}_{\{\mathcal{E}_1\}}\right] \le \mathbb{E}\left[\left(\sup_{\omega\in\mathcal{E}_1} \frac{1}{T_i^2}\right)\left(\sum_{s=1}^{T_i} Z_{ij}^{(s)} - p_{ij}\right)^2\right] \\
&\overset{(a)}{\le} \left(\frac{1}{T_i^* - A\tilde{e}_n/B}\right)^2 \mathbb{E}\left[\left(\sum_{s=1}^{T_i} Z_{ij}^{(s)} - p_{ij}\right)^2\right] \\
&\overset{(b)}{=} \left(\frac{1}{T_i^* - A\tilde{e}_n/B}\right)^2 \left(p_{ij}(1-p_{ij})\right) \mathbb{E}\left[T_i\right] \\
&\overset{(c)}{\le} \left(\frac{1}{T_i^* - A\tilde{e}_n/B}\right)^2 \left(p_{ij}(1-p_{ij})\right) \left(1 \times \left(T_i^* + \frac{(K-1)A\tilde{e}_n}{B}\right) + \delta \times n\right) \\
&\overset{(d)}{\le} \frac{p_{ij}(1-p_{ij})}{\left(T_i^*\right)^2} \left(1 + \frac{6A\tilde{e}_n}{BT_i^*}\right) \left(T_i^* + \frac{(K-1)A\tilde{e}_n}{B} + n\delta\right) \\
&= \frac{p_{ij}(1-p_{ij})}{\left(T_i^*\right)^2} \left(T_i^* + \frac{(K-1)A\tilde{e}_n}{B} + n\delta + \frac{6A\tilde{e}_n}{B} + \frac{6(K-1)A^2\tilde{e}_n^2}{B^2 T_i^*} + \frac{6A\tilde{e}_n n\delta}{BT_i^*}\right)
\end{aligned}$$

In the above display, **(a)** follows from the result of Lemma 1,
**(b)** follows from an application of Wald's Lemma,
**(c)** follows from the facts that $Pr(\mathcal{E}_1^c) \leq \delta$ , $T_i \leq n$ a.s., and the bounds on $T_i$ under the event $\mathcal{E}_1$ given by Lemma 1,
**(d)** follows from the fact that the function $x \mapsto 1/x^2$ is convex, and thus for the function lies below the chord joining the points $(1, 1/1^2)$ and $\left(1/2, 1/(1/2)^2\right)$, and the assumption that $n$ is large enough to ensure that $\frac{A\tilde{e}_n}{BT_i^*} < 1/2$. A sufficient condition for this is that $n > \frac{4(M_{\ell_2})^2}{\left(\lambda_{\min}^{(\ell_2)}\right)^2}$.

Next, we have the following:

$$\mathbb{E}\left[(\hat{p}_{ij}^2 - p_{ij}^2)^2 \mathbb{1}_{\{\mathcal{E}_1\}}\right] \leq \frac{(p_{ij}(1-p_{ij})}{T_i^*} \left(1 + \frac{n\delta}{T_i^*}\left(1 + \frac{6M_{\ell_2}\sqrt{n}}{T_i^*}\right) + \frac{M_{\ell_2}\sqrt{n}}{T_i^*}\left((K+5) + \frac{6(K-1)M_{\ell_2}\sqrt{n}}{\sqrt{T_i^*}}\right)\right).$$

Since $T_i^* = \lambda_i^{(\ell_2)} n$, we have

$$\mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^2 \mathbb{1}_{\{\mathcal{E}_1\}}\right] \leq \frac{p_{ij}(1-p_{ij})}{T_i^*} \left(1 + \frac{(K+5)M_{\ell_2}}{\lambda_i^{(\ell_2)}\sqrt{n}} + \frac{\delta}{\lambda_i^{(\ell_2)}}\left(1 + 6M_{\ell_2}\sqrt{n}\right) + \frac{6(K-1)\left(M_{\ell_2}\right)^2}{\left(\lambda_i^{(\ell_2)}\right)^2 n}\right)$$

Finally, summing up over the values of $j$ in the range $\{1, 2, \ldots, l\}$, we get

$$\sum_{j=1}^{l} \mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^2 \mathbb{1}_{\{\mathcal{E}_1\}}\right] \leq \sum_{j=1}^{l} \frac{p_{ij}(1-p_{ij})}{T_i^*} + \frac{(K+5)lM_{\ell_2}}{\left(\lambda_i^{(\ell_2)}\right)^2 n^{3/2}} + \mathcal{O}\left(\frac{\delta l\sqrt{\log n}}{\sqrt{n}} + \frac{\log(n)}{n^2}\right).$$

We can select $\delta$ small enough to ensure that the regret is of the order $\tilde{\mathcal{O}}\left(n^{-3/2}\right)$. A suitable choice for this is $\delta = n^{-5/2}$. □

## E. Analysis for $\ell_1$ loss

### E.1. Proof of Lemma 3

*Proof.* We can again write the event $\mathcal{E}_2 := \cap_t \cap_{i=1}^K \cap_{j=1}^l \mathcal{E}_2^{(t,i,j)}$. It suffices to show that $Pr\left(\mathcal{E}_2^{(t,i,j)}\right) \geq 1 - \delta_t$, and the result follows from an application of the union bound and the fact that $\sum_{t,i,j} \delta_t = \delta$.

To show that $Pr\left(\mathcal{E}_2^{(t,i,j)}\right) \geq 1 - \delta_t$, we employ (Maurer & Pontil, 2009, Theorem 10) to the collection of random variables $\left(Z_{ij}^{(s)}\right)_{s=1}^{T_{i,t}}$ defined as $Z_{ij}^{(s)} = \mathbb{1}_{\{X_i^{(s)}=j\}}$. Then we have $\hat{p}_{ij,t} = \frac{\sum_{s=1}^t Z_{ij}^{(s)}}{t}$, and $Var\left(Z_{ij}^{(s)}\right) = p_{ij}(1-p_{ij})$, and we obtain the required result by applying (Maurer & Pontil, 2009, Eq. (3) and (4)):

$$Pr\left(\left|\sqrt{\hat{p}_{ij,t}\left(1 - \hat{p}_{ij,t}\right)} - \sqrt{p_{ij}(1-p_{ij})}\right| > e_{ij,t}^{(\ell_1)}\right) \leq \delta_t$$

□

### E.2. Proof of Theorem 3

**Value of the constants** $A$, $B$ **and** $\tilde{e}_n$. We begin by obtain the values of the constants $A$, $B$ and $\tilde{e}_n$ from Lemma 1 corresponding to the $\ell_1$ loss function.

Recall the notation $\lambda_i^{(\ell_1)} := \frac{\left(c_i^{(\ell_1)}\right)^2}{\sum_{k=1}^K \left(c_k^{(\ell_1)}\right)^2}$, $C_{\ell_1}^2 = \sum_{k=1}^K \left(c_k^{(\ell_1)}\right)^2$, $\lambda_{\max}^{(\ell_1)} = \max_{1 \leq i \leq K} \lambda_i^{(\ell_1)}$ and $\lambda_{\min}^{(\ell_1)} = \min_{1 \leq i \leq K} \lambda_i^{(\ell_1)}$.

Then we have the following:

$$
\begin{aligned}
A &= \max_{1\leq i\leq K} \frac{1}{\sqrt{\widetilde{T}_i^*}} = \frac{1}{\sqrt{\lambda_{\min}^{(\ell_1)} n}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \\
B &= \min_{1\leq i\leq K} \frac{c_i^{(\ell_1)}}{2\left(\widetilde{T}_i^*\right)^{3/2}} = \frac{C_{\ell_1}}{\lambda_{\max} n^{3/2}} = \mathcal{O}\left(\frac{1}{n^{3/2}}\right), \\
\tilde{e}_n &= \max_{1\leq i\leq K} e_i^*) = \sqrt{\frac{2l^2\log(2/\delta_n)}{\widetilde{T}_i^*-1}} \overset{(a)}{\leq} \sqrt{\frac{4l^2\log(2/\delta_n)}{\widetilde{T}_i^*}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

In the above display, the inequality **(a)** follows from the assumption that $n$ is large enough to ensure that $\widetilde{T}_i^* \geq 3$ for all $1 \leq i \leq K$, which is implied by the assumption $n \geq 3/\lambda_{\min}^{(\ell_1)}$.

**Regret Derivation.** Since we use the approximate objective function for $\ell_1$ loss, we note that the regret can be written as follows:

$$
\mathcal{R}_n\left(\mathcal{A}_{\ell_1}, D_{\ell_1}\right) \leq \mathcal{L}_n\left(\mathcal{A}_{\ell_1}, D_{\ell_1}\right) - \varphi(c_i^{(\ell_1)}, \widetilde{T}_i^*) + 2\max_{1\leq i\leq K}|R_i\left(\widetilde{T}_i^*\right)|. \tag{19}
$$

We first bound the term in (19), $\mathcal{L}_n\left(\mathcal{A}_{\ell_1}, D_{\ell_1}\right) - \varphi\left(c_i^{(\ell_1)}, \widetilde{T}_i^*\right)$. Recall that in the final estimates of the probability mass function, we drop the $n+1$ from the subscript, i.e., we write $\hat{p}_{ij}$ and $\hat{P}_i$ instead of $\hat{p}_{ij,n+1}$ and $\hat{P}_{i,n+1}$. We next proceed as follows:

$$
\mathcal{L}_n\left(\mathcal{A}_{\ell_1}, D_{\ell_1}\right) = \mathbb{E}\left[\|\hat{P}_i - P_i\|_1\right] = \mathbb{E}\left[\|\hat{P}_i - P_i\|_1\left(\mathbb{1}_{\{\mathcal{E}_2\}} + \mathbb{1}_{\{\mathcal{E}_2^c\}}\right)\right] \leq \mathbb{E}\left[\|\hat{P}_i - P_i\|_1\mathbb{1}_{\{\mathcal{E}_2\}}\right] + l \times \delta.
$$

where the inequality follows from the fact that $\|\hat{P}_i - P_i\|_1 \leq l$ almost surely, and that $Pr\left(\mathcal{E}_2^c\right) \leq \delta$. Next, we expand the remaining term:

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{P}_i - P_i\|_1\mathbb{1}_{\{\mathcal{E}_2\}}\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{l}\sqrt{|\hat{p}_{ij}-p_{ij}|^2}\right)\mathbb{1}_{\{\mathcal{E}_2\}}\right] = \mathbb{E}\left[\left(\sum_{j=1}^{l}\sqrt{|\hat{p}_{ij}-p_{ij}|^2\mathbb{1}_{\{\mathcal{E}_2\}}}\right)\right] \\
&\overset{(a)}{\leq} \sum_{j=1}^{l}\sqrt{\frac{2}{\pi}\mathbb{E}\left[\left(\hat{p}_{ij}-p_{ij}\right)^2\mathbb{1}_{\{\mathcal{E}_2\}}\right]} + \mathcal{O}\left(\frac{1}{\left(\widetilde{T}_i^* - A\tilde{e}_n/B\right)^{3/2}}\right) \\
&\overset{(b)}{=} \sum_{j=1}^{l}\sqrt{\frac{2}{\pi}\mathbb{E}\left[\left(\hat{p}_{ij}-p_{ij}\right)^2\mathbb{1}_{\{\mathcal{E}_2\}}\right]} + \mathcal{O}\left(\frac{\sqrt{8}}{\left(\lambda_{\min}^{(\ell_1)}n\right)^{3/2}}\right) \\
&:= \sum_{j=1}^{l}\sqrt{\alpha_{ij}} + \mathcal{O}\left(\frac{\sqrt{8}}{\left(\lambda_{\min}^{(\ell_1)}n\right)^{3/2}}\right).
\end{aligned}
$$

The inequality **(a)** in the above display follows from the from an approximation relation between the mean absolute deviation and the standard deviation of Binomial distributions, as proved in (Blyth, 1980, Eq.(2.4)), while the inequality **(b)** follows from the definition of the optimal static allocation values $\widetilde{T}_i^* = \lambda_i^{(\ell_1)} n$ and the values of $A$, $B$ and $\tilde{e}_n$ derived above along with the assumption that $n$ is large enough to ensure that $\frac{A\tilde{e}_n}{B\lambda_{\min}^{(\ell_1)}} \leq 1/2$ or $n \geq 4\left(M_{\ell_1}\right)^2/\left(\lambda_{\min}^{(\ell_1)}\right)^2$.

Finally, we analyze the terms $\alpha_{ij}$:

$$\begin{aligned}\alpha_{ij} &\leq \mathbb{E}\left[\left(\sup_{\omega\in\mathcal{E}_2}\frac{1}{T_i^2}\right)\left(\sum_{s=1}^{T_i} Z_{ij}^{(s)} - p_{ij}\right)^2\right] \leq \frac{1}{\left(\widetilde{T}_i^* - \frac{A\tilde{e}_n}{B}\right)^2} p_{ij}\left(1-p_{ij}\right)\mathbb{E}\left[T_i\right] \\ &\leq \frac{p_{ij}(1-p_{ij})}{\left(\widetilde{T}_i^*\right)^2}\left(1 + 6\frac{A\tilde{e}_n}{B\widetilde{T}_i^*}\right)\left(\widetilde{T}_i^* + \frac{(K-1)A\tilde{e}_n}{B} + n\delta\right) \\ &= \frac{p_{ij}(1-p_{ij})}{\left(\widetilde{T}_i^*\right)^2}\left(\widetilde{T}_i^* + \frac{(K-1)A\tilde{e}_n}{B} + n\delta + 6\frac{A\tilde{e}_n}{B} + \frac{6(K-1)A^2\tilde{e}_n^2}{B^2\widetilde{T}_i^*} + \frac{6A\tilde{e}_n n\delta}{B\widetilde{T}_i^*}\right) \\ &\coloneqq \varphi\left(c_i^{(\ell_1)}, \widetilde{T}_i^*\right) + E_{ij},\end{aligned}$$

where $E_{ij}$ represents all the higher order terms. Thus, we get the following:

$$\sum_{j=1}^{l}\sqrt{\alpha_{ij}} \leq \sum_{j=1}^{l}\sqrt{\frac{p_{ij}(1-p_{ij})}{\widetilde{T}_i^*} + E_{ij}} \overset{(a)}{\leq} \sum_{j=1}^{l}\sqrt{\frac{p_{ij}(1-p_{ij})}{\widetilde{T}_i^*}} + \sum_{j=1}^{l}\sqrt{E_{ij}},$$

where **(a)** follows from the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y > 0$. Thus to complete the proof we need to analyze the term $E_{ij}$. First, we note that for the choice of the terms $A$, $B$ and $\tilde{e}_n$, we have

$$\frac{A\tilde{e}_n}{B} = \frac{\left(\sqrt{\frac{1}{\lambda_{\min}^{(\ell_1)}n}}\right)\left(\sqrt{\frac{4l^2\log(2/\delta_n)}{\lambda_{\min}^{(\ell_1)}n}}\right)}{\left(\frac{C_{\ell_1}}{\lambda_{\max}^{(\ell_1)}n^{3/2}}\right)} \coloneqq M_{\ell_1}\sqrt{n}.$$

We now rewrite $E_{ij}$ as follows:

$$\frac{E_{ij}}{p_{ij}(1-p_{ij})} = \frac{(K+5)M_{\ell_1}}{\left(\lambda_i^{(\ell_1)}\right)^2 n^{3/2}} + \frac{\delta}{n\left(\lambda_i^{(\ell_1)}\right)^2} + \frac{6(K-1)\left(M_{\ell_1}\right)^2}{\left(\lambda_i^{(\ell_1)}\right)^3 n^2} + \frac{6M_{\ell_1}\delta}{\left(\lambda_i^{(\ell_1)}\right)^3 n^{3/2}}.$$

By setting $\delta = 1/n$, we get the following:

$$\begin{aligned}\sum_{j=1}^{l}\sqrt{E_{ij}} &\leq \sum_{j=1}^{l}\left(\frac{\sqrt{p_{ij}(1-p_{ij})}}{\lambda_i^{(\ell_1)}}\left(\sqrt{\frac{(K+5)M_{\ell_1}}{n^{3/2}}} + \sqrt{\frac{\delta}{n} + \frac{6(K-1)\left(M_{\ell_1}\right)^2}{\lambda_i^{(\ell_1)}n^2} + \frac{6M_{\ell_1}\delta}{\lambda_i^{(\ell_1)}n^{3/2}}}\right)\right) \qquad (20) \\ &\overset{(a)}{\leq} \sqrt{\frac{(K+5)M_{\ell_1}}{n^{3/2}}}\sum_{j=1}^{l}\frac{\sqrt{p_{ij}(1-p_{ij})}}{\lambda_i^{(\ell_1)}} + \tilde{\mathcal{O}}\left(\frac{1}{n}\right) \\ &\overset{(b)}{\leq} \sqrt{\frac{(K+5)M_{\ell_1}}{n^{3/2}}}\left(\frac{l}{2\lambda_{\min}^{(\ell_1)}}\right) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right)\end{aligned}$$

where **(a)** follows from the assumption that $n$ is large enough to ensure that the first term dominates the remaining $\mathcal{O}\left(1/n\right)$ terms (a sufficient condition for this is that $n \geq \left(24\lambda_{\min}^{(\ell_1)}M_{\ell_1}\right)$, and **(b)** uses the fact that $p(1-p) \leq 1/4$ for all $p \in [0,1]$. Thus we have obtained the regret in tracking the approx-oracle solution, i.e.,

$$\mathcal{L}_n\left(\mathcal{A}_{\ell_1}, D_{\ell_1}\right) - \varphi\left(c_i^{(\ell_1)}, \widetilde{T}_i^*\right) \leq \sqrt{\frac{(K+5)M_{\ell_1}}{n^{3/2}}}\left(\frac{l}{2\lambda_{\min}^{(\ell_1)}}\right) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right). \qquad (21)$$

Finally, to obtain the approximation error in (19), we again apply (Blyth, 1980, Eq. (2.4)) to note that the approximation error, $\max_{i\in[K]}|\varphi(c_i^{(\ell_1)}, \widetilde{T}_i^*)\gamma_i\left(\widetilde{T}_i^*\right)$ is $\mathcal{O}\left(1/(\lambda_{\min}^{\ell_1})^{3/2}\right)$. This concludes the proof that the excess regret for the $\ell_1$ loss function is of the order of $\tilde{\mathcal{O}}\left(n^{-3/4}\right)$.

## F. Regret bound for $f$-divergence

### F.1. General Regret Bounds

To retrace the approach employed in Sections 4.2 and 4.3, and apply the general adaptive algorithm (Alg. 1), we first need to make sure that the approximate objective function $\varphi_i$ lies in $\mathcal{F}$, and then construct the requisite upper-bound. This step involves computations tailored to specific choices of $f$-divergence. Instead, we take an alternative approach and study the regret bound for any adaptive scheme for which we can define a particular high-probability event $\mathcal{E}_\delta$ (introduced below). This framework allows us to obtain a general decomposition of the regret into several components in Thm. 7. We study the particular case of KL-divergence in Sec. 4.4.1, for which we first show that $\varphi_i \in \mathcal{F}$ (see Eq. 9) and then construct the appropriate upper-bound necessary to implement Alg. 1 (see Lemma 10). We show that for this adaptive scheme, we can obtain the required high probability event, and thus, employing the general regret decomposition of Thm. 7, we derive explicit regret bound in Thm. 4. Similar results can be obtained for other commonly used $f$-divergences such as Hellinger distance.

We consider a general **adaptive allocation scheme** that is applied with the approximate objective function $\varphi_i$ and satisfies the condition: **(a1)** for any $\delta > 0$, we can define a $(1-\delta)$-probability event $\mathcal{E}_\delta$ under which $\tau_{0,i} \leq T_i \leq \tau_{1,i}$, for $i \in [K]$. Here $\tau_{0,i}$ and $\tau_{1,i}$ are non-negative constants that depend on $n$ and $\delta$, and $T_i$ is the (random) number of times $\mathcal{A}_f$ pulls arm $i$.

We now prove a lemma that bounds the moments of $W_{ij,T_i} = \sum_{s=1}^{T_i} \tilde{Z}_{ij}^{(s)}$ in the definition of $\varphi_i$ (Eq. 8) for an adaptive allocation scheme $\mathcal{A}_f$ satisfying **(a1)**.

**Lemma 9.** *Let $\mathcal{A}_f$ be an allocation scheme that satisfies **(a1)** and $\mathcal{E}_\delta$ be the corresponding high probability event. Then, for $m \geq 1$ and $i \in [K]$, we have*

$$\mathbb{E}\big[(W_{ij,T_i})^m \mathbb{1}_{\{\mathcal{E}\}}\big] \leq \mathbb{E}\big[(W_{ij,\tau_{0,i}})^m\big] + \beta_m^{(ij)}(\tau_{0,i}, \tau_{1,i}),$$

*where*

$$\beta_m^{(ij)}(\tau_{0,i}, \tau_{1,i}) := \sum_{k=1}^{m} (\tau_{1,i} - \tau_{0,i})^k \binom{m}{k} \mathbb{E}\big[(W_{ij,\tau_{0,i}})^{m-k}\big].$$

Lemma 9, proved in Appendix F.3, provides the bounds on the moments of $W_{ij,T_i}$ that will be used to upper-bound the term $\mathbb{E}[D_f^{(r)}(\hat{P}_i, P_i)]$ in the regret analysis of the adaptive scheme $\mathcal{A}_f$ in Theorem 7, which we now state.

**Theorem 7.** *Suppose $f$ satisfies **(f1)** and **(f2)**, and $\mathcal{A}_f$ satisfies **(a1)**. Then, we have*

$$\mathcal{R}_n(\mathcal{A}_f, D_f) \leq \max_{i \in [K]} \big(\varphi_i(\tau_{0,i}) - \varphi_i(T_i^*) + \Psi_i\big), \tag{22}$$

*where $\Psi_i := \left(\sum_{k=1}^3 \psi_{k,i}\right) + \psi_4$, and*

$$\psi_{1,i} := \big(\delta n^{\frac{r+1}{2}} + (\frac{\sqrt{n}}{\tau_{0,i}})^{r+1}\big) \sum_{j=1}^{l} \frac{C_{f,r+1} e^2 \big(3.2(r+1)\big)^{\frac{r+1}{2}}}{p_{ij}^{r+1}},$$

$$\psi_{2,i} := \sum_{m=1}^{r} \sum_{j=1}^{l} \frac{\beta_m^{(ij)}(\tau_{0,i}, \tau_{1,i})}{m!\ p_{ij}^{m-1}\ \tau_{0,i}^m}, \quad \psi_{3,i} := \sum_{m=1}^{r} \sum_{j=1}^{l} \frac{\delta}{m!\ p_{ij}^{m-1}},$$

$$\psi_4 := \max_{1 \leq i \leq K} \left( \sum_{j=1}^{l} 4 C_{f,r+1} (p_{ij} \widetilde{T}_i^*)^{-\frac{r+1}{2}} \right).$$

*Proof Outline.* From Eq. 13 in Proposition 1 in Appendix B, we have $\mathcal{R}_n(\mathcal{A}_f, D_f) \leq \big(\mathcal{L}_n(\mathcal{A}_f, D_f) - \varphi_i(\widetilde{T}_i^*)\big) + 2\max_{k \in [K]} |\mathbb{E}[R_{k,r+1}(\widetilde{T}_i^*)]|$. The second term is upper-bounded with $\psi_4$ by employing Lemma 4. Next, we decompose $\mathcal{L}_n(\mathcal{A}_f, D_f)$ into three terms: $\mathbb{E}[D_f^{(r)}(\hat{P}_i, P_i)\mathbb{1}_{\{\mathcal{E}_\delta\}}]$, $\mathbb{E}[D_f^{(r)}(\hat{P}_i, P_i)\mathbb{1}_{\{\mathcal{E}_\delta^c\}}]$ and $\mathbb{E}[R_{i,r+1}(T_i)]$ and upper-bound them with $(\varphi_i(\tau_{0,i}) + \psi_{2,i})$, $\psi_{3,i}$ and $\psi_{1,i}$ respectively. The details of these steps are given in Appendix F.4.

**Remark 4.** *The magnitude of the terms $\varphi_i(\tau_{0,i}) - \varphi_i(\widetilde{T}_i^*)$ and $\psi_{2,i}$ in* (22) *depend on how closely $(T_i)_{i=1}^K$ can match the approx-oracle allocation $(\widetilde{T}_i^*)_{i=1}^K$. The term $\psi_4$ is of $\mathcal{O}(n^{-(r+1)/2})$, while $\psi_{1,i} = \mathcal{O}(\delta n^{(r+1)/2} + (\sqrt{n}/\tau_{0,i})^{r+1})$. Finally, the term $\psi_{3,i}$ is of $\mathcal{O}(\delta)$.*

### F.2. Proof of Lemma 4

Note that for any fixed value of $\hat{p}_{ij}$, the remainder term $R_{ij,r+1}$ can be written in two ways as follows:

$$R_{ij,r+1} \overset{(a)}{=} \sum_{m \geq r+1} \frac{f^{(m)}(1)}{m! p_{ij}^{m-1}} (\hat{p}_{ij} - p_{ij})^m \overset{(b)}{=} \frac{f^{(r+1)}(z)}{m! p_{ij}^{r}} (\hat{p}_{ij} - p_{ij})^{r+1}. \tag{23}$$

The equality **(b)** is the Lagrange form of remainder, where the value $z$ is a function of $\hat{p}_{ij}/p_{ij}$.

Introduce the definition $Q_\epsilon \coloneqq \{q \ : \ |q - p_{ij}|/p_{ij} \geq \epsilon\}$ for a fixed $0 < \epsilon < 1$. In order to upper bound the expected value of $R_{ij,r+1}$, we consider two cases, **(1)** first when the random variable $\hat{p}_{ij}$ lies in $Q_\epsilon^c$, and **(2)** second, when $\hat{p}_{ij} \in Q_\epsilon$.

When $\hat{p}_{ij} \notin Q_\epsilon$, we use the equality **(a)** in (23) to note that

$$\begin{aligned} R_{ij,r+1}\mathbb{1}_{\{\hat{p}_{ij} \notin Q_\epsilon\}} &\leq p_{ij}\left(\frac{\hat{p}_{ij} - p_{ij}}{p_{ij}}\right)^{r+1}\left(\max_{m \geq r+1} \frac{f^{(m)}(1)}{m!}\right)\frac{1}{1-\epsilon} \\ \Rightarrow \mathbb{E}\left[R_{ij,r+1}\mathbb{1}_{\{\hat{p}_{ij} \notin Q_\epsilon\}}\right] &\leq \left(\frac{C_1}{p_{ij}^r(1-\epsilon)}\right)\mathbb{E}\left[\left(\hat{p}_{ij} - p_{ij}\right)^{r+1}\right]. \end{aligned} \tag{24}$$

Next we consider the second case. Here we note that for any $q \in Q_\epsilon$, the remainder can be written as (with dependence on $q$ made explicit):

$$\begin{aligned} R_{ij,r+1}(q) &= p_{ij}\left(f\left(\frac{q}{p_{ij}}\right) - \sum_{m=0}^{r} \frac{f^{(m)}(1)}{m!}\left(\frac{q - p_{ij}}{p_{ij}}\right)^m\right) \\ &= \frac{f^{(r+1)}(z_q)}{m!p_{ij}^r}\left(q - p_{ij}\right)^{r+1}. \end{aligned}$$

In the second equality, the term $z_q$ varies with $q$. Now, since the set $Q_\epsilon$ is compact, and by the local boundedness of the function $f$ (assumption **(f1)** in § 4.4), we have $\sup_{q \in Q_\epsilon} R_{ij,r+1} \coloneqq \gamma_\epsilon < \infty$. This implies the following:

$$\sup_{q \in Q_\epsilon} \frac{f^{(r+1)}(z_q)}{m!} \leq \frac{1}{p_{ij}}\left(\frac{\gamma_\epsilon}{\epsilon^{r+1}}\right) \coloneqq \frac{C_\epsilon}{p_{ij}}. \tag{25}$$

Using this inequality, we can proceed as follows:

$$\mathbb{E}\left[R_{ij,r+1}\mathbb{1}_{\{\hat{p}_{ij} \in Q_\epsilon\}}\right] \leq C_\epsilon \mathbb{E}\left[\left(\frac{\hat{p}_{ij} - p_{ij}}{p_{ij}}\right)^{r+1}\right] \tag{26}$$

Combining (24) and (26), we get

$$\mathbb{E}\left[R_{ij,r+1}\right] \leq \left(\inf_{0<\epsilon<1}\left(\frac{C_1 p_{ij}}{(1-\epsilon)} + C_\epsilon\right)\right)\mathbb{E}\left[\left(\frac{\hat{p}_{ij} - p_{ij}}{p_{ij}}\right)^{r+1}\right] \tag{27}$$

$$\overset{(a)}{\leq} \left(\inf_{0<\epsilon<1}\left(\frac{C_1 p_{ij}}{(1-\epsilon)} + C_\epsilon\right)\right)\left(\frac{3e^{2/e}(r+1)}{2}\right)^{r+1}\left(p_{ij}T_i\right)^{-(r+1)/2} \tag{28}$$

$$\coloneqq C_{f,r+1}(p_{ij}T_i)^{-(r+1)/2}. \tag{29}$$

In the inequality **(a)** above, we used the fact that the random variable $\hat{p}_{ij} - p_{ij}$ is subgaussian with parameter $\sigma = \sqrt{3/(2T_i)}$, and then used the upper bound on the $(r+1)^{th}$ moment of a subgaussian random variable (Rivasplata, 2012, Prop. 3.2).

### F.3. Proof of Lemma 9

Since $W_{ij,T_i} = \sum_{s=1}^{T_i} \tilde{Z}_{ij}^{(s)} = \sum_{s=1}^{T_i} Z_{ij}^{(s)} - T_i p_{ij}$, where $p_{ij} = \mathbb{E}\left[Z_{ij}^{(s)}\right]$ for all $s \geq 1$, we can upper bound $W_{ij,T_i}$ by upper bounding the first term and lower bounding the second term in its definition. This is achieved by exploiting the non-negativity of $Z_{ij}^{(s)}$ and the bounds on the random variable $T_i$ under the event $\mathcal{E}_\delta$. The rest of the proof then proceeds by using the binomial expansion of the expression obtained (i.e., the term bounding $W_{ij,T_i}$).

$$\begin{aligned}\mathbb{E}\left[(W_{ij,T_i})^m \mathbb{1}_{\{\mathcal{E}_\delta\}}\right] &= \mathbb{E}\left[\left(\sum_{s=1}^{T_i} Z_{ij}^{(s)} - T_i p_{ij}\right)^m \mathbb{1}_{\{\mathcal{E}_\delta\}}\right] \overset{(a)}{\leq} \mathbb{E}\left[\left(\sum_{s=1}^{\tau_{1,i}} Z_{ij}^{(s)} - \sum_{s=1}^{\tau_{0,i}} p_{ij}\right)^m \mathbb{1}_{\{\mathcal{E}_\delta\}}\right] \\ &\overset{(b)}{\leq} \mathbb{E}\left[\left(\sum_{s=1}^{\tau_{0,i}} \tilde{Z}_{ij}^{(s)} + \sum_{s=\tau_{0,i}+1}^{T_i} 1\right)^m \mathbb{1}_{\{\mathcal{E}_\delta\}}\right] = \mathbb{E}\left[\left(W_{ij,\tau_{0,i}} + (\tau_{1,i} - \tau_{0,i})\right)^m \mathbb{1}_{\{\mathcal{E}_\delta\}}\right] \\ &\leq \mathbb{E}\left[\left(W_{ij,\tau_{0,i}}\right)^m\right] + \sum_{k=1}^{m} (\tau_{1,i} - \tau_{0,i})^k \binom{m}{k} \mathbb{E}\left[\left(W_{ij,\tau_{0,i}}\right)^{m-k}\right].\end{aligned}$$

In the above display,
**(a)** follows from the fact that $Z_{ij}^{(s)} \geq 0$ a.s. and that under the event $\mathcal{E}_\delta$, we have $\tau_{0,i} \leq T_i \leq \tau_{1,i}$. This implies that $\sum_{s=1}^{T_i} Z_{ij}^{(s)} \leq \sum_{s=1}^{\tau_{1,i}} Z_{ij}^{(s)}$ and $T_i p_{ij} \geq \tau_{0,i} p_{ij}$.
**(b)** follows from the fact that $|\tilde{Z}_{ij}^{(s)}| \leq 1$ a.s.

### F.4. Proof of Theorem 7

Note that from (13) in Lemma 1 in Appendix B, we have $\mathcal{R}_n(\mathcal{A}_f, D_f) \leq \mathcal{L}_n(\mathcal{A}_f, D_f) - \varphi_i(\widetilde{T}_i^*) + 2\max_{k\in[K]} |R_{k,r+1}(\widetilde{T}_i^*)|$.

The risk term can be further decomposed as follows:

$$\begin{aligned}\mathcal{L}_n\left(\mathcal{A}_f, D_f\right) &= \mathbb{E}\left[D_f^{(r)}(\hat{P}_i, P_i) + R_{i,r+1}(T_i)\right] \\ &\leq \mathbb{E}\left[D_f^{(r)}\left(\hat{P}_i, P_i\right)\left(\mathbb{1}_{\{\mathcal{E}_\delta\}} + \mathbb{1}_{\{\mathcal{E}_\delta^c\}}\right)\right] + \mathbb{E}\left[R_{i,r+1}(T_i)\right].\end{aligned}$$

We first obtain an upper bound on the term term $\mathcal{L}_n\left(\mathcal{A}_f, D_f\right)$. Suppose $\hat{P}_i$ is the empirical estimate of $P_i$ constructed from the samples collected through the adaptive scheme $\mathcal{A}_{(f)}$ after $n$ rounds. Then we have the following:

$$\begin{aligned}\mathcal{L}_n\left(\mathcal{A}_f, D_f\right) &= \mathbb{E}\left[D_f\left(\hat{P}_i, P_i\right)\right] = \mathbb{E}\left[D_f^{(r)}(\hat{P}_i, P_i) + R_{i,r+1}(T_i)\right] \qquad (30) \\ &= \mathbb{E}\left[\sum_{m=1}^{r}\sum_{j=1}^{l} \frac{f^{(m)}(1)}{m!p_{ij}^{m-1}} \left(\hat{p}_{ij} - p_{ij}\right)^m + \sum_{m\geq r+1}\sum_{j=1}^{l} \frac{f^{(m)}(1)}{m!p_{ij}^{m-1}} \left(\hat{p}_{ij} - p_{ij}\right)^m\right] \\ &:= \textbf{term1} + \textbf{term2} \qquad (31)\end{aligned}$$

We consider the two terms above separately.

$$\begin{aligned}\textbf{term1} &= \mathbb{E}\left[D_f^{(r)}\left(\hat{P}_i, P_i\right)\left(\mathbb{1}_{\{\mathcal{E}_\delta\}} + \mathbb{1}_{\{\mathcal{E}_\delta^c\}}\right)\right] = \mathbb{E}\left[\sum_{m=1}^{r}\sum_{j=1}^{l}\left(\frac{f^{(m)}(1)}{m!p_{ij}^{m-1}} \frac{\left(\sum_{s=1}^{T_{i,n}} \tilde{Z}_{ij}^{(s)}\right)^m}{T_{i,n}^m}\right)\left(\mathbb{1}_{\{\mathcal{E}_\delta\}} + \mathbb{1}_{\{\mathcal{E}_\delta^c\}}\right)\right] \\ &\overset{(a)}{\leq} \sum_{m=1}^{r}\sum_{j=1}^{l} \frac{f^{(m)}(1)}{m!p_{ij}^{m-1}}\left(\frac{1}{\tau_{0,i}^m}\mathbb{E}\left[\left(\sum_{s=1}^{\tau_{0,i}} \tilde{Z}_{ij}^{(s)}\right)^m\right]\right) + \sum_{m=1}^{r}\sum_{j=1}^{l}\left(\frac{f^{(m)}(1)}{\tau_{0,i}^m m!p_{ij}^{m-1}}\beta_m^{(ij)}\left(\tau_{0,i}, \tau_{1,i}\right) + \frac{f^{(m)}(1)\delta}{m!p_{ij}^{m-1}}\right) \\ &= \varphi_i\left(\tau_{0,i}\right) + \psi_{2,i} + \psi_{3,i}.\end{aligned}$$

The inequality **(a)** in the above display uses the following facts:
(i) Under the event $\mathcal{E}_\delta$, the term $T_i$ is lower bounded by $\tau_{0,i}$ and thus $1/T_i^m$ is upper bounded by $1/\tau_{0,i}^m$
(ii) We use Lemma 9 to upper bound the moment of $\mathbb{E}\left[\left(\sum_{s=1}^{\tau_{0,i}} \tilde{Z}_{ij}^{(s)}\right)^m\right]$, which gives us the additional term consisting of $\beta_m^{(ij)}\left(\tau_{0,i}, \tau_{1,i}\right)$.
(iii) Under the event $\mathcal{E}_\delta^c$, we can upper bound the required moment by its worst case value of 1 multiplied by the bound on the probability of $\mathcal{E}_\delta^c$, i.e., $\delta$.

Next, we need to get the upper bound on the second term in (31). For any $t \geq 1$, we recall the notation $W_{ij,T_i} = \sum_{s=1}^{T_i} \tilde{Z}_{ij}^{(s)}$.

Since **term2** is the expectation of the remainder of the $r$-term approximation of $D_f\left(\hat{P}_i, P_i\right)$ with $P_i$ as constructed by the adaptive scheme, we can write the following:

$$\begin{aligned}\text{term2} &\overset{(a)}{\leq} \mathbb{E}\left[C_{f,r+1}\left(\frac{W_{ij,T_i}}{p_{ij}T_i}\right)^{r+1}\right] = C_{f,r+1}\left(\mathbb{E}\left[\left(\frac{W_{ij,T_i}}{p_{ij}T_i}\right)^{r+1}\left(\mathbb{1}_{\{\mathcal{E}_\delta\}} + \mathbb{1}_{\{\mathcal{E}_\delta^c\}}\right)\right]\right) \\ &\overset{(b)}{\leq} \frac{C_{f,r+1}}{(p_{ij})^{r+1}}\mathbb{E}\left[(W_{ij,T_i})^{r+1}\right]\left(\delta + \frac{1-\delta}{\tau_{0,i}^{r+1}}\right)\end{aligned}$$

In the above display,
**(a)** employs the upper bound on the remainder term derived in (27) in the proof of Lemma 9,
**(b)** uses the fact that $T_i \geq \tau_{0,i}\mathbb{1}_{\{\mathcal{E}_\delta\}} + \mathbb{1}_{\{\mathcal{E}_\delta^c\}}$.

It remains to obtain an appropriate upper bound on the term $\mathbb{E}\left[(W_{ij,T_i})^{r+1}\right]$. We first observe the following:

$$W_{ij,T_i} \leq \max_{1\leq s\leq n} W_{ij,s} \ := \mathcal{W}_{ij,n},$$

where the inequality holds in an almost sure sense, and it follows from the fact that by definition $T_i \leq n$ almost surely for all $1 \leq i \leq K$. Next, using the fact that $(W_{ij,s})_{s\geq 1}$ is a martingale sequence, we obtain by an application of Doob's $L_p$ maximal inequality (Durrett, 2019, Theorem 4.4.6), the following inequality:

$$\begin{aligned}\mathbb{E}\left[(W_{ij,T_i})^{r+1}\right] &\leq \mathbb{E}\left[(\mathcal{W}_{ij,n})^{r+1}\right] \leq \left(\frac{r+1}{r+1-1}\right)^{r+1}\mathbb{E}\left[|W_{ij,n}|^{r+1}\right] \\ &\overset{(a)}{\leq} \left(\frac{r+1}{r+1-1}\right)^{r+1}\left(\sqrt{\frac{3n}{2}}e^{1/e}\sqrt{r+1}\right)^{r+1} \\ &\leq e^2\left(3.2(r+1)\right)^{(r+1)/2} n^{(r+1)/2} := C_{(r)}n^{(r+1)/2}.\end{aligned}$$

The inequality **(a)** in the above display follows from the observation that $W_{ij,n}$ is a subgaussian random variable with $\sigma = \sqrt{3n/2}$, and that the $r+1$ moment of a $\sigma$-subgaussian random variable is upper bounded by $\left(\sigma e^{1/e}\sqrt{r+1}\right)^{r+1}$ (Rivasplata, 2012, Prop. 3.2). Thus we finally obtain

$$\text{term2} \leq \frac{C_{f,r+1}}{(p_{ij})^{r+1}}\left(\delta + \frac{1-\delta}{\tau_{0,i}^{r+1}}\right)C_{(r)}n^{(r+1)/2} = \frac{C_{f,r+1}C_{(r)}}{(p_{ij})^{r+1}}\left(\delta n^{(r+1)/2} + \left(\frac{\sqrt{n}}{\tau_{0,i}}\right)^{r+1}\right) := \psi_{1,i}.$$

To summarize, we have shown that the risk $\mathcal{L}_n\left(\mathcal{A}_f, D_f\right)$, i.e., the expected $f$-divergence between the empirical estimate $\hat{P}_i$ constructed using the adaptive sampling scheme $\mathcal{A}$ with a budget of $n$ samples can be upper bounded as

$$\mathcal{L}_n\left(\mathcal{A}_f, D_f\right) \leq \max_{1\leq i\leq K}\left(\varphi_i\left(\tau_{0,i}\right) + \psi_{1,i} + \psi_{2,i} + \psi_{3,i}\right) \tag{32}$$

We can now apply the regret decomposition bound given in (13) in Proposition 1 in Appendix B to get the following:

$$\begin{aligned}\mathcal{R}_n\left(\mathcal{A}_f, D_f\right) &\leq \mathcal{L}_n\left(\mathcal{A}_f, D_f\right) - \varphi_i\left(\widetilde{T}_i^*\right) + 2\max_{1\leq k\leq K}|R_{i,r+1}(\widetilde{T}_i^*)| \\ &\leq \mathcal{L}_n\left(\mathcal{A}_f, D_f\right) - \varphi_i\left(\widetilde{T}_i^*\right) + \psi_4 \\ &\leq \max_{1\leq i\leq K}\left(\varphi_i(\tau_{0,i}) - \varphi_i\left(\widetilde{T}_i^*\right) + \Psi_i\right).\end{aligned}$$

In the last inequality which follows from an application of (32), we used the fact that $\varphi_i\left(\widetilde{T}_i^*\right)$ and $\psi_4$ do not depend on $i$. This completes the proof.

## G. Analysis for KL divergence

### G.1. Proof of Lemma 10

Having obtained the approximate objective function $\varphi(c_i^{(\mathrm{KL})}, T_i)$ for the tracking problem (4), we now need to construct high probability confidence intervals for the parameters $c_i^{(\mathrm{KL})}$. We present the required concentration result under the assumption that the distributions lie in the $\eta$-interior of the $(l-1)$-dimensional simplex, $\Delta_l^{(\eta)}$, defined in Section 2.

**Lemma 10.** *For a given $\delta \in (0,1)$, recall $e_{ij,t} = \sqrt{\log(2/\delta_t)/2T_{i,t}}$ introduced in Lemma 2. Assume that $\hat{p}_{ij,t} \geq 7e_{ij,t}/2$. Then, the following inequalities hold for $P_i \in \Delta_l^{(\eta)}$:*

$$\frac{1}{p_{ij}} \leq \frac{1}{\hat{p}_{ij,t} - e_{ij,t}} \leq \frac{1}{p_{ij}} + \frac{4e_{ij,t}}{\eta\, p_{ij}}\ , \tag{33}$$

*which implies that*

$$\sum_{j=1}^{l} \frac{1}{p_{ij}} \leq \sum_{j=1}^{l} \frac{1}{\hat{p}_{ij,t} - e_{ij,t}} \leq \sum_{j=1}^{l} \frac{1}{p_{ij}} + \frac{l}{\eta^2}\sqrt{\frac{8l^2 \log(2/\delta_t)}{T_{i,t}}}.$$

Lemma 10 allows us to define high probability upper bounds on the parameters $c_i^{(\mathrm{KL})}$ as $u_{i,t}^{(\mathrm{KL})} := \big(\sum_{j=1}^{l} \frac{1}{\hat{p}_{ij,t} - e_{ij,t}} - 1\big)/12$, if $\hat{p}_{ij,t} \geq 7e_{ij,t}/2$, and $u_{i,t}^{(\mathrm{KL})} = +\infty$, otherwise. The first inequality of (33) follows directly from the definition of the event $\mathcal{E}_3$, under which $p_{ij} \geq \hat{p}_{ij,t} - e_{ij,t}$.

We next claim that under the conditions of the lemma, we must have $\hat{p}_{ij,t} - e_{ij,t} > \eta/2$. We prove this statement by contradiction. Assume that $\eta/2 \geq \hat{p}_{ij,t} - e_{ij,t} \geq 7/2e_{ij,t} - e_{ij,t} = 5/2e_{ij,t}$, or equivalently $5e_{ij,t} \leq \eta$. Since $\hat{p}_{ij,t} + e_{ij,t} \geq p_{ij}$ under the event $\mathcal{E}_1$, we also have the following chain of inequalities:

$$\eta < p_{ij} \leq \hat{p}_{ij,t} + e_{ij,t} = \hat{p}_{ij,t} - e_{ij,t} + 2e_{ij,t} \leq \eta/2 + 2e_{ij,t}$$

which implies that $e_{ij,t} > \eta/4$ or equivalently $4e_{ij,t} > \eta$. This gives us the required contradiction.

We now invoke the convexity of the mapping $x \mapsto 1/x$ to claim the following:

$$\begin{aligned}\frac{1}{\hat{p}_{ij,t} - e_{ij,t}} &\leq \frac{1}{p_{ij}} + \left(\frac{2/\eta - 1/p_{ij}}{\eta/2 - p_{ij}}\right)(\hat{p}_{ij,t} - e_{ij,t} - p_{ij}) \\ &= \frac{1}{p_{ij}} + \frac{2\left(p_{ij} - \hat{p}_{ij,t} + e_{ij,t}\right)}{\eta p_{ij}} \leq \frac{1}{p_{ij}} + \frac{4e_{ij,t}}{\eta p_{ij}} \leq \frac{1}{p_{ij}} + \sqrt{\frac{32\log(2/\delta_t)}{T_{i,t}\eta^4}}.\end{aligned}$$

The result in (10) follows by taking the summation over $1 \leq j \leq l$.

### G.2. Deviation of $\widetilde{T}_i^*$ from the uniform allocation

Before proceeding to the proof of Theorem 4, we first present a result that shows that the optimal static allocation for the approximate tracking objective does not deviate much from the uniform allocation (i.e., the first order approximation).

**Lemma 11.** *Suppose $P_i \in \Delta_l^{(\eta)}$ for $1 \leq i \leq K$. Define $T_0 = n/K$, and let $\widetilde{T}_i^*$ for $1 \leq i \leq K$ be the optimal static allocation according to the first two terms of (33). For $1 \leq i \leq K$, define $c_i^{(KL)} := \frac{1}{12}\left(\sum_{j=1}^{l} \frac{1}{p_{ij}} - 1\right)$ and let $c_{\max}^{(KL)}$ and $c_{\min}^{(KL)}$ denote the maximum and minimum values of $c_i^{(KL)}$ as $i$ varies from $1$ to $K$. Then we have the following:*

$$\left|\widetilde{T}_i^* - T_0\right| \leq K\frac{c_{\max}^{(KL)} - c_{\min}^{(KL)}}{l-1} \quad \forall\, 1 \leq i \leq K.$$

*Proof.* We will use the notation $\varphi_i(T)$ to denote $\varphi(c_i, T)$ in this proof.

We first assume that $\tilde{T}_i^* < T_0$. The following two statements follow from the fact that $\varphi_i$ is decreasing.

- $\varphi_i(\tilde{T}_i^*) - \varphi_i(T_0) \geq 0$. This is because $\varphi_i$ is decreasing in $T$.
- $g_i(T_0)(\tilde{T}_i^* - T_0) \geq 0$. This is because $g_i(T_0) \leq 0$ as $\varphi_i$ is decreasing.

Next, by convexity of $\varphi_i$ we have

$$\begin{aligned} &\varphi_i(T_0) + g_i(T_0)(\tilde{T}_i^* - T_0) \le \varphi_i(\tilde{T}_i^*) \Rightarrow \; g_i(T_0)(\tilde{T}_i^* - T_0) \le \varphi_i(\tilde{T}_i^*) - \varphi_i(T_0) \\ &\quad \Rightarrow |g_i(T_0)(\widetilde{T}_i^* - T_0)| \le |\varphi_i(\widetilde{T}_i^*) - \varphi_i(T_0)| \end{aligned} \tag{34}$$

Thus, if we define $U = \{i : \widetilde{T}_i^* < T_0\}$, then we have $\max_{i\in U} |\widetilde{T}_i^* - T_0| \le \max_{i \in U} \frac{|\varphi_i(T_0) - \varphi_i(\widetilde{T}_i^*)|}{|g_i(T_0)|} \overset{(a)}{\le} \frac{c_{\max} - c_{\min}}{l-1} := Q$, where **(a)** is shown in the proof of Lemma 11.

Next, define $V = \{i : \widetilde{T}_i^* \ge T_0\} = [K] \setminus U$. Then we know that $\sum_{i=1}^K (\widetilde{T}_i^* - T_0) = 0$, which means that

$$\max_{i \in V} (\widetilde{T}_i^* - T_0) \le \sum_{i \in V} (\widetilde{T}_i^* - T_0) = \sum_{j \in U} (T_0 - \widetilde{T}_i^*) \le |U| Q \le KQ.$$

Thus we have the following:

$$\max_{i \in U} |\widetilde{T}_i^* - T_0| \le Q, \quad \text{and} \quad \max_{i \in V} |\widetilde{T}_i^* - T_0| \le KQ.$$

which implies the required result. Note, in the above we have assumed that the set $U$ is non-empty. In case $U$ is an empty set, then we must have $\widetilde{T}_i^* = T_0$ for all $i \in [K]$ and the result follows trivially. □

**Remark 5.** *We can show that by using the uniform allocation, the regret incurred in learning distribution in terms of KL-divergence is $\mathcal{O}\left(n^{-2}\right)$. This regret is asymptotically slower than the $\mathcal{O}\left(n^{-5/2}\right)$ rate presented in Theorem 4. However, under some parameter regimes, this rate may actually be faster than the one presented in Theorem 4 due to the larger hidden constants in front of the $\mathcal{O}\left(n^{-5/2}\right)$ regret bound of Theorem 4.*

### G.3. Proof of Theorem 4

We assume that $n$ is large enough to ensure that $\widetilde{T}_i^* \ge n/(2K)$ for $1 \le i \le K$. A sufficient condition for this is $n \ge \frac{K(l-\eta)}{6(l-1)}$.

**Value of Constants $A$, $B$ and $\tilde{e}_n$.** We have the following:

$$\begin{aligned} A &= \max_{1 \le i \le K} \left. \frac{\partial \varphi(c, \widetilde{T}_i^*)}{\partial c} \right|_{c = c_i^{(KL)}} = \max_{1 \le i \le K} \frac{1}{\left(\widetilde{T}_i^*\right)^2} \le \frac{4K^2}{n^2}. \\ B &= \max_{1 \le i \le K} \left. \frac{\partial \varphi(c_i^{(KL)}, T)}{\partial T} \right|_{T = \widetilde{T}_i^*} = \min_{1 \le i \le K} \frac{l-1}{\left(2\widetilde{T}_i^*\right)^2} + \frac{2c_i^{(KL)}}{\left(\widetilde{T}_i^*\right)^3} \ge \frac{(l-1)K^2}{n^2}. \\ \tilde{e}_n &= \max_{1 \le i \le K} \frac{1}{\eta^3} \sqrt{\frac{12 \log(\delta_n)}{\widetilde{T}_i^*}} \le \frac{1}{\eta^3} \sqrt{\frac{24K \log(\delta_n)}{n}}. \end{aligned}$$

Thus we have

$$\frac{A\tilde{e}}{B} \le \frac{4\sqrt{24K\log(\delta_n)}}{(l-1)\eta^3\sqrt{n}} := M^{(KL)} \frac{1}{\sqrt{n}}. \tag{35}$$

We next proceed to analyze the regret of the adaptive scheme when compared to the oracle static allocation scheme. As suggested by Theorem 7, we separately obtain upper bounds on the three terms $\psi_{1,i}$, $\psi_4$ and $\mathbb{E}\left[D_f^{(r)}\left(\hat{P}_i, P_i\right)\right]$.

#### G.3.1. Bound on $\boldsymbol{\psi_{1,i}}$.

We proceed as follows:

$$\psi_{1,i} = \sum_{j=1}^{l} \frac{C_{KL,r+1} C_{(r)}}{p_{ij}^{r+1}} \left( \delta n^{(r+1)/2} + \left(\frac{3K}{\sqrt{n}}\right)^{r+1} \right) \le \frac{lC_{KL,r+1}, C_{(r)}}{\eta^{r+1}} \left( \delta n^{(r+1)/2} + (3K)^{(r+1)} n^{-(r+1)/2} \right).$$

In the above display, $C_{KL,r+1}$ is the instance of the constant $C_{f,r+1}$ for the case of KL-divergence. Since $r=5$, we see that an appropriate choice of $\delta$ above is $\delta=(3K/n)^6$, which gives us

$$\psi_{1,i}=\frac{2lC_{KL,6}C_{(5)}(3K)^6}{\eta^6}n^{-3}=\mathcal{O}\left(n^{-3}\right). \tag{36}$$

### G.3.2. BOUND ON $\psi_4$.

We proceed as follows:

$$\psi_4=\max_{1\le i\le K}\left(\sum_{j=1}^{l}C_{KL,r+1}\left(\widetilde{T}_i^*p_{ij}\right)^{-(r+1)/2}\right)\overset{(a)}{\le}\sum_{j=1}^{l}C_{KL,r+1}\left(\eta n/(2K)\right)^{-(r+1)/2}.$$

In the above display, we used the fact that since $P_i\in\Delta_l^{(\eta)}$ we must have $p_{ij}\ge\eta$ and that $n$ is large enough to ensure that $\widetilde{T}_i^*\ge n/2K$. For the value of $r=5$ in the Taylor's expansion of the KL-divergence, we get the following bound on $\psi_4$

$$\psi_4\le lC_{KL,6}\left(\frac{2K}{\eta n}\right)^3. \tag{37}$$

### G.3.3. BOUND ON $\psi_{3,i}$

We have the following:

$$\psi_{3,i}=\delta\sum_{m=1}^{r}\sum_{j=1}^{l}\frac{f^{(m)}(1)}{m!p_{ij}^{m-1}}\le\frac{\delta l5}{\eta^4}=\frac{5l(3K)^6}{\eta^4}n^{-6}. \tag{38}$$

The first inequality in the above display uses the following facts: since $f(x)=x\log x$ for KL-divergence, we have $f^{(m)}(1)/m!\le 1$, and that since $P_i\in\Delta_l^{(\eta)}$ we have $p_{ij}\ge\eta$. Finally the second inequality uses the value of $\delta=(3K)^6n^{-6}$.

### G.3.4. BOUND ON $\varphi_r\left(\tau_{0,i}\right)-\varphi_r\left(\widetilde{T}_i^*\right)$.

Next we analyze the contribution of the term: $\varphi_r\left(\tau_{0,i}\right)-\varphi_r\left(\widetilde{T}_i^*\right)$ to the regret. We proceed as follows:

$$\varphi_r\left(\tau_{0,i}\right)-\varphi_r\left(\widetilde{T}_i^*\right):=\frac{l-1}{2\tau_{0,i}}-\frac{l-1}{2\widetilde{T}_i^*}+\frac{c_i^{(KL)}}{\left(\tau_{i,n}\right)^2}-\frac{c_i^{(KL)}}{\left(\widetilde{T}_i^*\right)^2}$$

Consider the first term $(l-1)/(2\tau_{0,i})$.

$$\begin{aligned}\frac{l-1}{2\tau_{0,i}}&=\frac{l-1}{2\widetilde{T}_i^*\left(1-\frac{\widetilde{T}_i^*-\tau_{0,i}}{\widetilde{T}_i^*}\right)}\overset{(a)}{\le}\frac{l-1}{2\widetilde{T}_i^*}\left(1+2\frac{\widetilde{T}_i^*-\tau_{0,i}}{\widetilde{T}_i^*}\right)\\&\le\frac{l-1}{2\widetilde{T}_i^*}+\frac{(l-1)(\tau_{1,i}-\tau_{0,i})}{\left(\widetilde{T}_i^*\right)^2}\overset{(b)}{\le}\frac{l-1}{2\widetilde{T}_i^*}+\frac{4K^2(l-1)\left(c_{\max}^{(KL)}-c_{\min}^{(KL)}\right)}{n^{5/2}}.\end{aligned} \tag{39}$$

In the above display,
**(a)** follows from the convexity of the mapping $x\mapsto 1/x$, and the graph of this mapping lies below the chord connecting $(1,1)$ and $(1/2,2)$.
**(b)** follows from the assumption that $n$ is large enough to ensure that $\widetilde{T}_i^*\ge n/(2K)$ for all $1\le i\le K$.

Proceeding in a similar way, we can obtain the following bound on the second term $\frac{c_i^{(KL)}}{\tau_{0,i}^2}$:

$$\frac{c_i^{(KL)}}{\tau_{0,i}^2}\le\frac{c_i^{(KL)}}{\left(\widetilde{T}_i^*\right)^2}+\frac{32K^3c_i^{(KL)}\left(c_{\max}^{(KL)}-c_{\min}^{(KL)}\right)}{n^{7/2}}. \tag{40}$$

Together, (39) and (40) imply that

$$\begin{aligned}\varphi_r\left(\tau_{0,i}\right) - \varphi_r\left(\widetilde{T}_i^*\right) &\leq \frac{4K^2(l-1)\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{n^{5/2}} + \frac{32K^3 c_i^{(KL)}\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{n^{7/2}} \\ &\leq \frac{8K^2(l-1)\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{n^{5/2}}, \end{aligned} \tag{41}$$

where the second inequality follows from the assumption that $n \geq 8Kc_{\max}^{(KL)}/(l-1)$.

### G.3.5. Bound on the term $\boldsymbol{\psi_{2,i}}$.

Finally, we obtain the required bounds on the term $\psi_{2,i} := \sum_{m=1}^{r}\sum_{j=1}^{l}\beta_m^{(ij)}/\left(\tau_{0,i}^m m! p_{ij}^{m-1}\right)$, where $\beta_m^{(ij)}\left(\tau_{0,i},\tau_{1,i}\right)$ was introduced in Lemma 9. We consider several cases:

$\boldsymbol{m = 1.}$ This is the simplest case.

$$\sum_{j=1}^{l}\frac{\beta_1^{(ij)}}{\tau_{0,i}} = \sum_{j=1}^{l}\frac{1}{\tau_{0,i}}\left(\left(\tau_{1,i} - \tau_{0,i}\right) \times 1 \times \mathbb{E}\left[W_{\tau_{0,i}}\right]\right) = 0,$$

which follows from the fact that $\mathbb{E}\left[W_{\tau_{0,i}}\right] = 0$.

$\boldsymbol{m = 2.}$

$$\begin{aligned}\sum_{j=1}^{l}\frac{1}{\tau_{0,i}^2 2! p_{ij}}\beta_m^{(ij)} &= \sum_{j=1}^{l}\frac{1}{\tau_{0,i}^2 2! p_{ij}}\left(\left(\tau_{1,i} - \tau_{0,i}\right)\binom{2}{1}\mathbb{E}\left[W_{\tau_{0,i}}\right] + \left(\tau_{1,i} - \tau_{0,i}\right)^2\binom{2}{2} \times 1\right) \\ &= \sum_{j=1}^{l}\frac{\left(\tau_{1,i} - \tau_{0,i}\right)^2}{2\tau_{0,i}^2 p_{ij}} \leq \frac{\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)^2 9K^2 l}{n^3\eta} \end{aligned} \tag{42}$$

assumed that $n$ is large enough to ensure that $\tau_{0,i} \geq n/(3K)$ which is true if $n$ is large enough to ensure that $n \geq \left(\frac{24K\sqrt{24K\log(1/\delta_n)}}{(l-1)\eta^3}\right)^{2/3}$.

.

$\boldsymbol{m \geq 3.}$ Finally we consider the case where $m \geq 3$.

$$\begin{aligned}\sum_{j=1}^{l}\frac{1}{\tau_{0,i}^m m! p_{ij}^{m-1}}\beta_m^{(ij)} &= \sum_{j=1}^{l}\frac{1}{\tau_{0,i}^m m! p_{ij}^{m-1}}\left(\sum_{k=1}^{m}\left(\tau_{1,i} - \tau_{0,i}\right)^k\binom{m}{k}\mathbb{E}\left[W_{\tau_{0,i}}^{m-k}\right]\right) \\ &\overset{(a)}{=} \sum_{j=1}^{l}\frac{1}{\tau_{0,i}^m m! p_{ij}^{m-1}}\left(\sum_{k=1}^{m-2}\left(\tau_{1,i} - \tau_{0,i}\right)^k\binom{m}{k}\mathbb{E}\left[W_{\tau_{0,i}}^{m-k}\right] + \left(\tau_{1,i} - \tau_{0,i}\right)^m \times 1\right) \\ &\leq \sum_{j=1}^{l} p_{ij}\frac{\tau_{1,i} - \tau_{0,i}}{\tau_{0,i}^m m! \eta^m}\left((m-2)m!\left(\sqrt{m-2}e^{1/e}\sqrt{3\tau_{0,i}/2}\right)^{m-2}\right) \\ &\qquad + \frac{\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)^m (2K)^m l}{\eta^m n^{7/2}} \\ &= \frac{a_m\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{\eta^m n^{-5/2}} + \frac{\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)^m (2K)^m l}{\eta^m n^{7/2}}. \end{aligned} \tag{43}$$

where $a_m := (m-2)^{m/2}(3K)^m e^{(m-2)/e}(3/2)^{m/2-1}$ is implicitly defined in the last equality.

Thus we can obtain the following bound on the term $\psi_{2,i}$ using (42) and (43).

$$\begin{aligned}
\psi_{2,i} &\leq \frac{\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)^2 9K^2 l}{n^3 \eta} + (r-3)\left(\frac{a_r\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{\eta^r n^{-5/2}} + \frac{\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)^r (2K)^r l}{\eta^r n^{7/2}}\right) \\
&= \frac{(r-3)a_r\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{\eta^r n^{-5/2}} + \tilde{\mathcal{O}}\left(n^{-3}\right). \qquad (44)
\end{aligned}$$

The final bound on the regret for the adaptive sampling scheme with loss $D_{KL}$ is obtained by summing up the contributions outlined in equations (36), (37), (38), (42) and (44), we get that

$$\mathcal{R}_n\left(\mathcal{A}_{KL}, D_{KL}\right) \leq \frac{8K^2(l-1)\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{n^{5/2}} + \frac{(r-3)a_5\left(c_{\max}^{(KL)} - c_{\min}^{(KL)}\right)}{\eta^5 n^{5/2}} + \tilde{\mathcal{O}}\left(n^{-3}\right). \qquad (45)$$

This result holds if $n$ is large enough to ensure all of the following:

$$\begin{aligned}
& n \geq \frac{K(l-\eta)}{6(l-1)} \\
& n \geq 8Kc_{\max}^{(KL)}/(l-1) \quad \text{and} \\
& n \geq \left(\frac{24K\sqrt{24K\log(1/\delta_n)}}{(l-1)\eta^3}\right)^{2/3}
\end{aligned} \qquad (46)$$

### G.4. Derivations for $D_{\chi^2}$

The $\chi^2$-distance between two discrete distributions $P$ and $Q$ is defined as $D_{\chi^2}(P,Q) = \sum_{j=1}^{l}\frac{(p_j-q_j)^2}{q_j}$. If $\hat{P}_i$ denotes the empirical estimator of a distribution $P_i$ using $T_i$ i.i.d. samples, then we can compute the expected value of $D_{\chi^2}$ between $\hat{P}_i$ and $P_i$ as follows:

$$\begin{aligned}
\mathbb{E}\left[D_{\chi^2}\left(\hat{P}_i, P_i\right)\right] &= \sum_{j=1}^{l}\mathbb{E}\left[\frac{\left(\hat{p}_{ij} - p_{ij}\right)^2}{p_{ij}}\right] = \sum_{j=1}^{l}\mathbb{E}\left[\frac{\hat{p}_{ij}^2 + p_{ij}^2 - 2p_{ij}\hat{p}_{ij}}{p_{ij}}\right] \\
&= \sum_{j=1}^{l}\left(\mathbb{E}\left[\frac{\hat{p}_{ij}^2}{p_{ij}}\right] - p_{ij}\right) = \sum_{j=1}^{l}\left(\frac{p_{ij}(1-p_{ij})}{T_i p_{ij}} - p_{ij}\right) \\
&= \frac{l-1}{T_i} - 1.
\end{aligned}$$

Thus the tracking objective function is given by $\varphi(c_i, T_i) = \frac{l-1}{2T_i}$. Note that the objective function can be computed exactly and lies in the function class $\mathcal{F}$. Hence we have $T_i^* = \tilde{T}_i^*$. Furthermore, since the objective function only depends on the support size of the distributions, and we assume all $P_i \in \Delta_l$ for the same $l$, the optimal allocation is the uniform allocation.

## H. Regret bound for Separation Distance

### H.1. Proof of Lemma 6

*Proof.* Recall that $Z_{ij}^{(t)}$ is the indicator that the output of the $s^{th}$ pull of arm $i$ was $x_j$. and define $V_j := \frac{\sum_{s=1}^{T_i} p_{ij} - Z_{ij}^{(s)}}{\sqrt{T_i} p_{ij}}$. Next, let $\rho : \mathbb{R}^l \mapsto \{1, 2, \ldots, l\}$ represent the $\arg\max$ operation which breaks ties by returning the smallest index, i.e., for $\vec{v} = (v_1, v_2, \ldots, v_l) \in \mathbb{R}^l$, $\rho\left(\vec{v}\right) = \min\{j : v_j = \max_{1 \leq k \leq l} v_k\}$. With this definition of $\rho$, and with $\vec{V} := (V_1, V_2, \ldots, V_l)$, we introduce the events $\Omega_j := \{\rho(\vec{V}) = j\}$ for $j = 1, 2, \ldots, l$.

We can now proceed as follows:

$$
\begin{aligned}
\mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] &\coloneqq \mathbb{E}\left[\max_{1\le j\le l}\left(1-\frac{\hat{p}_{ij}}{p_{ij}}\right)\right] = \mathbb{E}\left[\max_{1\le j\le l}\sum_{t=1}^{T_i}\frac{p_{ij}-Z_{ij}^{(t)}}{\sqrt{p_{ij}(1-p_{ij})T_i}}\sqrt{\frac{1-p_{ij}}{p_{ij}T_i}}\right] \\
&= \frac{1}{\sqrt{T_i}}\mathbb{E}\left[\max_{1\le j\le l} V_j\left(\sum_{k=1}^{l}\mathbb{1}_{\Omega_k}\right)\right] \overset{(a)}{=} \frac{1}{\sqrt{T_i}}\mathbb{E}\left[\sum_{j=1}^{l} V_j\mathbb{1}_{\Omega_j}\right] \\
&\overset{(b)}{\le} \frac{1}{\sqrt{T_i}}\sum_{j=1}^{l}\mathbb{E}\left[V_j\mathbb{1}_{\{V_j\ge 0\}}\right] \overset{(c)}{\le} \sqrt{\frac{1}{2\pi T_i}}\left(\sum_{j=1}^{l}\sqrt{\frac{1-p_{ij}}{p_{ij}}}\right) + C_i^{(s)}\frac{1}{T_i}.
\end{aligned}
$$

In the above display, **(a)** follows from the fact that on the event $\Omega_j$, $\max_{1\le k\le l} V_k = V_j$, **(b)** follows from the observation that $\Omega_j \subset \{V_j \ge 0\}$, and **(c)** follows from an application of Theorem 3.2 of (Ross, 2011) and the term $C_i^{(s)}$ is given by

$$
C_i^{(s)} = \sum_{j=1}^{l}\left((1-p_{ij})(1+2p_{ij}^2-2p_{ij}) + \left(2(1-p_{ij})\left((1-p_{ij})^3+p_{ij}^3\right)/\pi\right)^{1/2}\right). \tag{47}
$$

Next, in order to obtain the lower bound we need some additional notation. For any subset $S$ of $[l]$, such that $|S| < l$, we introduce the terms

$$
\mathfrak{p}_{i,S} = \sum_{j\in S} p_{ij}, \quad \text{and} \quad \hat{\mathfrak{p}}_{i,S} = \sum_{j\in S}\hat{p}_{ij}. \tag{48}
$$

Define the event that $\Omega_0 \coloneqq \{\hat{\mathfrak{p}}_i \ge \mathfrak{p}_i\}$. Then we have the following:

$$
\begin{aligned}
\mathbb{E}\left[\max_{1\le j\le l}\left(1-\frac{\hat{p}_{ij}}{p_{ij}}\right)\right] &\overset{(a)}{\ge} \mathbb{E}\left[\max\left(1-\frac{\hat{\mathfrak{p}}_{i,S}}{\mathfrak{p}_{i,S}},\ 1-\frac{1-\hat{\mathfrak{p}}_{i,S}}{1-\mathfrak{p}_{i,S}}\right)\right] \\
&= \mathbb{E}\left[\left(1-\frac{\hat{\mathfrak{p}}_{i,S}}{\mathfrak{p}_{i,S}}\right)\mathbb{1}_{\{\Omega_0\}} + \left(1-\frac{1-\hat{\mathfrak{p}}_{i,S}}{1-\mathfrak{p}_{i,S}}\right)\mathbb{1}_{\{\Omega_0^c\}}\right] \\
&\overset{(c)}{\ge} \sqrt{\frac{1}{2\pi T_i}}\left(\sqrt{\frac{1-\mathfrak{p}_{i,S}}{\mathfrak{p}_{i,S}}} + \sqrt{\frac{\mathfrak{p}_{i,S}}{1-\mathfrak{p}_{i,S}}}\right) - \frac{\tilde{C}_i^{(s)}(\mathfrak{p}_{i,S}}{T_i}.
\end{aligned}
$$

In the above display,
**(a)** follows from the argument that '*bunching together*' probability mass can only result in a lower value of the term inside the expectation as formally proved in Lemma 12. **(c)** follows from an application of Theorem 3.2 of (Ross, 2011) and the term $\tilde{C}_i^{(s)}(S)$ is defined as follows:

$$
\tilde{C}_i^{(s)}(S) \coloneqq \sum_{p\in\{\mathfrak{p}_{i,S},1-\mathfrak{p}_{i,S}\}}\left((1-p)(1+2p^2-2p) + \left(2(1-p)\left((1-p)^3+p^3\right)/\pi\right)^{1/2}\right) \tag{49}
$$

Finally, the result follows by first introducing the definition $\tilde{c}_i \coloneqq \max\left\{\left(\sqrt{\frac{1-\mathfrak{p}_{i,S}}{\mathfrak{p}_{i,S}}} + \sqrt{\frac{\mathfrak{p}_{i,S}}{1-\mathfrak{p}_{i,S}}}\right)\ \mid\ S\subset[l],\ 1\le|S|<l\right\}$. Let $S^*$ denote the subset of $[l]$ at which the maximum in the definition of $\tilde{c}_i^{(s)}$ is achieved. Using this we define the remaining term to complete the proof

$$
\tilde{C}_i^{(s)} \coloneqq \tilde{C}_i^{(s)}\left(S^*\right). \tag{50}
$$

□

**Lemma 12.** *Given two probability distributions $P = (p_1, \ldots, p_l)$ and $Q = (q_1, \ldots, q_l)$ in $\Delta_l$ (the $l-1$ dimensional probability simplex), a set $S \subset [l]$ such that $1 \le |S| < l$, and $\mathfrak{p}_S$ and $\mathfrak{q}_S$ be defined as in* (48)*. Then we have the following:*

$$
D_s\left(P, Q\right) \ge \max\left(1-\frac{\mathfrak{p}_S}{\mathfrak{q}_S},\ 1-\frac{1-\mathfrak{p}_S}{1-\mathfrak{q}_S}\right).
$$

*Proof.* Suppose $j_0$ is the index of the maximizer in the definition of $D_s$, i.e., $1 - p_{j_0}/q_{j_0} \geq 1 - p_j/q_j$ for $j \in [l]$, $j \neq j_0$. Equivalently, we have $p_{j_0}/q_{j_0} \leq p_j/q_j$ for all $j \in [l]$, $j \neq j_0$. We then have the following:

$$\frac{p_{j_0}}{q_{j_0}} \leq \sum_{j \in S} \left(\frac{p_j}{q_j}\right) \frac{q_j}{\sum_{j' \in S} q_{j'}} = \frac{\mathfrak{p}_S}{\mathfrak{q}_S}, \qquad \text{and} \qquad \frac{p_{j_0}}{q_{j_0}} \leq \sum_{j \in S^c} \left(\frac{p_j}{q_j}\right) \frac{q_j}{\sum_{j' \in S^c} q_{j'}} = \frac{1 - \mathfrak{p}_S}{1 - \mathfrak{q}_S}.$$

This implies that $D_s\left(P, Q\right) = (1 - p_{j_0}/q_{j_0}) \geq \max\{1 - \mathfrak{p}_S/\mathfrak{q}_S,\ 1 - (1 - \mathfrak{p}_S)/(1 - \mathfrak{q}_S)\}$, which completes the proof.

□

### H.2. Proof of Lemma 7

Assume that the event $\mathcal{E}_1$ defined in Lemma 2 holds, that $\hat{p}_{ij,t} \geq (7e_{ij,t})/2$ for all $1 \leq j \leq l$ and that the distribution $P_i$ lies in $\Delta_l^{(\eta)}$ for $1 \leq i \leq K$.

From the proof of Lemma 10, under the above assumptions we have

$$\frac{1}{p_{ij}} - 1 \leq \frac{1}{\hat{p}_{ij,t} - e_{ij,t}} - 1 \leq \frac{1}{p_{ij}} - 1 + \frac{1}{\eta^2}\sqrt{\frac{8\log(2/\delta_t)}{T_{i,t}}}.$$

We can then proceed as follows:

$$\sqrt{\frac{1}{\hat{p}_{ij,t} - e_{ij,t}} - 1} \overset{(a)}{\leq} \sqrt{\frac{1}{p_{ij}} - 1} + \frac{1}{\eta}\left(\frac{8\log(2/\delta_t)}{T_{i,t}}\right)^{1/4} \quad \text{and}$$
$$\sqrt{\frac{1}{\hat{p}_{ij,t} - e_{ij,t}} - 1} \overset{(b)}{\leq} \sqrt{\frac{1}{p_{ij}} - 1} + \frac{1}{2\eta^{5/2}}\sqrt{\frac{8\log(2/\delta_t)}{T_{i,t}}}.$$

In the above display,
**(a)** follows from the fact that $\sqrt{z_1 + z_2} \leq \sqrt{z_1} + \sqrt{z_2}$ for $z_1, z_2 \geq 0$, and
**(b)** uses the fact that the function $h_2(x) = \sqrt{x}$ is concave, and thus majorized by its first order approximation at the point $x = 1/p_{ij} - 1$. Furthermore, since the derivative of $h_2(x)$ is $1/\left(2\sqrt{x}\right)$, we also use the fact that $p_{ij} \leq 1 - \eta$ as $P_i \in \Delta_l^{(\eta)}$.

Define the terms $a_{i,t}$ and $b_{i,t}$ as follows:

$$a_{i,t} := \left(\frac{8\log(2/\delta_t)}{T_{i,t}}\right)^{1/4} \qquad b_{i,t} := \frac{l}{\eta} a_{i,t} \min\left(1, \frac{a_{i,t}}{2\eta^{3/2}}\right). \tag{51}$$

Then we have the following:

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

We will refer to the $a_{i,n}$ and $b_{i,n}$ corresponding to the oracle allocation scheme (i.e., with $T_{i,t} = T_i^*$) as $a_{i,n}^*$ and $b_{i,n}^*$ respectively and furthermore, define

$$a_n^* := \max_{1 \leq i \leq K} a_{i,n}^*, \quad b_n^* := \max_{1 \leq i \leq K} b_{i,n}^* \tag{52}$$

### H.3. Proof of Theorem 5

**Calculate the values $A$, $\tilde{e}_n$ and $B$.** As before, we have $\widetilde{T}_i^* = \lambda_i^{(s)} n$ where we have $\lambda_i^{(s)} \frac{(c_i^{(s)})^2 n}{(C_s)^2}$ and $(C_s)^2 = \sum_{k=1}^K c_k^2$. We also introduce $\lambda_{\min}^{(s)} = \min_{1 \le i \le K} \lambda_i^{(s)}$ and $\lambda_{\max}^{(s)} = \max_{1 \le i \le K} \lambda_i^{(s)}$. We proceed as follows:

$$
\begin{aligned}
A &\coloneqq \max_{1 \le i \le K} \left. \frac{\partial \varphi\left(c, \widetilde{T}_i^*\right)}{\partial c} \right|_{c = c_i^{(s)}} = \frac{1}{\sqrt{\lambda_{\min}^{(s)} n}} \\
B &\coloneqq \max_{1 \le i \le K} \left. \frac{\partial \varphi(c_i^{(s)}, T)}{\partial T} \right|_{T = \widetilde{T}_i^*} = \min_{1 \le i \le K} \frac{c_i^{(s)}}{2\left(\widetilde{T}_i^*\right)^{3/2}} = \frac{C_s}{\lambda_{\max}^{(s)} n^{3/2}} \\
\tilde{e}_n &= b_n^*,
\end{aligned}
$$

where $b_n^*$ is defined in (52). In the rest of this section, we will use the notation $E \coloneqq \frac{A \tilde{e}_n}{B}$. Note that by definition we have

$$
E = n \frac{\lambda_{\max}^{(s)}}{\sqrt{\lambda_{\min}^{(s)}} C_s} \frac{l a_n^*}{\eta} \min\left(1, \frac{a_n^*}{2\eta^{3/2}}\right). \tag{53}
$$

Introduce the term

$$
N_0 \coloneqq \min\{n \ge 1 \;:\; (n / \log(2/\delta_n)) \ge 2/(\lambda_{\min}^{(s)} \eta^6)\}, \tag{54}
$$

where $\delta_n$ is defined in Lemma 2. Then for $n \le N_0$, the term $E$ is $\tilde{\mathcal{O}}\left(n^{1/2}\right)$, while for larger values of $n$, $E$ is $\tilde{\mathcal{O}}\left(n^{3/4}\right)$.

**Regret bound derivation.** Since we are using an approximate objective function, we employ the decomposition of the regret given in (13) in the statement of Proposition 1 in Appendix B:

$$
\mathcal{R}_n\left(\mathcal{A}_s, D_{\mathrm{s}}\right) \le \mathcal{L}_n\left(\mathcal{A}_s, D_{\mathrm{s}}\right) - \varphi\left(c_i^{(s)}, \widetilde{T}_i^*\right) + 2 \max_{1 \le i \le K} \left| R_i\left(\widetilde{T}_i^*\right)\right|. \tag{55}
$$

We first note that the remainder term can be upper bounded by an application of Lemma 6 as

$$
\begin{aligned}
2 \max_{1 \le i \le K} \left| R_i\left(\widetilde{T}_i^*\right)\right| &\le 2 \max_{1 \le i \le K} \left( \left(c_i^{(s)} - \tilde{c}_i^{(s)}\right) \sqrt{\frac{1}{2\pi \widetilde{T}_i^*}} + \frac{C_i^{(s)} - \tilde{C}_i^{(s)}}{\widetilde{T}_i^*} \right) \\
&= 2\left(c_i^{(s)} - \tilde{c}_i^{(s)}\right) \sqrt{\frac{1}{2\pi \lambda_{\min}^{(s)}}} + \mathcal{O}\left(\frac{1}{n}\right).
\end{aligned} \tag{56}
$$

First we get an upper bound on $\mathcal{L}_n\left(\mathcal{A}_s, D_{\mathrm{s}}\right)$. As in the earlier sections, we consider the probability $1 - \delta$ event under which we have $\tau_{0,i} \le T_i \le \tau_{1,i}$, and $\tau_{1,i} - \tau_{0,1} \le 2A\tilde{e}_n / B$. Then we replace the random sum in the computation of the expectation with a sum of a (deterministic) constant number of terms.

$$
\begin{aligned}
\mathcal{L}_n\left(\mathcal{A}_s, D_s\right) \coloneqq \mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] &= \frac{1}{\sqrt{T_i}} \mathbb{E}\left[\max_{1 \le j \le l} \sum_{t=1}^{T_i} \frac{p_{ij} - Z_{ij}^{(t)}}{\sqrt{p_{ij}(1 - p_{ij}) T_i}} \sqrt{\frac{1 - p_{ij}}{p_{ij}}}\right] \\
&\overset{(a)}{\le} \frac{\delta}{\eta} + \frac{1}{\sqrt{\tau_{0,i}}} \mathbb{E}\left[\max_{1 \le j \le l} \sum_{t=1}^{\tau_{0,i}} \frac{p_{ij} - Z_{ij}^{(t)}}{\sqrt{p_{ij}(1 - p_{ij}) \tau_{0,i}}} \sqrt{\frac{1 - p_{ij}}{p_{ij}}}\right] + c_i^{(s)} \sqrt{\frac{2(\tau_{1,i} - \tau_{0,i})}{\tau_{0,i}^2}} \\
&\overset{(b)}{\le} \frac{\delta}{\eta} + \varphi(\tau_{0,i}, c_i^{(s)}) + \frac{C_i^{(s)}}{\tau_{0,i}} + c_i^{(s)} \sqrt{\frac{4E}{\tau_{0,i}^2}}
\end{aligned}
$$

In the above display,
**(a)** follows from the fact that the event $\mathcal{E}_1$ (defined in Lemma 2) occurs with probability at least $1 - \delta$, that the separation distance is always upper bounded by $1/\eta$ for $P_i \in \Delta_l^{(\eta)}$, and that under the event $\mathcal{E}_1$, we have $\tau_{0,i} \le T_i \le \tau_{1,i}$,
**(b)** follows from Lemma 7 and the definition of the term $E$ in (53).

Next, by exploiting the convexity of the mappings $x \mapsto 1/x$ and $x \mapsto 1/\sqrt{x}$, we can obtain the following relations.

$$\varphi\left(\tau_{0,i}, c_i^{(s)}\right) = \frac{c_i^{(s)}}{\sqrt{\widetilde{T}_i^* \left(1 - \frac{\widetilde{T}_i^* - \tau_{0,i}}{\widetilde{T}_i^*}\right)}} \leq \frac{c_i^{(s)}}{\sqrt{\widetilde{T}_i^*}} + \frac{2c_i^{(s)} E}{\left(\lambda_{\text{min}}^{(s)} n\right)^{3/2}} \tag{57}$$

$$\frac{C_i^{(s)}}{\tau_{0,i}} \leq \frac{C_i^{(s)}}{\widetilde{T}_i^*} + \frac{2C_i^{(s)} E}{\left(\widetilde{T}_i^*\right)^2} \leq \frac{C_i^{(s)}}{\widetilde{T}_i^*} + \frac{2C_i^{(s)} E}{\left(\lambda_{\text{min}}^{(s)} n\right)^2} \tag{58}$$

$$\frac{2c_i^{(s)}\sqrt{E}}{\tau_{0,i}} \leq \frac{2c_i^{(s)}\sqrt{E}}{\widetilde{T}_i^*} + \frac{4c_i^{(s)} E^{3/2}}{\left(\widetilde{T}_i^*\right)^2} \tag{59}$$

In the above display, we have assumed that $n$ is large enough to ensure that $(\widetilde{T}_i^* - \tau_{0,i})/\widetilde{T}_i^* \leq 1/2$ for all $1 \leq i \leq K$. A sufficient condition for this is that

$$E \leq (\lambda_{\text{min}}^{(s)} n)/2, \tag{60}$$

where $E$ is defined in (53).

Combining the relations in the previous display, we get the following bound for the expected separation distance between the empirically constructed $\hat{P}_i$ and the true distribution $P_i$.

$$\mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] - \varphi\left(c_i^{(s)}, \widetilde{T}_i^*\right) \leq \frac{\delta}{\eta} + \frac{C_i^{(s)} + 2c_i^{(s)}\sqrt{E}}{\widetilde{T}_i^*} + \frac{2c_i^{(s)} E}{\left(\lambda_{\text{min}}^{(s)} n\right)^{3/2}} + \frac{2C_i^{(s)} E}{\left(\lambda_{\text{min}}^{(s)} n\right)^2} + \frac{4c_i^{(s)} E^{3/2}}{\left(\lambda_{\text{min}}^{(s)} n\right)^2} \tag{61}$$

The final expression for regret stated in Theorem 5 follows by plugging (61) and (56) in the regret decomposition given in (55).

Now we show that the first term in RHS of (55) is either $\tilde{\mathcal{O}}\left(n^{-3/4}\right)$ or $\tilde{\mathcal{O}}\left(n^{-5/8}\right)$ depending on whether $n \geq$ or $< N_0$.

**Case 1: $n \geq N_0$.** In this case we have $E = \left(\frac{\lambda_{\text{max}}^{(s)} (a_n^*)^2}{C\sqrt{\lambda_{\text{min}}^{(s)}} 2\eta^{5/2}}\right) n^{1/2} := M_1\sqrt{n}$. Then we have

$$\mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] - \varphi\left(c_i^{(s)}, \widetilde{T}_i^*\right) \leq \frac{2c_i^{(s)}\sqrt{M_1}}{\lambda_{\text{min}}^{(s)} n^{3/4}} + \frac{\delta}{\eta} + \frac{1}{n}\left(\frac{C_i^{(s)}}{\lambda_{\text{min}}^{(s)}} + \frac{2c_i^{(s)} M_1}{\left(\lambda_{\text{min}}^{(s)}\right)^{3/2} n} + \frac{4M_1^2}{\left(\lambda_{\text{min}}^{(s)}\right)^2}\right) + \frac{2C_i^{(s)} M_1}{\left(\lambda_{\text{min}}^{(s)}\right)^2 n^{3/2}}. \tag{62}$$

If $\delta \leq \eta/n$, then we have that

$$\mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] - \varphi\left(c_i^{(s)}, \widetilde{T}_i^*\right) \leq \frac{2c_i^{(s)}\sqrt{M_1}}{\lambda_{\text{min}}^{(s)} n^{3/4}} + \tilde{\mathcal{O}}\left(\frac{1}{n}\right). \tag{63}$$

**Case 2: $n < N_0$.** In this case, we have $E = n^{3/4} \frac{\lambda_{\text{max}}^{(s)}}{\sqrt{\lambda_{\text{min}}^{(s)}} C_s} \frac{l a_n^*}{\eta} := M_2 n^{3/4}$, which implies that

$$\mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] - \varphi\left(c_i^{(s)}, \widetilde{T}_i^*\right) \leq \frac{2c_i^{(s)}\sqrt{M_2}}{\lambda_{\text{min}}^{(s)}} n^{-5/8} + \frac{\delta}{\eta} + \frac{2C_i^{(s)} M_2 n^{-3/4}}{\left(\lambda_{\text{min}}^{(s)}\right)^2} + \frac{2C_i^{(s)} M_2 n^{-5/4}}{\left(\lambda_{\text{min}}^{(s)}\right)^2} + \frac{4c_i^{(s)} M_2^{3/2} n^{-7/8}}{\left(\lambda_{\text{min}}^{(s)}\right)^2}, \tag{64}$$

which implies that in this case we have

$$\mathbb{E}\left[D_s\left(\hat{P}_i, P_i\right)\right] - \varphi\left(c_i^{(s)}, \widetilde{T}_i^*\right) \leq \frac{2c_i^{(s)}\sqrt{M_2}}{\lambda_{\text{min}}^{(s)}} n^{-5/8} + \tilde{\mathcal{O}}\left(n^{-3/4}\right). \tag{65}$$

# I. Deferred Proofs from Section 5

In this section, we present the proof of Theorem 6 and Corollary 1 which were stated in Section 5.

## I.1. Proof of Theorem 6

For any adaptive allocation scheme $\mathcal{A}$, recall that $T_1$ and $T_2$ refer to the number of samples drawn from the first and second arms respectively. Since $K = 2$, we have $T_1 + T_2 = n$ where $n$ is the sampling budget. Denote by $\Omega = \{0,1\}^n$, the observation space. The allocation scheme $\mathcal{A}$ along with the distributions of the two arms induce a probability measure on $\Omega$. We use $\mathbb{P}_1$ and $\mathbb{P}_2$ to represent the two probability measures on $\Omega$ corresponding to the two problem instances $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively. Finally recall that we use $(\widetilde{T}_i^*)_{i=1}^2$ to represent the oracle allocation scheme corresponding to the objective functions $\varphi(c_i, T_i) = c_i/T_i^\alpha$ for $\alpha > 0$. For the case of problem $\mathcal{P}_1$, we have $c_1 = \kappa\left(()\, p_0\right)$ and $c_2 = \kappa\left(()\, p_0 - \epsilon\right)$ and these terms are swapped for the problem $\mathcal{P}_2$. Note that since $K = 2$, we have $|\widetilde{T}_1^* - n/2| = |\widetilde{T}_2^* - n/2|$ for both problem instances.

Now, without loss of generality we assume that $\kappa\left(p_0\right) < \kappa\left(p_0 - \epsilon\right)$, and define the event $\mathcal{E} \coloneqq \{T_1 < n/2\}$. The other case, (i.e., $\kappa\left(p_0\right) < \kappa\left(p_0 - \epsilon\right)$) can be handled in an analogous manner by considering the event $\mathcal{E}^c$. By an application of the standard change-of-measure result (Kaufmann & Garivier, 2017, Lemma 1), we have

$$\mathrm{kl}\left(\mathbb{P}_1\left(\mathcal{E}\right), \mathbb{P}_2\left(\mathcal{E}\right)\right) \leq \mathbb{E}_1\left[T_1\right] \mathrm{kl}\left(p_0, p_0 - \epsilon\right) + \mathbb{E}_1\left[T_2\right] \mathrm{kl}\left(p_0 - \epsilon, p_0\right). \tag{66}$$

In the above display, we use the notation $\mathrm{kl}(p,q)$ to represent the kl-divergence between $\mathrm{Ber}\left(p\right)$ and $\mathrm{Ber}\left(q\right)$ random variables. Next, we note that for any $p, q \in (0,1)$ we have $2(p-q)^2 \leq \mathrm{kl}(p,q)$ by an application of Pinsker's inequality. Furthermore, due to the concavity of $\log(x)$, we can also show that $\mathrm{kl}(p,q) \leq \frac{(p-q)^2}{q(1-q)}$. Combining these two observations with (66), along with the assumption that $\epsilon < p_0 - 1/2$, we have

$$2\left(\mathbb{P}_1\left(\mathcal{E}\right) - \mathbb{P}_2\left(\mathcal{E}\right)\right)^2 \leq \mathbb{E}_1\left[T_1\right] \frac{\epsilon^2}{(p_0 - \epsilon)(1 - p_0 + \epsilon)} + \mathbb{E}_1\left[T_2\right] \frac{\epsilon^2}{p_0(1-p_0)} \leq \frac{2\epsilon^2}{1 - p_0}\left(\mathbb{E}_1\left[T_1 + T_2\right]\right) = \frac{2\epsilon^2 n}{1 - p_0}.$$

From the above display, we can conclude that

$$|\mathbb{P}_1\left(\mathcal{E}\right) - \mathbb{P}_2\left(\mathcal{E}\right)| \leq \epsilon\sqrt{\frac{n}{1-p_0}}. \tag{67}$$

Introducing the notation $\delta = \frac{1 - \epsilon\sqrt{n/(1-p_0)}}{2}$, we note that the inequality (67) implies that at least one the following two statements is true: **(1)** $\mathbb{P}_1\left(\mathcal{E}^c\right) \geq \delta$, or **(2)** $\mathbb{P}_2\left(\mathcal{E}\right) \geq \delta$. With the notation $\tau \coloneqq |\widetilde{T}_i^* - n/2|$, we note that under the event $\mathcal{E}$ for the problem instance $\mathcal{P}_1$, we have $|T_1 - \widetilde{T}_1^*| \geq \tau$ and under the event $\mathcal{E}^c$ for problem instance $\mathcal{P}_2$ we have $|T_2 - \widetilde{T}_1^*| \geq \tau$. This implies that we have the following:

$$\max_{\mathcal{P}_1, \mathcal{P}_2} \max_{i=1,2} \mathbb{E}\left[|T_i - \widetilde{T}_i^*|\right] \geq \delta\tau = \left(\frac{\left(1 - \epsilon\sqrt{n/(1-p_0)}\right)\tau}{2}\right). \tag{68}$$

Finally, the result of the statement follows from the following two observations:

$$\begin{aligned} &|\widetilde{T}_1^* - \widetilde{T}_2^*| = 2|\widetilde{T}_1^* - n/2| = 2\tau \text{ and} \\ &|\widetilde{T}_1^* - \widetilde{T}_2^*| = n\left|\frac{\left(\kappa\left(p_0\right)\right)^{1/\alpha} - \left(\kappa\left(p_0 - \epsilon\right)\right)^{1/\alpha}}{\left(\kappa\left(p_0\right)\right)^{1/\alpha} + \left(\kappa\left(p_0 - \epsilon\right)\right)^{1/\alpha}}\right|. \end{aligned}$$

## I.2. Proof of Corollary 1

We first note that with a choice of $\epsilon = \frac{1}{4\sqrt{n}}$ and $p_0 = 3/4$, we have $\delta = \frac{1 - \epsilon\sqrt{n/(1-p_0)}}{2} = 1/4$, and the RHS of (68) becomes $\tau/4$. To complete the proof, it suffices to show that $|\kappa\left(p_0\right)^{1/\alpha} - \kappa\left(p_0 - \epsilon\right)^{1/\alpha}| = \Omega(\epsilon)$, for the $\kappa$ corresponding to the three loss functions $D_{\ell_2}, D_{\ell_1}$ and $D_{\mathrm{s}}$. The result then follows from the definition of $\tau$ and the choice of $\epsilon$. We consider the three cases separately.

**$\ell_2^2$-distance.** In this case, we have $\kappa(p) = 1-p^2-(1-p)^2$ and $\alpha = 1$. Thus we have $|\kappa(p_0)-\kappa(p_0-\epsilon)| = \epsilon-2\epsilon^2 \geq \epsilon/2$, where the last inequality follows from the fact that for $\epsilon = 1/4\sqrt{n}$, we have $\epsilon^2 \leq \epsilon/2$.

**$\ell_1$-distance.** In this case, we have $\kappa(p) = 2\sqrt{p(1-p)}$ and $\alpha = 1/2$. Thus, we have for $p_0 = 3/4$ $|\kappa(p_0)^{1/\alpha} - \kappa(p_0-\epsilon)^{1/\alpha}| = 4(\epsilon/2-\epsilon^2) \geq \epsilon$. In the last inequality we used the fact that for the choice of $\epsilon$ mentioned earlier, we have $\epsilon^2 \leq \epsilon/4$.

**Separation distance.** In this case, we have $\kappa(p) = \sqrt{1/p-1} + \sqrt{1/(1-p)+1}$ and $\alpha = 1/2$. Thus, we have for $p_0 = 3/4$, $|\kappa(p_0)^{1/\alpha} - \kappa(p_0-\epsilon)^{1/\alpha}| = \frac{1-2(1-p_0)}{p_0(1-p_0)} - \frac{1-2(1-p_0+\epsilon)}{(p_0-\epsilon)(1-p_0+\epsilon)} := \zeta(1/4) - \zeta(1/4+\epsilon)$, where we have defined $\zeta(z) = \frac{1-2x}{x(1-x)}$. Note that for $x \in [1/4, 1/2]$ the function $\zeta$ is decreasing and convex, with $\zeta(1/2) = 0$. Thus we have $\zeta(1/4) - \zeta(1/4+\epsilon) \geq \left(\frac{0-\zeta(1/4)}{1/2-1/4}\right)(1/4-(1/4+\epsilon)) = \frac{2\epsilon}{3}$. In the last inequality, we used the fact that $\zeta(x)$ is majorized by the straight line joining $(1/4, \zeta(1/4))$ and $(1/2, 0)$ due to the convexity of $\zeta$ in this domain.

Thus for all the three distances, we have $\tau = (n/2)\frac{|\kappa(p_0)^{1/\alpha}-\kappa(p_0-\epsilon)^{1/\alpha}|}{|\kappa(p_0)^{1/\alpha}+\kappa(p_0-\epsilon)^{1/\alpha}|} = \Omega(\epsilon n) = \Omega(\sqrt{n})$ as required.

## J. Details of Extension to Continuous Distributions in Section 7

We introduce the notation $\mathcal{X} = [0,1]$ and let $L^2(\mathcal{X})$ denote the set of square integrable functions on $\mathcal{X}$. Suppose $(\psi_j)_{j=1}^l$ are ortho-normal functions in $L^2(\mathcal{X})$ and define $\mathcal{G}_l = \text{span}(\{\psi_j \ : \ 1 \leq j \leq l\})$. Now suppose $(P_i)_{i=1}^K$ are continuous distributions on $\mathcal{X}$ which admit density functions $(\nu_i)_{i=1}^K$ with respect to the uniform distribution (denoted by $\mu$), and further assume that the density functions lie in $\mathcal{G}_l$.

Suppose we have $\nu_i = \sum_{j=1}^l a_{ij}\psi_j$, then we have the following:

$$\int_{\mathcal{X}} |\nu_i|^2 d\mu = \sum_{j=1}^l a_{ij}^2 = \|a_i\|_2^2, \quad \text{and} \quad a_{ij} = \int_{\mathcal{X}} \nu_i \psi_j d\mu := \langle \nu_i, \, \psi_j \rangle.$$

Thus in order to estimate $\nu_i$ in squared-error sense, it suffices to estimate the coefficients $(a_{ij})_{j=1}^l$ and we can employ the projection estimators defined as:

$$\hat{a}_{ij} = \frac{1}{T_i}\sum_{s=1}^{T_i} \psi_j(X_i^{(s)}), \quad \text{and} \quad \hat{\nu}_i = \sum_{j=1}^l \hat{a}_{ij}\psi_j.$$

Finally, due to the orthonormality of the basis functions, we have the following:

$$\mathbb{E}\left[\int_{\mathcal{X}} (\hat{\nu}_i - \nu_i)^2 d\mu\right] = \mathbb{E}\left[\|\hat{a}_i - a_i\|_2^2\right] = \frac{1}{T_i}\sum_{j=1}^l \text{var}(\psi_j(X_i)) := \frac{c_i}{T_i}.$$

Since the resulting objective function lies in $\mathcal{F}$, we can construct an appropriate upper-bound for the parameters $c_i$ based on the choice $(\psi_j)_{j=1}^l$ and proceed as described in Section 4.2 to obtain the regret bounds.