Learning a Hidden Random Variable from Correlated Bandit Samples

Samarth Gupta
Dept. of ECE
Carnegie Mellon University
Pittsburgh, PA 15213
Email: samarthg@andrew.cmu.edu

Gauri Joshi
Dept. of ECE
Carnegie Mellon University
Pittsburgh, PA 15213
Email: gaurij@andrew.cmu.edu

Osman Yağan
Dept. of ECE
Carnegie Mellon University
Pittsburgh, PA 15213
Email: oyagan@ece.cmu.edu

***Abstract*—We consider the task of learning the probability distribution p_X of a discrete random variable X whose samples are not observed directly. At each time slot, we choose one of the possible K functions, $g_1, \ldots, g_K$ and observe the corresponding sample $g_i(X)$. The goal is to estimate the probability distribution of X by generating minimum number of samples. This problem is relevant in inference under non-precise information and privacy preserving statistical estimation. We reveal the conditions on the functions under which asymptotically consistent estimation is possible. We also derive lower bounds on the estimation error as a function of total samples and show that it is *order-wise* achievable. Finally, we propose an iterative algorithm that chooses the function to observe at each step and combines the obtained samples to estimate p_X. We demonstrate the performance of this algorithm across different scenarios through simulations.**

***Index Terms*—distribution learning, hidden random variable, multi-arm bandits, correlated arms**

I. Introduction

The modern world is rich with various types of data such as images, video, cloud job execution traces, social network data, and crowdsourced survey data. These data can provide invaluable insights into the underlying random phenomenon which are generally not directly observable due to privacy concerns, or imprecise measurement mechanisms. For example, if we want to estimate the income distribution of a population, their salary data may not be public. However, it may be possible to estimate the income distribution using surveys about their spending on luxury goods, or whether their income is above or below some given thresholds.

In this work we seek to design techniques to use indirect and correlated samples to estimate the probability distribution of a hidden random phenomenon. We consider a stylized model, shown in Fig. 1, where a hidden variable X can be sampled through functions $g_1(X), \ldots, g_K(X)$, referred to as *arms*. Our objective is to accurately estimate the probability distribution of X with the minimum number of samples; see Section section II for a precise definition of the problem.

Learning the distribution of a random variable from its samples is a well-studied research problem [1], [2] in theoretical computer science. Some works [3], [4] are interested in finding the min-max or worst case loss for various loss functions; e.g., L2-loss and Kullback-Liebler (KL) divergence. Unlike the majority of the literature on distribution learning,

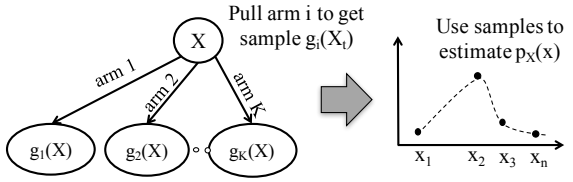

Fig. 1: *At step t we pull some arm i and observe $g_i(X_t)$, where X_t is an i.i.d. realization of the hidden variable X. Our objective is use the samples to estimate the distribution $p_X(x)$.*

here we assume that only functions $g_i(X)$ of the samples can be observed instead of direct samples of X. Inferring a hidden random variable from indirect samples is related to many works in estimation theory [5], where the objective is to estimate a set of parameters θ using observations $y_1, \ldots, y_n$ that follow a model $p(Y|\theta)$. In our problem, the hidden variable X is analogous to the parameter θ while samples $g_i(X_t)$ correspond to the observations $y_1, \ldots, y_n$.

A key difference between our model and typical parameter estimation problems is that we decide on the arm to be pulled in each time slot to obtain the corresponding sample $g_i(X_t)$. This aspect is closely related to the multi-arm bandit (MAB) sequential decision-making framework [6]. In the classical MAB framework [7], each arm gives a reward according to some unknown distribution that is independent across arms, and the objective is to maximize the total reward. In contrast, in our problem, the arms $g_1(X), \ldots, g_K(X)$, are correlated through the common hidden variable X. Contextual bandits [8]–[10] consider a context vector x for each arm that governs its reward distribution, unlike a common hidden variable X as considered in this paper.

To the best of our knowledge, this is the first work to consider the problem of using sequential, indirect samples to learn the distribution of a hidden random variable. Our main contributions are: 1) conditions on the functions $g_1(X), g_2(X), \ldots g_K(X)$ for asymptotically consistent estimation of the hidden distribution, 2) lower bound on the estimation error, and 3) proposing an algorithm to choose arms and combine their samples, which successfully eliminates redundant arms. Simulations indicate that the error performance of our algorithm is better than several baseline strategies.

II. Problem Formulation

Consider a discrete random variable X that can take values from a finite alphabet $\{x_1, x_2, \ldots, x_n\}$ with an *unknown* probability distribution $\{p_1, p_2, \ldots, p_n\}$. Throughout this paper, we assume $p_i > 0, \forall i$. Our objective is to estimate this probability distribution using a sequence of independent samples from K functions $\{g_1, g_2, \ldots, g_K\}$, where each g_i is a mapping from $\{x_1, x_2, \ldots, x_n\}$ to $\mathbb{R}$; throughout, we refer to these functions also as *arms*. More precisely, with $\{X_t : t = 1, 2, \ldots\}$ denoting a sequence of independent and identically distributed (i.i.d.) realizations of X, we shall choose and observe only one of the K possible outcomes $g_1(X_t), \ldots, g_K(X_t)$, at each step $t \in \mathcal{N}$. Broadly speaking, for a given set of functions $\{g_1, g_2, \ldots, g_K\}$, our goal is to derive an efficient and *powerful* algorithm i) to decide which function will be observed at each iteration step, and ii) to come up with an estimate $\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n\}$ of the true probability distribution based on these observations. Ultimately, we aim to minimize the mean-squared error, formally defined below.

Definition 1 (Estimation Error)**.** *The error in estimating $\{p_1, p_2, \ldots, p_n\}$ at step t (i.e., after observing t samples) is defined as*

$$\varepsilon(t) = \mathbb{E}\left[\sum_{i=1}^{n} (\hat{p}_i(t) - p_i)^2\right]. \tag{1}$$

Here, $\hat{p}_i(t)$ denotes the estimation obtained after observing t samples $g_{c_1}(X_1), g_{c_2}(X_2), \ldots, g_{c_t}(X_t)$, where $c_j \in \{1, \ldots, K\}$ is the arm pulled at step j. We now give two examples to illustrate and clarify the problem formulation.

Example 1. *Fig. 2 shows an example in which X takes three possible values $\{x_1, x_2, x_3\}$, and there are three arms, g_1, g_2 and g_3. Output of g_1, g_2 and g_3 corresponding to x_1, x_2, x_3 are illustrated in Fig. 2. In arm 1, output b can come from either x_2 or x_3. This ambiguity exists in output d (between x_1 and x_3) in g_2 and in output f (between x_1 and x_2) in g_3.*

Example 2. *Fig. 3 illustrates an example with two arms, with each arm showing outputs corresponding to $\{x_1, x_2, x_3, x_4\}$. Arm 1 has ambiguity coming from output of x_2 and x_3, whereas arm 2 exhibits ambiguity in the output of x_1, x_2 and x_3. For this set of functions it is not possible to* accurately *estimate p_2 and p_3, as we will prove in Section III-A.*

We note that if a function g_i is invertible, then every output sampled from g_i will be uniquely matched to a single value (say, x_j) that X can take without any ambiguity. In those cases, it would be optimal (in the sense of minimizing $\varepsilon(t)$ for each t) to pull g_i at every step. We formally prove a more general version of this result in Theorem 2.

III. Results

A. Conditions for asymptotically consistent estimation

A natural question is whether it is possible to estimate the true distribution *accurately* when the number t of steps grows unboundedly large. The answer to this question depends on the functions $g_1, g_2, \ldots, g_K$ as we now show.

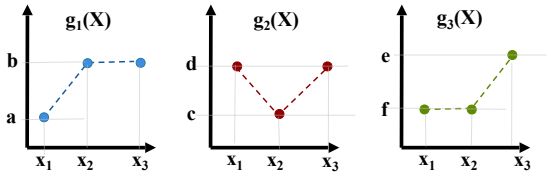

Fig. 2: *An example where it is possible to estimate $\{p_1, p_2, p_3\}$ (asymptotically consistently) although no arm is invertible.*

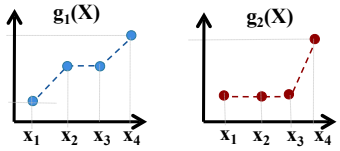

Fig. 3: *An example where it is not possible to get consistent estimation due to the ambiguity between p_2 and p_3.*

Definition 2 (Asymptotically consistent estimation)**.** *Given a random variable X and arms $\{g_1, g_2, \ldots, g_K\}$, we call an estimation asymptotically consistent if* $\lim_{t\to\infty} \varepsilon(t) = 0$.

For each $k = 1, \ldots, K$, let $\{o_{1,k}, o_{2,k}, \ldots, o_{m_k,k}\}$ denote the set of possible outcomes (i.e., the domain) of the function g_k; evidently, m_k is the number of distinct outputs of g_k. We find it useful to construct a matrix A_k with m_k rows and n columns, where

$$A_k(i,j) = \begin{cases} 1, & \text{if } g_k(x_j) = o_{i,k} \\ 0, & \text{otherwise,} \end{cases}$$

for each $i = 1, \ldots, m_k$ and $j = 1, \ldots, n$. Informally, $A_k(i,j) = 1$ if output $o_{i,k}$ could have been generated by x_j in arm k. We refer A_k as the *Sample Generation Matrix* for arm k. Let the matrix A be given by $A = [A_1^\intercal, A_2^\intercal, \ldots, A_K^\intercal]^\intercal$; the size of A is $m \times n$, where $m = m_1 + \ldots + m_K$. The corresponding matrices $A^{\text{Example-1}}$ and $A^{\text{Example-2}}$ for Examples 1 and 2, respectively are shown below.

$$A^{\text{Example-1}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A^{\text{Example-2}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Theorem 1. *It is possible to achieve asymptotically consistent estimation if and only if rank$(A) = n$.*

Proof of Theorem 1. Recall that $o_{i,k}$ represents the i^{th} distinct output of arm k. Let $p_{o_{i,k}}$ denote the probability of observing $o_{i,k}$ each time arm k is pulled. Consider the system of linear equations below relating these probabilities to the probability distribution of X:

$$p_{o_{i,k}} = \sum_{z=1}^{n} A_k(i,z) p_z, \qquad \begin{matrix} k = 1, \ldots, K \\ i = 1, \ldots, m_k \end{matrix} \tag{2}$$

Suppose now that A is full rank. In order to construct an asymptotically consistent estimate of $p_X = \{p_1, \ldots, p_n\}$, assume that arms are pulled in a round robin manner. Thus, at step t we will have $\frac{t}{K}$ samples from each arm. With $n_{o_{i,k}}$ denoting the number of times $o_{i,k}$ is observed in t steps, we let $\hat{p}_{o_{i,k}}(t) = \frac{n_{o_{i,k}}}{t/K}$ be the estimate of $p_{o_{i,k}}$ at step t. By virtue of Strong Law of Large Numbers, we have $\hat{p}_{o_{i,k}}(t) \to p_{o_{i,k}}$ almost surely as t goes to infinity. Put differently, the estimates $\hat{p}_{o_{i,k}}(t)$ are asymptotically consistent. Given that A is full rank, the estimates $\hat{p}_{o_{i,k}}(t)$ can be used to obtain a unique solution of $\{p_1, p_2, \ldots, p_n\}$ from the system of equations (2). Given that n is finite, this unique solution will constitute an asymptotically consistent estimation of p_X as well.

Conversely, if $\text{rank}(A) < n$, it is not possible to obtain a unique solution of the system of equations in (2). This implies that even if consistent estimation of each $p_{o_{i,k}}$ is possible, it is not possible to achieve asymptotically consistent estimation of the probability distribution $\{p_1, \ldots, p_n\}$ of X. □

Clearly, $\text{rank}(A^{\text{Example-1}}) = n$ while $\text{rank}(A^{\text{Example-2}}) < n$. Thus, asymptotically consistent estimation is possible for the set of functions in Example 1 but not in Example 2.

B. Eliminating redundant functions/arms

Recall the definition of sample generation matrix A_k for each arm k given in Section III-A.

Definition 3 (Subset Arm)**.** *An arm r is said to be a* subset *of another arm s if the row space of A_r is a subset of the row space of A_s.*

Informally, this means that all information produced by arm r can be generated by arm s. For example, in Fig. 3 we see that arm 2 generates information about $p_1+p_2+p_3$, while arm 1 generates information about p_1 and p_2+p_3 separately; also, both arms generate information about p_4 separately. Therefore, information produced by arm 2 can be generated by arm 1. This observation is made precise next.

Theorem 2. *If an arm r is a subset of some other arm s, then it is suboptimal (for the purposes of minimizing $\varepsilon(t)$) to pull arm r at any round t.*

Proof of Theorem 2. Suppose we have two arms r and s with sample generation matrices A_r and A_s respectively. Let arm r be a subset of arm s according to Definition 3.

Samples from arm s give us $n_{o_{i,k}}$ which can be used to estimate the output probabilities $p_{o_{i,k}}$ as given in the proof of Theorem 1. Since the rows A_r can be generated by A_s, we can use arm s samples to generate estimates for the output probabilities of arm r. For example, if rows i and j of A_s can be added to generate a row k of A_r, then the output probability $p_{o_{r,k}}$ is given by

$$p_{o_{r,k}} = \sum_{z=1}^{n} A_r(k,z)p_z = \sum_{z=1}^{n}(A_s(i,z) + A_s(j,z))p_z$$
$$= p_{o_{s,i}} + p_{o_{s,j}}$$

Therefore, pulling arm s is at least as good as pulling arm r. Hence, pulling arm r is suboptimal. □

By Theorem 2, if an *invertible* arm exists, all other arms will be its subset. This leads to the following corollary.

Corollary 1. *If there is an invertible arm, then the optimal action (for the purpose of minimizing $\varepsilon(t)$) is to pull the invertible arm at every round.*

Remark 1. *The proofs of Theorem 2 and Corollary 1 are not specific to the error metric given in Definition 1. Thus, both results hold true under other error metrics as well; e.g., L1 norm, KL divergence etc.*

C. Lower bounds for unbiased estimators

We first derive a *naive* lower bound on the estimation error.

Theorem 3 (Naive Lower Bound)**.** *Estimation error of any unbiased estimator for the problem in Section II is lower bounded by $\sum_{i=1}^{n} \frac{p_i(1-p_i)}{t}$.*

Proof of Theorem 3. From Corollary 1, we know that it is optimal to always pull the invertible arm if there exists one. It is also clear that the optimal error can only decrease when an additional arm is included in the set of possible arms we can choose. Thus, for the purpose of deriving a lower bound on the estimation error, we can assume the existence of an invertible arm.

We define $\hat{p}_i(t) = \frac{n_{x_i}}{t}$, as the corresponding empirical estimator (which is also the maximum likelihood estimator), where n_{x_i} is the number of times the output corresponding to x_i was observed (from the invertible arm) in t steps. Under this scenario, the estimation error is given by

$$\varepsilon(t) = \sum_{i=1}^{n} \text{Var}\left[\hat{p}_i(t)\right] = \sum_{i=1}^{n} \frac{p_i(1-p_i)}{t}, \tag{3}$$

which also gives the minimum possible variance for any unbiased estimator (given the samples from the invertible arm). Using this fact and Corollary 1, we establish Theorem 3.

□

Remark 2. *The lower bound in Theorem 3 is achieved if an invertible arm exists.*

Next, we derive a lower bound on the error of any unbiased estimator given the number of times each arm is pulled.

Theorem 4 (Lower bound Given the Number of Pulls)**.** *Let $t_1, t_2, \ldots, t_K$ be the number of times arms $g_1, \ldots, g_K$ are pulled, respectively. The estimation error of any unbiased estimator satisfies*

$$\varepsilon(t) \geq tr(I(\theta)^{-1}) + \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} I(\theta)^{-1}(i,j), \tag{4}$$

where $I(\theta)$ is an $n-1 \times n-1$ matrix with entries

$$I_{i,j}(\theta) = \sum_{k=1}^{K}\sum_{\ell=1}^{m_K} \frac{-t_k A_k(\ell,i)A_k(\ell,j)(1-A_k(\ell,n))}{p_{o_{\ell,k}}} + \frac{-t_k(1-A_k(\ell,i))(1-A_k(\ell,j)A_k(\ell,n)}{p_{o_{\ell,k}}}. \quad (5)$$

Proof. We use the Cramer-Rao bound [11], [12] that provides a lower bound on the covariance matrix of any unbiased estimator of an unknown deterministic parameter. Since $\sum_{i=1}^{n} p_i = 1$ it suffices to estimate any $n-1$ of the parameters $\{p_1, p_2, \ldots, p_n\}$. Let these parameters ($\theta = \{\theta_1, \theta_2, \ldots, \theta_{n-1}\}$) be $\{p_1, p_2, \ldots, p_{n-1}\}$. Let $\mathcal{D}_t$ be the event that after t slots, we observe the i^{th} output from arm k $n_{o_{i,k}}$ times, for all $k \in [1, K]$, and $i \in [1, m_k]$.

We evaluate the log likelihood $L(\mathcal{D}_t;\theta)$ of observed data $\mathcal{D}_t$ with respect to θ, We then compute the $n-1 \times n-1$ Fisher information matrix, $I(\theta)$, whose $(i,j)^{th}$ entry is given by $-\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(\mathcal{D}_t;\theta)|t_1,\ldots,t_K\right]$. For our problem, we obtain a closed form expression of $I_{i,j}(\theta)$ given in (5).

The Cramer-Rao lower bound on covariance matrix of θ for any unbiased estimator is then given by $I(\theta)^{-1}$. Our objective is to minimize $\sum_{i=1}^{n} \mathrm{Var}\left[\hat{p}_i\right]$, which can be bounded as

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{Var}\left[\hat{p}_i\right] &= \sum_{i=1}^{n-1} \mathrm{Var}\left[\hat{p}_i\right] + \mathrm{Var}\left[\hat{p}_n\right], \\
&\geq tr(I(\theta)^{-1}) + \mathrm{Var}\left[1 - \sum_{i=1}^{n-1} \hat{p}_i\right], \\
&= tr(I(\theta)^{-1}) + \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \mathrm{Cov}(\hat{p}_i, \hat{p}_j), \\
&\geq tr(I(\theta)^{-1}) + \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} I(\theta)^{-1}(i,j).
\end{aligned}$$

□

D. Orderwise Achievability

In section III-C, we showed that $\epsilon(t) = \Omega\left(\frac{1}{t}\right)$. We now show that this lower bound is achievable if $rank(A) = n$.

Theorem 5 (Order-wise Achievability)**.** *It is possible to achieve estimation error of $O\left(\frac{1}{n}\right)$ if $rank(A) = n$.*

Proof of Theorem 5. In order to show achievability we divide the t pulls equally across all K arms; i.e., each arm is pulled $\frac{t}{K}$ times. For each $k = 1, \ldots, K$ and $i = 1, \ldots, m_k$ let $\hat{p}_{o_{i,k}}(t) = \frac{n_{o_{i,k}}}{t/K}$ be the estimate for $p_{o_{i,k}}$. From these estimates, we can generate estimates $\{\hat{p}_1(t), \ldots, \hat{p}_n(t)\}$ by solving the system of equations described by (2). More precisely, with $\hat{Y} := [\hat{p}_1(t), \hat{p}_2(t), \ldots, \hat{p}_n(t)]^\intercal$ and $\hat{P}_o = [\hat{p}_{o_{1,1}}(t), \ldots, \hat{p}_{o_{m_1,1}}(t), \ldots \hat{p}_{o_{m_K,K}}(t)]^\intercal$, we can solve

$$A\hat{Y} = \hat{P}_o. \quad (6)$$

First, we show that the estimates $\hat{p}(t)$ are unbiased. Let Y be the list of true probabilities $[p_1, p_2, \ldots, p_n]^\intercal$ and P_o be the list of true probabilities of observations, i.e., $P_o = [p_{o_{1,1}}(t), \ldots, p_{o_{m_1,1}}(t), \ldots p_{o_{m_K,K}}(t)]^\intercal$. The solution of (6) is given by $\hat{Y} = A^+\hat{P}_o$, where A^+ is the Moore-Penrose inverse of the matrix A. Thus, we get

$$\mathbb{E}\left[\hat{Y}\right] = \mathbb{E}\left[A^+\hat{P}_o\right] = A^+\mathbb{E}\left[\hat{P}_o\right] = A^+P_o,$$

upon using the fact that the estimates $\hat{p}_{o_{i,k}} = \frac{n_{o_{i,k}}}{t/K}$ are unbiased. The desired result $\mathbb{E}\left[\hat{Y}\right] = Y$ is now established as we note that $A^+P_o = Y$ in view of (2).

Next, we derive a bound on the estimation error $\varepsilon(t)$. We remark that our estimates are of the form $\hat{p}_i = A^+(i)\hat{P}_o$ where $A^+(i)$ is the i^{th} row of the Moore-Penrose inverse of A. It is also easy to see that (e.g., by Chebshev inequality) the variance of each empirical estimator $\hat{p}_{o_{i,k}} = \frac{n_{o_{i,k}}}{t/K}$ is $O\left(\frac{K}{t}\right)$. With $m = m_1 + \ldots + m_K$ denotes the (finite) number of rows in A, we then get

$$\begin{aligned}
\varepsilon(t) &= \sum_{i=1}^{n} \mathrm{Var}\left[\hat{p}_i(t)\right] \\
&= \sum_{i=1}^{n} \mathrm{Var}\left[\sum_{j=1}^{m} A^+(i,j)\hat{P}_o(j)\right] \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{m} \left(A^+(i,j)\right)^2 \mathrm{Var}\left[\hat{P}_o(j)\right] \\
&= O\left(\frac{1}{t}\right)
\end{aligned}$$

where the inquality follows from the fact that elements in $\hat{P}_o$ are negatively correlated since $\sum_{i=1}^{n} \hat{p}_{o_{i,k}} = 1$ for each $k = 1, \ldots, K$. □

IV. Proposed Algorithm

The design of an algorithm to minimize the estimation error can be divided into two parts: 1) producing the estimate of the distribution $\hat{p}_X(t)$ based on the samples observed till step t, and 2) deciding which arm to pull at each time t. Algorithm 1 describes our proposed algorithm. In Section IV-A and Section IV-B, we describe the two parts of our algorithm in detail.

A. Combining estimates from observations

We present a method to estimate $\{p_1, \ldots, p_n\}$ given $t_1, \ldots, t_K$, i.e., the number of times arms $g_1, \ldots, g_k$ are pulled until time t, respectively.

It is known that the Maximum Likelihood Estimate $\hat{\theta}$ of a parameter θ behaves as $N(\theta, I(\theta)^{-1})$ asymptotically, where $I(\theta)$ is the Fisher Information matrix; here $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. This means that MLE estimator is asymptotically consistent and it achieves the Cramer-Rao lower bound. This motivates us to use the maximum likelihood estimation for predicting $\hat{p}_1(t), \hat{p}_2(t), \ldots, \hat{p}_n(t)$ given samples observed until step t.

Recall that we defined $n_{o_{i,k}}$ as the number of times i^{th} output from arm k, i.e., $o_{i,k}$, is observed. Let $\hat{p}_{o_{i,k}}(t)$ be the

Algorithm 1 Sequential Distribution Estimation

Input: $\{x_1, x_2, \ldots, x_n\}$, Functions $\{g_1, g_2 \ldots g_K\}$ where $g_i : \{x_1, x_2, \ldots, x_n\} \to \mathbb{R}$. Total number of rounds, T.
Initialize: $n_{o_{i,k}} = 1, \forall i, k.\ \hat{p}_i(0) = \frac{1}{n}, \forall i$.
for $t = 1 : T$ **do**
  $c_t = \arg\min_k \tilde{V}(c_t | c_{1:t-1}, \hat{p}(t-1))$
  Pull arm c_t, observe output y_t
  if $y_t = o_{i,k}$ **then**
    $n_{o_{i,k}} = n_{o_{i,k}} + 1$
  end if
  Obtain estimates $\hat{p}_i(t)$ by obtaining fixed point solution of the set of equations described by

$$\hat{p}_j(t) = \frac{1}{t} \sum_{k=1}^{K} \sum_{i=1}^{m_k} n_{o_{i,k}} \frac{A_k(i,j)\hat{p}_j(t)}{\hat{p}_{o_{i,k}}(t)}, \quad j = 1, 2, \ldots, n.$$

end for

probability of observing output $o_{i,k}$ under the probability distribution $\hat{p}(t) = \{\hat{p}_1(t), \hat{p}_2(t), \ldots, \hat{p}_n(t)\}$. The log likelihood of $\mathcal{D}_t$ with respect to the probability distribution $\hat{p}(t)$ is given by

$$L(\mathcal{D}_t; \hat{p}(t)) = \sum_{k=1}^{K} \sum_{i=1}^{m_k} n_{o_{i,k}} \log(\hat{p}_{o_{i,k}}(t)). \tag{7}$$

where, $\hat{p}_{o_{i,k}} = \sum_{z=1}^{n} A_k(i,z)\hat{p}_z$.

In order to obtain the maximum likelihood estimate of $\hat{p}(t)$, we take the derivative of $L(\mathcal{D}_t; \hat{p}(t))$ and equate it to zero under the constraint $\sum_{i=1}^{n} \hat{p}_i(t) = 1$. This provides us a set of equations described by

$$\hat{p}_j(t) = \frac{1}{t} \sum_{k=1}^{K} \sum_{i=1}^{m_k} n_{o_{i,k}} \frac{A_k(i,j)\hat{p}_j(t)}{\hat{p}_{o_{i,k}}(t)}, \quad j = 1, 2, \ldots, n. \tag{8}$$

Observe that these set of equations are in the form of $x = f(x)$ and thus can be solved numerically by finding a fixed point. Since the log likelihood function is concave in $\hat{p}(t)$, the solution from the set of equations described above maximizes the log likelihood function.

B. Deciding which arm to pull

The previous section describes a method to generate estimates given the observations $\mathcal{D}_t$. In this section, we look at the task of deciding which arm to pull in each round, i.e., choosing c_t, where c_t is a decision variable indicating the chosen arm in round t. Formally, $c_t = k$, if arm k is pulled in round t.

If we had an analytic expression for estimation error $\varepsilon(t)$ at each round t, we could find the arm that minimizes the estimation error. However, in absence of an invertible arm, it is hard to obtain an analytic expression of ε, due to which we resort to a heuristic approach.

We define $E_i(t)$ as the "effective" number of samples observed for x_i until time t. In particular, we let

$$E_i(t) \triangleq \sum_{k=1}^{K} q_{k,i} t_k, \tag{9}$$

where $q_{k,i}$ is a parameter that measures the *quality* of arm k in estimating the probability of x_i, and t_k is the number of times arm k is pulled until time t.

In what follows, the quality metric $q_{k,i}$ is defined to achieve two fundamental goals. First, we would like to ensure that enough arms are sampled to achieve *asymptotically consistent estimation* if $\text{rank}(A) = n$. Secondly, we would like to ensure that an arm is never pulled if it is a subset of another arm. With these in mind, in our algorithm we choose $q_{k,j}$ as follows,

$$q_{k,j} = \begin{cases} \sum_{i=1}^{m_k} \frac{A_k(i,j)}{\sum_{\ell=1}^{n} A(i,\ell)}, & \text{if } k = \arg\max_r \sum_{i=1}^{m_r} \frac{A_r(i,j)}{\sum_{\ell=1}^{n} A(i,\ell)} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Observe that for a single invertible arm, $E_i(t) = t$ for all i since $q_{k,j} = 1, \forall j$. The intuition behind this choice of $q_{k,j}$ is to ensure that the most *informative* arm corresponding to each x_i is pulled. For a Bernoulli random variable with parameter p, the variance in estimating p from n samples is given by $\frac{p(1-p)}{n}$. This expression motivates us to use the method described below for deciding which arm to pull.

Let $c_{1:t} = [c_1, c_2, \ldots c_t]$ be the vector of indices of the arms pulled until time t. Given $c_{1:t}$, we can compute $t_1, t_2, \ldots, t_K$. Our decision in round $t+1$ will be based on comparing the value of the function

$$\tilde{V}(c_{t+1} = k | c_{1:t}, \hat{p}(t)) \triangleq \sum_{i=1}^{n} \frac{\hat{p}_i(t)(1 - \hat{p}_i(t))}{E_i(t) + q_{k,i}} \tag{11}$$

across all arms, which is a heuristic estimate of the error if $c_{t+1} = k$ (i.e., if arm k is picked next), given $c_{1:t}$ and $\hat{p}(t)$. In particular, at each round we choose the arm that minimizes $\tilde{V}$, i.e.,

$$c^*_{t+1} = \arg\min_k \tilde{V}(c_{t+1} = k | c_{1:t}, \hat{p}(t)), \tag{12}$$

with ties broken uniformly at random.

Lemma 1. *The proposed algorithm never picks an arm that is a subset of another arm.*

Proof. Suppose arm r is a subset of arm s. By (10), arm s has higher quality than arm r, that is, $q_{s,i} \geq q_{r,i}$ for all x_i. Thus, given $E_i(t)$ and $\hat{p}(t)$ in (11), picking arm s instead of arm r will always result in a smaller value of $\tilde{V}$. Hence, the proposed algorithm will always choose arm s over arm r. □

From Lemma 1, it follows that when there is an invertible arm, the method described above always picks it. Notice that our algorithm is following the principles mentioned in Theorem 2 and Corollary 1 to minimize the estimation error.

C. Simulations

In this section, we demonstrate the performance of our algorithm under different scenarios. We compare the estimation error of our algorithm with the Cramer Rao lower bound evaluated in Section III-C. Recall that Cramer Rao bound gives a lower bound on the estimation error given the choice of $\{t_1, t_2, \ldots, t_K\}$. To evaluate the lower bound after a total of T time slots, we find the Cramer Rao bound for all

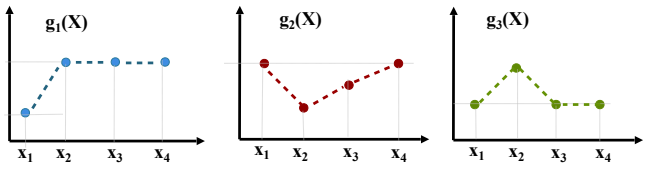

Fig. 4: *Example of a set of functions* $\{g_1, g_2, \ldots, g_K\}$, *where arm* g_3 *is a subset of arm* g_2. *Probability distribution* $\{p_1, p_2, p_3, p_4\}$ *is* $[0.1, 0.2, 0.3, 0.4]$.

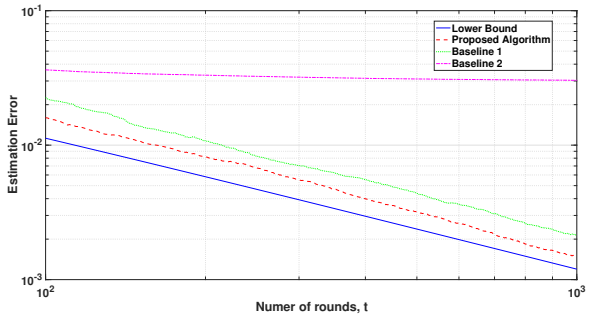

Fig. 5: *Comparison of our policy against the described baseline algorithms for the example in Fig. 4*

combinations of $\{\alpha_1 T, \alpha_2 T, \ldots, \alpha_K T\}$ where $\sum_{i=1}^{K} \alpha_i = 1$, and take the minimum over all such combinations. We iterate $\{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ for all possible values between 0 to 1 with a precision of 0.01. Note that the existence of an algorithm that achieves the Cramer Rao lower bound is not guaranteed.

For comparison purposes, in simulations, we also include the estimation error of two baseline algorithms. First baseline algortihm (Baseline 1) selects arms in a round robin manner, and produces estimate as described in Section IV-A. Baseline 2 algorithm selects arms as described in Section IV-B and produces estimate $\hat{p}_j(t) = \frac{\tilde{t}_j}{t}$, where

$$\tilde{t}_j = \sum_{k=1}^{K} \sum_{i=1}^{m_k} n_{o_{i,k}} \frac{A_k(i,j)}{\sum_{j=1}^{n} A(i,j)}.$$

Informally, Baseline 2 algorithm keeps a pseudo-count of the number of occurrences of each x_i. When the output could have been generated from multiple x_is, it increases pseudo count of all such x_is by an equal amount such that total increase in pseudo count is 1. Fig. 5 shows that the performance of our proposed algorithm is superior to that of the two considered baseline algorithms.

V. Concluding Remarks

We consider the problem of learning the distribution p_X of a hidden random variable X, using indirect samples from the functions $g_1(X)$, $g_2(X)$, $\ldots g_K(X)$, referred to as *arms*. The samples are obtained in a multi-arm bandit fashion, by choosing one of the K arms in each time slot. Several applications where we wish to infer properties of a hidden random phenomenon using indirect or imprecise observations fit into our framework. We determine conditions for asymptotically consistent estimation of p_X and obtain lower bounds on the estimation error. Using insights from this analysis, we propose an algorithm to choose arms and combine their samples. This algorithm eliminates redundant arms and gives lower estimation error than some intuitive baseline algorithms.

Ongoing work includes obtaining stronger guarantees on the performance of our algorithm. A generalization of the problem is to consider the sample $g_k(X)$ as the reward for pulling arm k, and design an algorithm to maximize the total expected reward. Analysis of such algorithms would require regret analysis, similar to multi-arm bandit problems.

Acknowledgements

This work was supported in part by the Department of Electrical and Computer Engineering at Carnegie Mellon University and by the National Science Foundation through grant CCF #1617934.

References

[1] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, pp. 1134–1142, Nov. 1984.

[2] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie, "On the learnability of discrete distributions," in *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, pp. 273–282, 1994.

[3] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Conference on Learning Theory*, pp. 1066–1100, 2015.

[4] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Transactions on Information Theory*, vol. 61, pp. 6343–6354, Nov 2015.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[6] S. Bubeck, N. Cesa-Bianchi, *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[7] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[8] L. Zhou, "A survey on contextual multi-armed bandits," *CoRR*, vol. abs/1508.03326, 2015.

[9] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory*, pp. 39–1, 2012.

[10] P. Sakulkar and B. Krishnamachari, "Stochastic contextual bandits with known reward functions," *CoRR*, vol. abs/1605.00176, 2016.

[11] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, pp. 81–91, 1945.

[12] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton University Press, 1946.