Contextual Bandits with Stochastic Experts

Rajat Sen
The University of Texas
at Austin

Karthikeyan Shanmugam
IBM Research,
Thomas J. Watson Center

Sanjay Shakkottai
The University of Texas
at Austin

Abstract

We consider the problem of contextual bandits with stochastic experts, which is a variation of the traditional stochastic contextual bandit with experts problem. In our problem setting, we assume access to a class of *stochastic experts*, where each expert is a conditional distribution over the arms given a context. We propose upper-confidence bound (UCB) algorithms for this problem, which employ two different importance sampling based estimators for the mean reward for each expert. Both these estimators leverage *information leakage* among the experts, thus using samples collected under all the experts to estimate the mean reward of any given expert. This leads to *instance dependent* regret bounds of $\mathcal{O}\left(\lambda(\boldsymbol{\mu})\mathcal{M}\log T/\Delta\right)$, where $\lambda(\boldsymbol{\mu})$ is a term that depends on the mean rewards of the experts, Δ is the smallest gap between the mean reward of the optimal expert and the rest, and $\mathcal{M}$ quantifies the information leakage among the experts. We show that under some assumptions $\lambda(\boldsymbol{\mu})$ is typically $\mathcal{O}(\log N)$. We implement our algorithm with stochastic experts generated from cost-sensitive classification oracles and show superior empirical performance on real-world datasets, when compared to other state of the art contextual bandit algorithms.

1 Introduction

Modern machine learning applications like recommendation engines [24, 11, 25], computational advertising [28, 10], A/B testing in medicine [29, 30] are inherently online. In these settings the task is to take sequential decisions that are not only profitable but also enable the system to learn better in future. For instance in a computational advertising system, the task is to sequentially place advertisements on users' webpages with the dual objective of learning the preferences of the users and increasing the click-through rate on the fly. A key attribute of these systems is the well-known *exploration* (searching the space of possible decisions for better learning) and *exploitation* (taking decisions that are more profitable) trade-off. A principled method to capture this trade-off is the study of multi-armed bandit problems [12].

K-armed stochastic bandit problems have been studied for several decades. These are formulated as a sequential process, where at each time step any one of the K-arms can be selected. Upon selection of the k-th arm, the arm returns a stochastic reward with an expected reward of μ_k. Starting from the work of [21], a major focus has been on *regret*, which is the difference in the total reward that is accumulated from the *genie* optimal policy (one that always selects the arm with the maximum expected reward) from that of the chosen online policy. The current state-of-art algorithms achieve a regret of $O((K/\Delta)\log T)$ [12, 7, 4, 5], which is order-wise optimal [21]. Here, Δ corresponds to the gap in expected reward between the best arm and the next best one.

Additional side information can be incorporated in this setting through the framework of contextual bandits. In the stochastic setting, it is assumed that at each time-step nature draws $(x, r_1, ..., r_K)$ from a fixed but unknown distribution. Here, $x \in \mathcal{X}$ represents the context vector, while $r_1, ..., r_K$ are the rewards of the K-arms [22]. The context x is revealed to the policy-designer, after which she decides to choose an arm $a \in \{1, 2, ..., K\}$. Then, the reward r_a is revealed to the policy-designer. In the computational advertising example, the context can be thought of as the browsing history, age, gender etc. of an user arriving in the system, while $r_1, ..., r_K$ are generated according to the probability of the user clicking on each of the K advertisements. The task here is to learn a *good*

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

mapping from the space of contexts $\mathcal{X}$ to the space of arms $[K] = \{1, 2, ..., K\}$ such that when the decisions are taken according to that mapping, the mean reward observed is high.

A popular model in the stochastic contextual bandits literature is the *experts* setting [3, 18, 22]. The task is to compete against the best *expert* in a class of experts $\Pi = \{\pi_1, ..., \pi_N\}$, where each expert $\pi \in \Pi$ is a function mapping $\mathcal{X} \to [K]$. The mean reward of an expert π is defined as $\mathbb{E}\left[r_{\pi(X)}\right]$, where X is the random variable denoting the context and the expectation is taken over the unknown distribution over $(x, r_1, ..., r_K)$. The best expert is naturally defined as the expert with the highest mean reward. The expected difference in rewards of a genie policy that always chooses the best expert and the online algorithm employed by the policy-designer is defined as the regret. This problem has been well-studied in the literature, where a popular approach is to reduce the contextual bandit problem to supervised learning techniques through argmin-oracles [8]. This leads to powerful algorithms with instance-independent regret bounds of $\mathcal{O}\left(\sqrt{KT\text{polylog}(N)}\right)$ at time T [3, 18].

In practice the class of experts are generated online by training cost-sensitive classification oracles [3, 18]. Once trained, the resulting classifiers/oracles can provide reliable confidence scores given a new context, especially if they are well-calibrated [17]. These confidence scores effectively are a K-dimensional probability vector, where the k^{th} entry is the probability of the classifier/oracle choosing the k^{th} arm as the best, given a context. Motivated by this observation, we propose a variation of the traditional experts setting, which we term contextual bandits with *stochastic experts*. We assume access to a class of *stochastic experts* $\Pi = \{\pi_1, ..., \pi_N\}$, which are *not deterministic*. Instead, each expert $\pi \in \Pi$, is a conditional probability distribution over the arms given a context. For an expert $\pi \in \Pi$ the conditional distribution is denoted by $\pi_{V|X}(v|x)$ where $V \in [K]$ is the random variable denoting the arm chosen and X is the context. An additional benefit is that this setting allows us to derive regret bounds in terms of *closeness* of these soft experts quantified by divergence measures, rather than in terms of the total number of arms K.

As before, the task is to compete against the expert in the class with the highest mean reward. The expected reward of a stochastic expert π is defined as $\mathbb{E}_{X,V\sim\pi(V|X)}\left[r_V\right]$, i.e the mean reward observed when the arm is drawn from the conditional distribution $\pi(V|X)$. We propose upper-confidence (UCB) style algorithms for the contextual bandits with stochastic experts problem, that employ two importance sampling based estimators for the mean rewards under various experts. We prove *instance-dependent* regret guarantees for our algorithms. The main contributions of this paper are listed in the next section.

1.1 Main Contributions

The contributions of this paper are three-fold:

(i) **(Importance Sampling based Estimators):** The key components in our approach are two importance sampling based estimators for the mean rewards under all the experts. Both these estimators are based on the observation that samples collected under one expert can be reweighted by likelihood/importance ratios and averaged to provide an estimate for the mean reward under another expert. This sharing of information is termed as *information leakage* and has been utilized before under various settings [23, 27, 10]. The first estimator that we use is an adaptive variant of the well-known clipping technique, which was proposed in [27]. The estimator is presented in Eq. (3). However, we carefully adapt the clipping threshold in an online manner, in order to achieve regret guarantees.

We also propose an importance sampling variant of the classical median of means estimator (see [26, 13]). This estimator is also designed to utilize the samples collected under all experts together to estimate the mean reward under any given expert. We define the estimator in Eq. (6). To the best of our knowledge, importance sampling has not been used in conjunction with the median of means technique in the literature before. We provide novel confidence guarantees for this estimator which depends on chi-square divergences between the conditional distributions under the various experts. This may be of independent interest.

(ii) **(Instance Dependent Regret Bounds):** We propose the contextual bandits with stochastic experts problem. We design two UCB based algorithms for this problem, based on the two importance sampling based estimators mentioned above. We show that utilizing the *information leakage* between the experts leads to regret guarantees that scale sub-linearly in N, the number of experts. The information leakage between any two experts in the first estimator is governed by a pairwise log-divergence measure (Def. 2). For the second estimator, chi-square divergences (Def. 3) characterize the leakage.

We show that the regret of our UCB algorithm based on these two estimators scales as [1]: $\mathcal{O}\left(\frac{\lambda(\boldsymbol{\mu})\mathcal{M}}{\Delta}\log T\right)$.

[1] Tighter regret bounds are derived in Theorems 1 and 2. Here, we only mention the Corollaries of our approach, that are easy to state.

Here, $\mathcal{M}$ is related to the largest pairwise divergence values under the two divergence measures used. Δ is the gap between the mean rewards of the optimal expert and the second best. $\lambda(\boldsymbol{\mu})$ is a parameter that only depends on the gaps between mean rewards of the optimum experts and various sub-optimal ones. It is a normalized sum of difference in squares of the gaps of adjacent sub-optimal experts ordered by their gaps. Under the assumption that the suboptimal gaps (except that of the second best arm) are uniformly distributed in a bounded interval, we can show that the parameter $\lambda(\boldsymbol{\mu})$ is $O(\log N)$ in expectation. We define this parameter explicitly in Section 6.

For the clipped estimator we show that $\mathcal{M} = M^2 \log^2(1/\Delta)$ where M is the largest pairwise log-divergence associated with the clipped estimator. For the median of means estimator, $\mathcal{M} = \sigma^2$ where σ^2 is the largest pairwise chi-squared divergence.

Naively treating each expert as an arm would lead to a regret scaling of $\mathcal{O}(N \log T/\Delta)$. However, this ignores information leakage. Existing instance-independent bounds for contextual bandits scale as $\sqrt{KT\text{poly}\log(N)}$ [3]. Our problem dependent bounds have a near optimal dependence on Δ and does not depend on K, the numbers of arms. However, it depends on the divergence measure associated with the information leakage in the problem (M or σ parameters). Besides our analysis, we empirically show that this divergence based approach rivals or performs better than very efficient heuristics for contextual bandits (like bagging etc.) on real-world data sets.

(*iii*) **(Empirical Validation):** We empirically validate our algorithm on three real world data-sets [20, 19, 1] against other state of the art contextual bandit algorithms [22, 3] implemented in Vowpal Wabbit [2]. In our implementation, we use online training of cost-sensitive classification oracles [8] to generate the class of stochastic experts. We show that our algorithms have better regret performance on these data-sets compared to the other algorithms.

2 Related Work

Contextual bandits has been studied in the literature for several decades, starting with the simple setting of discrete contexts [12], to linear contextual bandits [16] and finally the general experts setting [18, 3, 22, 6, 9]. In this work, we focus on the experts setting. Contextual bandits with experts was first studied in the adversarial setting, where there are algorithms with the optimal regret scaling $\mathcal{O}(\sqrt{KT\log N})$ [6].

In this paper, we are more interested in the stochastic version of the problem, where the context and the rewards of the arms are generated from an unknown but fixed distribution. The first strategies to be explored in this setting were explore-then-commit and epsilon-greedy [22] style strategies that achieve a regret scaling of $\mathcal{O}\left(\sqrt{K\log N}T^{2/3}\right)$ in the instance-independent case. Following this there have been several efforts to design adaptive algorithms that achieve a $\mathcal{O}(\sqrt{KT\text{polylog}(N)})$ instance-independent regret scaling. Notable among these are [18, 3]. These algorithms map the contextual bandit problem to supervised learning and assume access to cost-sensitive classification oracles. These algorithms have been heavily optimized in Vowpal Wabbit [2].

We study the contextual bandits with stochastic experts problem, where the experts are not deterministic functions mapping contexts to arms, but are conditional distributions over the arms given a context. We show that we can achieve instance-dependent regret guarantees for this problem, that can scale as $\mathcal{O}\left((\mathcal{M}\log N/\Delta)\log T\right)$ under some assumptions. Here, Δ is the gap between the mean reward of the best expert and the second best and $\mathcal{M}$ is a divergence term between the experts. Our algorithms are based on importance sampling based estimators which leverage information leakage among stochastic experts. We use an adaptive clipped importance sampling estimator for the mean rewards of the experts, that was introduced in [27]. In [27], the estimator was studied in a best-arm/pure explore setting, while we study a cumulative regret problem where we need to adjust the parameters of the estimator in an online manner. In addition, we introduce an importance sampling based median of means style estimator in this paper, that can leverage the information leakage among experts.

3 Problem Setting and Definitions

The general stochastic contextual bandit problem with K arms is defined as a sequential process for T discrete time-steps [22], where T is the time-horizon of interest. At each time $t \in \{1, 2, \cdots, T\}$ nature draws a vector $(x_t, r_1(t), ..., r_K(t))$ from an unknown but fixed probability distribution. Here, $r_i(t) \in [0, 1]$ is the reward of arm i. The context vector $x_t \in \mathcal{X}$ is revealed to the policy-designer, whose task is then to choose an arm out the K possibilities. Only the reward $r_{v(t)}(t)$ of the chosen arm $v(t)$, is then revealed to the policy-designer. We will use $r_{v(t)}$ in place of $r_{v(t)}(t)$ for notational convenience.

Stochastic Experts: We consider a class of stochastic experts $\Pi = \{\pi_1, \cdots, \pi_N\}$, where each π_i is a conditional probability distribution $\pi_{V|X}(v|x)$ where $V \in [K]$ is the random variable denoting the arm chosen and X is the context. We will use the shorthand

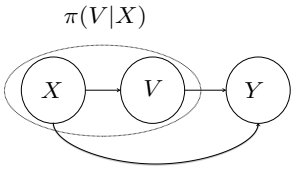

Figure 1: Bayesian Network denoting the joint distribution of the random variables at a given time-step, under our contextual bandit setting. X denotes the context, V denotes the chosen arm, while Y denotes the reward from the chosen arm that also depends on the context observed. The distribution of the reward given the chosen arm and the context, and the marginal of the context remain fixed over all time slots. However, the conditional distribution of the chosen arm given the context is dependent on the stochastic expert at that time-step.

$\pi_i(V|X)$ to denote the conditional distribution corresponding to expert i, for notational convenience. The observation model at each time step t is as follows: *(i)* A context x_t is observed. *(ii)* The policy-designer chooses a stochastic expert $\pi_{k(t)} \in \Pi$. An arm $v(t)$ is drawn from the probability distribution $\pi_{k(t)}(V|x_t)$, by the policy-designer. *(iv)* The stochastic reward $y_t = r_{v(t)}$ is revealed.

The joint distribution of the random variables X, V, Y denoting the context, arm chosen and reward observed respectively at time t, can be modeled by the Bayesian Network shown in Fig. 1. The joint distribution factorizes as follows, $p(x, v, y) = p(y|v, x)p(v|x)p(x)$ (1), where $p(y|v, x)$ (the reward distribution given the arm and the context), and $p(x)$ (marginal distribution of the context) is determined by the nature's distribution and are fixed for all time-steps $t = 1, 2, ..., T$. On the other hand $p(v|x)$ (distribution of the arm chosen given the context) depends on the expert selected at each round. At time t, $p(v|x) = \pi_{k(t)}(v|x)$ that is the conditional distribution encoded by the stochastic expert chosen at time t. Now we are at a position to define the objective of the problem.

Regret: The objective in our contextual bandit problem is to perform as well as the best expert in the class of experts. We will define $p_k(x, v, y) \triangleq p(y|v, x)\pi_k(v|x)p(x)$ as the distribution of the corresponding random variables when the expert chosen is $\pi_k \in \Pi$. The expected reward of expert k is now denoted by, $\mu_k = \mathbb{E}_{p_k(x,v,y)}[Y]$, where $\mathbb{E}_{p(.)}$ denotes expectation with respect to distribution $p(.)$. The best expert is given by $k^* = \arg\max_{k\in[N]} \mu_k$. The objective is to minimize the *regret* till time T, which is defined as $R(T) = \sum_{t=1}^{T} \left(\mu^* - \mu_{k(t)}\right)$, where $\mu^* = \mu_{k^*}$. Note that this is analogous to the regret definition for the deterministic *expert* setting [22]. Let us define $\Delta_k \triangleq \mu^* - \mu_k$ as the optimality gap in terms of expected reward, for expert k. Let $\boldsymbol{\mu} \triangleq \{\mu_1, ..., \mu_N\}$. Now we will define some divergence metrics that will be important in describing our algorithms and theoretical guarantees.

3.1 Divergence Metrics

In this section we will define some f-divergence metrics that will be important in analyzing our estimators. Similar divergence metrics were defined in [27] to analyze the clipped estimator in (3) in the context of a best arm identification problem. In addition to the divergence metric in [27], we will also define the chi-square divergence metric which will be useful in analyzing the median of means based estimator (6). First, we define conditional f-divergence.

Definition 1. *Let $f(\cdot)$ be a non-negative convex function such that $f(1) = 0$. For two joint distributions $p_{X,Y}(x, y)$ and $q_{X,Y}(x, y)$ (and the associated conditionals), the conditional f-divergence is given by:* $D_f(p_{X|Y}\|q_{X|Y}) = \mathbb{E}_{q_{X,Y}}\left[f\left(\frac{p_{X|Y}(X|Y)}{q_{X|Y}(X|Y)}\right)\right]$.

Recall that π_i is a conditional distribution of V given X. Thus, $D_f(\pi_i\|\pi_j)$ is the conditional f-divergence between the conditional distributions π_i and π_j. Note that in this definition the marginal distribution of X is the marginal of X given by nature's inherent distribution over the contexts. In this work we will be concerned with two specific f-divergence metrics that are defined as follows.

Definition 2. *(M_{ij} measure) [27] Consider the function $f_1(x) = x\exp(x-1) - 1$. We define the following log-divergence measure: $M_{ij} = 1+\log(1+D_{f_1}(\pi_i\|\pi_j))$, $\forall i, j \in [N]$.*

The M_{ij}-measures will be crucial in analyzing one of our estimators (clipped estimator) defined in Section 5.

Definition 3. *(σ_{ij} measure) $D_{f_2}(\pi_i\|\pi_j)$ is known as the chi-square divergence between the respective conditional distributions, where $f_2(x) = x^2 - 1$. Let $\sigma_{ij}^2 = 1 + D_{f_2}(\pi_i\|\pi_j)$.*

The σ_{ij}-measures are important in analyzing our second estimator (median of means) defined in Section 5.

4 A Meta-Algorithm

In this section, we propose a general upper-confidence bound (UCB) style strategy that utilizes the structure of the problem to converge to the best expert much faster than a naive UCB strategy that treats each expert as an arm of the bandit problem. One of the key observations in this framework is that rewards collected under one expert can give us valuable information about the mean under another expert, owing

to the Bayesian Network factorization of the joint distribution of X, V and Y. We propose two estimators for the mean rewards of different experts, that leverage this information leakage among experts, through importance sampling. These estimators are defined in Section 5. We propose a meta-algorithm (Algorithm 1) that is designed to use these estimators and the corresponding confidence intervals, to control regret.

Algorithm 1 D-UCB: Divergence based UCB for contextual bandits with stochastic experts

1: For time step $t = 1$, observe context x_1 and choose a random expert $\pi \in \Pi$. Play an arm drawn from the conditional distribution $\pi(V|x_1)$.
2: **for** $t = 2, ..., T$ **do**
3: Observe context x_t
4: Let $k(t) = \arg\max_k U_k(t-1) \triangleq \hat{\mu}_k(t-1) + s_k(t-1)$.
5: Select an arm $v(t)$ from the distribution $\pi_{k(t)}(V|x_t)$.
6: Observe the reward $Y(t)$.
7: **end for**

Here, $\hat{\mu}_k(t)$ denotes an estimate for the mean reward for expert k at time t, while $s_k(t)$ denotes the upper confidence bound for the corresponding estimator at time t. We propose two estimators that utilize all the samples observed under various experts to provide an estimate for the mean reward under expert k.

The first estimator denoted by $\hat{\mu}^c_k(t)$ (Section 5, Eq. (3)) is a clipped importance sampling estimator inspired by [27]. If this estimator is used, then $s_k(t)$ is set as in Equation. (4).

The second estimator denoted by $\hat{\mu}^m_k(t)$ (Section 5, Eq. (6)) is a median of means based importance sampling estimator. If this estimator is used, then $s_k(t)$ is set as in Equation. (7).

5 Estimators and Confidence Bounds

In this section we define two estimators for estimating the mean rewards under a given expert. Both these estimators can effectively leverage the information leakage between samples collected under various experts, through importance sampling. One key observation that enables us in doing so is the following equation,

$$\mu_k = \mathbb{E}_{p_j(x,v,y)}\left[Y\frac{\pi_k(V|X)}{\pi_j(V|X)}\right]. \tag{2}$$

This has been termed as *information leakage* and has been leveraged before in the literature [27, 23, 10] in best-arm identification settings. Recall that the subscript $p_j(x, v, y)$ denotes that the expectation is taken under the joint distribution in (1), where $p(v|x) = \pi_j(v|x)$ i.e. under the distribution imposed by expert π_j. However, even under this distribution we can technically estimate the mean reward under expert π_k. The above equation is the motivation behind our estimators. Now, we will introduce our first estimator.

Clipped Estimator: This estimator was introduced in [27] in the context of a pure exploration problem. Here, we analyze this estimator in a cumulative regret setting, where the parameters of the estimator need to be adjusted differently. Let $n_i(t)$ denote the number of times expert i has been invoked by Algorithm 1 till time t, for all $i \in [N]$. We define the fraction $\nu_i(t) \triangleq n_i(t)/t$. We will also define $\mathcal{T}_i(t)$ as the subset of time-steps among $\{1, .., t\}$, in which the expert i was selected. Let $\hat{\mu}^c_k(t)$ be the estimate of the mean reward of expert k from samples collected till time t. The estimator is given by,

$$\hat{\mu}^c_k(t) = \frac{1}{Z_k(t)}\sum_{j=1}^{N}\sum_{s\in\mathcal{T}_j(t)}\frac{1}{M_{kj}}Y_j(s)\frac{\pi_k(V_j(s)|X_j(s))}{\pi_j(V_j(s)|X_j(s))} \times \mathbb{1}\left\{\frac{\pi_k(V_j(s)|X_j(s))}{\pi_j(V_j(s)|X_j(s))} \le 2\log(2/\epsilon(t))M_{kj}\right\}. \tag{3}$$

Here, $A_j(s)$ is the value of the random variable A at time s drawn using expert j, where A can be the r.v's X,Y or V. We set $Z_k(t) = \sum_j n_j(t)/M_{kj}$. $\epsilon(t)$ is an adjustable term which controls the bias-variance trade-off for the estimator.

Intuition: The clipped estimator is a weighted average of the samples collected under different experts, where each sample is scaled by the importance ratio as suggested by (2). We also clip the importance ratios which are larger than a clipper level. This clipping introduces bias but decreases variance. The clipper level is carefully chosen to trade-off bias and variance. The clipper level values and the weights are dependent on the divergence terms M_{kj}'s. When the divergence M_{kj} is large, it means that the samples from expert j is not valuable for estimating the mean for expert k. Therefore, a weight of $1/M_{kj}$ is applied. Similarly, the clipper level is set at $2\log(2/\epsilon(t))M_{kj}$ to restrict the aditive bias to $\epsilon(t)$.

The upper confidence term in Algorithm 1 for the estimator $\hat{\mu}^c_k(t)$ is chosen as,

$$s^c_k(t) = \frac{3}{2}\beta(t) \tag{4}$$

at time t, where $\beta(t)$ is such that,

$\frac{\beta(t)}{\log(2/\beta(t))} = \frac{\sqrt{c_1 t\log t}}{Z_k(t)}$. We set $c_1 = 16$ in our analysis.

The upper confidence bound is derived using Lemma 1.

Lemma 1. *Consider the estimator in Eq. (3). We have the following confidence bound at time t,*

$$\mathbb{P}\left(\mu_k - \delta - \epsilon(t)/2 \leq \hat{\mu}_k^c(t) \leq \mu_k + \delta\right) \geq 1 - 2\exp\left(-\frac{\delta^2 t}{8(\log(2/\epsilon(t)))^2}\left(\frac{Z_k(t)}{t}\right)^2\right).$$

The lemma is implied by Theorem 4 in [27]. We include the proof in Section A in the appendix. The lemma shows that the clipped estimator can pool samples from all experts, in order to estimate the mean under expert k. The variance of the estimator depends on $Z_k(t)$, which depends on the log-divergences and number of times each expert has been invoked.

Median of Means Estimator: Now we will introduce our second estimator which is based on the well-known median of means technique of estimation. Median of means estimators are popular for statistical estimation when the underlying distributions are heavy-tailed [13]. The estimator for the mean under the k^{th} expert at time t is obtained through the following steps: (i) We divide the total samples into $l(t) = \lfloor c_2 \log(1/\delta(t)) \rfloor$ groups, such that the fraction of samples from each expert is preserved. We choose $c_2 = 8$ for our analysis. Let us index the groups as $r = 1, 2..., l(t)$. This means that there are at least $\lfloor n_i(t)/l(t) \rfloor$ samples from expert i in each group. (ii) We calculate the empirical mean of expert k from the samples in each group through importance sampling. (iii) The median of these means is our estimator.

Now we will setup some notation. Let $\mathcal{T}_j^{(r)} \subseteq \{1, 2, ..., t\}$ be the indices of the samples from expert j that lie in group r. Let $W_k(r,t) = \sum_i n_i(r,t)/\sigma_{ki}$, where $n_i(r,t)$ is the number of samples from expert i in group r. Let $n(r,t) = \sum_i n_i(r,t)$. Then the mean of expert k estimated from group r is given by,

$$\hat{\mu}_k^{(r)}(t) = \frac{1}{W_k(r,t)} \sum_{j=1}^{N} \sum_{s \in \mathcal{T}_j^{(r)}} \frac{1}{\sigma_{kj}} Y_j(s) \frac{\pi_k(V_j(s)|X(s))}{\pi_j(V_j(s)|X_j(s))}. \tag{5}$$

The median of means estimator for expert k is then given by,

$$\hat{\mu}_k^m(t) \triangleq \text{median}\left(\hat{\mu}_k^{(1)}(t), \cdots, \hat{\mu}_k^{(l(t))}(t)\right). \tag{6}$$

Intuition: The mean of every group is a weighted average of samples from each expert, rescaled by the importance ratios. This is similar to the clipped estimator in Eq. (3). However, here the importance ratios are not clipped at a particular level. In this estimator, the bias-variance trade-off is controlled by taking the median of means from $l(t)$ groups. The number of groups $l(t)$ needs to be carefully set in-order to control the bias-variance trade-off.

The upper confidence bound used in conjunction with this estimator at time t is given by,

$$s_k^m(t) = \frac{1}{W_k(t)} \sqrt{\frac{c_3 \log(1/\delta(t))}{t}} \tag{7}$$

where $W_k(t) = \min_{r \in [l(t)]} W_k(r,t)/n(r,t)$ and $\delta(t)$ is set as $1/t^2$ in our algorithm. This choice is inspired by the following lemma.

Lemma 2. *Let $\delta(t) \in (0,1)$. Then the estimator in (6) has the following confidence bound,*

$$\mathbb{P}\left(|\hat{\mu}_k^m(t) - \mu_k| \leq \frac{1}{W_k(t)} \sqrt{\frac{c_3 \log(1/\delta(t))}{t}}\right) \geq 1 - \delta(t). \tag{8}$$

We provide the proof of this lemma in Section B in the appendix. The constant c_3 is 64.

6 Theoretical Results

In this section, we provide *instance dependent* regret guarantees for Algorithm 1 for the two estimators proposed - a) The clipped estimator (3) and b) The median of means estimator (6). Let $\Delta = \min_{k \neq k^*} \Delta_k$ be the gap in the expected reward between the optimum expert and the second best. We define a parameter $\lambda(\boldsymbol{\mu})$, later in the section, that depends only on the gaps of the expected rewards of various experts from the optimal one.

For the Algorithm 1 that uses the clipped estimator, regret scales as $\mathcal{O}(\lambda(\boldsymbol{\mu}) M^2 \log^2(6/\Delta) \log T/\Delta)$. Similarly, for the case of the median of means estimator, regret scales as $\mathcal{O}(\lambda(\boldsymbol{\mu}) \sigma^2 \log T/\Delta)$. Here M is the maximum log-divergence and σ^2 is the maximum chi-square divergence between two experts, respectively.

When the gaps between the optimum expert and sub-optimal ones are distributed uniformly at random in $[\delta_2, 1]$ ($\delta_2 > 0$), we show that the $\lambda(\boldsymbol{\mu})$ parameter is at most $O(\log N)$ in expectation. In contrast, if the experts were used as separate arms, a naive application of UCB-1 [5] bounds would yield a regret scaling of $\mathcal{O}\left(\frac{N}{\Delta} \log T\right)$. This can be prohibitively large when the number of experts are large.

For ease of exposition of our results, let us re-index the experts using indices $\{(1), (2), ..., (N)\}$ such that $0 =$

$\Delta_{(1)} \leq \Delta_{(2)} \leq ... \leq \Delta_{(N)}$. The regret guarantees for our clipped estimator are provided under the following assumption.

Assumption 1. *Assume the log-divergence terms (2) are bounded for all $i, j \in [N]$. Let $M = \max_{i,j} M_{ij}$.*

Now we are at a position to present one of our main theorems that provides regret guarantees for Algorithm 1 using the estimator (3).

Theorem 1. *Suppose Assumption 1 holds. Then the regret of Algorithm 1 at time T using estimator (3), is bounded as follows:*

$$R(T) \leq \frac{C_1 M^2 \log^2(6/\Delta_{(N)}) \log T}{\Delta_{(N)}} + \frac{\pi^2}{3}\left(\sum_{i=2}^{N} \Delta_{(i)}\right) + \sum_{k=2}^{N-1} \frac{C_1 M^2 \log^2(6/\Delta_{(k)}) \log T}{\Delta_{(k)}} \left(1 - \frac{\gamma(\Delta_{(k)})}{\gamma(\Delta_{(k+1)})}\right)$$

Here, C_1 is an universal constant and $\gamma(x) = \frac{x^2}{\log^2(6/x)}$.

We defer the proof of Theorem 1 to Appendix A. We now present Theorem 2 that provides regret guarantees for Algorithm 1 using the estimator (6). The theorem holds under the following assumption.

Assumption 2. *Assume the chi-square terms (3) are bounded for all $i, j \in [N]$. Let $\sigma = \max_{i,j} \sigma_{ij}$.*

Theorem 2. *Suppose Assumption 2 holds. Then the regret of Algorithm 1 at time T using estimator (6), is bounded as follows:*

$$R(T) \leq \frac{C_2 \sigma^2 \log T}{\Delta_{(N)}} + \sum_{k=2}^{N-1} \frac{C_2 \sigma^2 \log T}{\Delta_{(k)}} \left(1 - \frac{\Delta^2_{(k)}}{\Delta^2_{(k+1)}}\right) + \frac{\pi^2}{3}\left(\sum_{i=2}^{N} \Delta_{(i)}\right)$$

Here, C_2 is an universal constant.

The proof of Theorem 2 has been deferred to Appendix B. Now, we will delve deeper into the instance dependent terms in Theorems 1 and 2. The proofs of Theorem 1 and 2 imply the following corollary.

Corollary 1. *Let $\lambda(\boldsymbol{\mu}) \triangleq 1 + \sum_{k=2}^{N-1} \left(1 - \frac{\Delta^2_{(k)}}{\Delta^2_{(k+1)}}\right)$. We have the following regret bounds:*

(i) For Algorithm 1 with estimator (3),

$$R(T) \leq \mathcal{O}\left(\frac{M^2 \log^2(6/\Delta_{(2)}) \log T}{\Delta_{(2)}} \min\left(\lambda(\boldsymbol{\mu}), \frac{1}{\Delta_{(2)}}\right)\right).$$

(ii) Similarly for Algorithm 1 with estimator (6),

$$R(T) \leq \mathcal{O}\left(\frac{\sigma^2 \log T}{\Delta_{(2)}} \min\left(\lambda(\boldsymbol{\mu}), \frac{1}{\Delta_{(2)}}\right)\right).$$

Corollary 1 leads us to our next result. In Corollary 2 we show that when the Δ gaps are uniformly distributed, then the $\lambda(\boldsymbol{\mu})$ is $O(\log N)$, in expectation.

Corollary 2. *Consider a generative model where $\Delta_{(3)} \leq ... \leq \Delta_{(N)}$ are the order statistics of $N-2$ random variables drawn i.i.d uniform over the interval $[\Delta_{(2)}, 1]$. Let p_Δ denote the measure over these Δ's. Then we have the following:*

(i) For Algorithm 1 with estimator (3),

$$\mathbb{E}_{p_\Delta}[R(T)] = \mathcal{O}\left(\frac{M^2 \log N \log^2(1/\Delta_{(2)}) \log T}{\Delta_{(2)}}\right).$$

(ii) For Algorithm 1 with estimator (6),

$$\mathbb{E}_{p_\Delta}[R(T)] = \mathcal{O}\left(\frac{\sigma^2 \log N \log T}{\Delta_{(2)}}\right).$$

Remark 1. *Note that our guarantees do not have any term containing K - the number of arms. This dependence is implicitly captured in the divergence terms among the experts. In fact when the number of arms K is very large, we expect our divergence based algorithms to perform comparatively better than other algorithms, whose guarantees explicitly depend on K. This phenomenon is observed in practice in our empirical validation on real world data-sets in Section 7. We also show empirically, that the term $\lambda(\boldsymbol{\mu})$ grows very slowly with the number of experts on real-world data-sets. This empirical result is included in Appendix D.*

7 Empirical Results

In this section, we will empirically test our algorithms on three real-world multi-class classification datasets, against other state of the art algorithms for contextual bandits with experts. Any multi-class classification dataset can be converted into a contextual bandit scenario, where the features are the contexts. At each time-step, the feature (context) of a sample point is revealed, following which the contextual bandit algorithm chooses one of the K classes, and the reward observed is 1 if its the correct class otherwise it is 0. This is bandit feedback as the correct class is never revealed, if not chosen. This method has been widely used to benchmark contextual bandit algorithms [9, 3], and is in fact implemented in Vowpal Wabbit [2].

Our algorithm is run in batches. At the starting of each batch, we add experts trained on prior data through cost-sensitive classification oracles [8] and also update the divergence terms between experts, which are estimated from data observed *so far*. During each batch, Algorithm 1 is deployed with the current set of experts. The pseudo-code for this procedure is provided in Algorithm 2. We use XgBoost [15] and Logistic Regression in scikit-learn [14] with calibration, as the base classifiers for our cost-sensitive oracles. Bootstrapping is used to generate different experts. At the starting of each batch 4 new experts are added. The constants are set as $c_1 = 1, c_2 = 4$ and

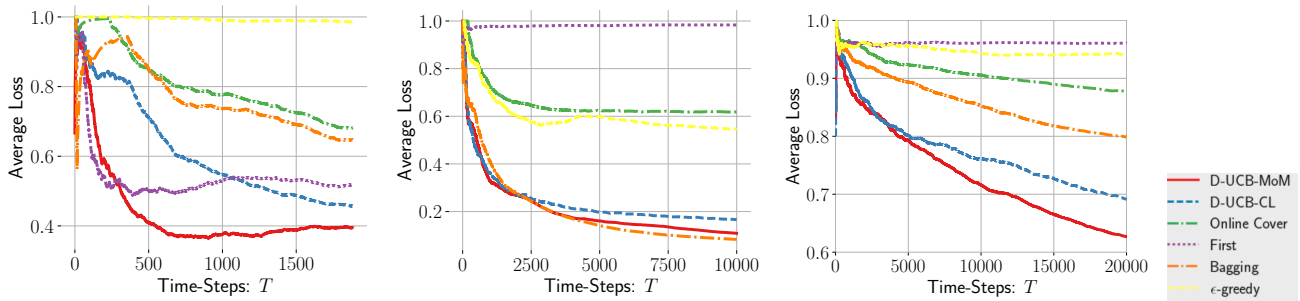

Figure 2: In all these plots, the average progressive validation loss/error till time T has been plotted as a function of time T. (a) Performance of the algorithms on the Drug Consumption dataset [19]. (b) Performance of the algorithms on the Stream Analytics dataset [1]. (c) Performance of the algorithms on the Letters dataset [20]. (d) Legend.

.

Algorithm 2 Batched D-UCB with cost-sensitive classification experts

1: Let $\Pi = \{\pi_1\}$, which is an expert that chooses arms randomly. For time steps $t = 1$ to $3K$, choose an arm sampled from expert π_1. $t = 3K + 1$.
2: Add experts to Π trained on observed data and update divergences.
3: **while** $t <= T$ **do**
4: **for** $s = t$ to $t + \mathcal{O}(\sqrt{t})$ **do**
5: Deploy Algorithm 1 with experts in Π.
6: **end for**
7: Let $t = t + \mathcal{O}(\sqrt{t})$. Add experts to Π trained on observed data and update divergences.
8: **end while**

$c_3 = 2$ in practice. All the settings are held fixed over all three data-sets, *without any parameter tuning*. We provide more details in Appendix D. In the appendix we also show that the gap dependent term in our theoretical bounds grows much slower compared to UCB-1 bounds (Fig. 3), as the number of experts increase in the stream analytics dataset [1]. An implementation of our algorithm can be found here.

We compare against Vopal Wabbit implementations of the following algorithms: (i) ϵ-greedy [22] - parameter set at '–epsilon 0.06'. (ii) First (Greedily selects best expert) - parameter set at '–first 100'. (iii) Online Cover [3] - parameter set at '–cover 5' (iv) Bagging (Simulates Thompson Sampling through bagged classifiers) - parameter set at '–bag 7'.

Drug Consumption Data: This dataset [19] is a part of UCI repository. It has data from 1885 respondents with 32 dimensional continuous features (contexts) and their history of drug use. There are 19 drugs under study (19 arms). For each entry, if the bandit algorithm selects the drug most recently used, the reward observed is 1, o.w. 0 reward is observed. The performance of the algorithms are shown in Fig. 2a. We see that D-UCB (Algorithm 2) with median of moments clearly performs the best in terms of average loss, followed by D-UCB with the clipped estimator. D-UCB-MoM converges to an average loss of 0.4 within 1885 samples.

Stream Analytics Data: This dataset [1] has been collected using the stream analytics client. It has 10000 samples with 100 dimensional mixed features (contexts). There are 10 classes (10 arms). For each entry, if the bandit algorithm selects the correct class, the reward observed is 1, o.w. 0 reward is observed. The performance of the algorithms are shown in Fig. 2b. In this data-set bagging performs the best closely followed by the two versions of D-UCB (Algorithm 2). Bagging is a strong competitor empirically, however this algorithm lacks theoretical guarantees. Bagging converges to an average loss of 8%, while D-UCB-MoM converges to an average loss of 10%.

Letters Data: This dataset [20] is a part of the UCI repository. It has 20000 samples of hand-written English letters, each with 17 hand-crafted visual features (contexts). There are 26 classes (26 arms) corresponding to 26 letters. For each entry, if the bandit algorithm selects the correct letter, the reward observed is 1, o.w. 0 reward is observed. The performance of the algorithms are shown in Fig. 2c. Both versions of D-UCB significantly outperform the others. The median of moments based version converges to an average loss of 0.62, while the clipped version converges to an average loss of 0.68.

Acknowledgment: This work is partially supported by NSF SaTC grant 1704778, ARO grant W911NF-17-1-0359, and the US DoT supported D-STOP Tier 1 University Transportation Center.

References

[1] Stream analytics dataset. https://stream-machinelearning.s3.amazonaws.com/vw-files/stream_labeled_data.vw. Accessed: 2017-10-10.

[2] Vowpal wabbit. https://github.com/JohnLangford/vowpal_wabbit. Accessed: 2017-10-10.

[3] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

[4] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[5] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[6] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[7] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[8] Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM, 2009.

[9] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

[10] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[11] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *International Conference on Neural Information Processing*, pages 324–331. Springer, 2012.

[12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[14] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

[15] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[16] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[17] Ira Cohen and Moises Goldszmidt. Properties and benefits of calibrated classifiers. In *PKDD*, volume 3202, pages 125–136. Springer, 2004.

[18] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

[19] Elaine Fehrman, Awaz K Muhammad, Evgeny M Mirkes, Vincent Egan, and Alexander N Gorban. The five factor model of personality and evaluation of drug consumption risk. In *Data Science*, pages 231–242. Springer, 2017.

[20] Peter W Frey and David J Slate. Letter recognition using holland-style adaptive classifiers. *Machine learning*, 6(2):161–182, 1991.

[21] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[22] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[23] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

[24] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. Scene: a scalable two-stage personalized news recommendation system. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 125–134. ACM, 2011.

[25] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[26] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*, 2017.

[27] Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3057–3066, International Convention Centre, Sydney, Australia, 2017. PMLR.

[28] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

[29] Cem Tekin, Onur Atan, and Mihaela Van Der Schaar. Discover the expert: Context-adaptive expert selection for medical diagnosis. *IEEE Transactions on Emerging Topics in Computing*, 3(2):220–234, 2015.

[30] Cem Tekin, Jinsung Yoon, and Mihaela van der Schaar. Adaptive ensemble learning with confidence bounds for personalized diagnosis. In *AAAI Workshop: Expanding the Boundaries of Health Informatics Using AI*, 2016.