Discover Relevant Sources : A Multi-Armed Bandit Approach

Onur Atan*, Mihaela van der Schaar†
Department of Electrical Engineering, University of California Los Angeles
Email: *oatan@ucla.edu, †mihaela@ee.ucla.edu

Abstract

Existing work on online learning for decision making takes the information available as a given and focuses solely on choosing the best actions *given* this information. Instead, in this paper, the decision maker needs to simultaneously learn both what decisions to make and what source(s) of information to consult/gather data from in order to inform its decisions such that its reward is maximized. We formalize this dual-learning and online decision making problem as a multi-armed bandit problem. If it were known in advance which sources were relevant for which decisions, the problem would be simple - but they are not. We propose algorithms that discover the relevant source(s) over time, while simultaneously learning what actions to take based on the information revealed by the selected source(s). Our algorithm resembles that of the well-known UCB algorithm but adds to it the online discovery of what specific sources are relevant to consult to inform specific decisions. We prove logarithmic regret bounds and also provide a matching lower bound on the number of times a wrong source is selected, which is achieved by RS-UCB for specific cases. The proposed algorithm can be applied in many applications including clinical decision assist systems for medical diagnosis, recommender systems, actionable intelligence, etc. where observing the complete information of a patient or a consumer or consulting all the available sources to gather intelligence is not feasible.

I. Introduction

Decision makers have access to vast amounts of different types of information. For instance, for medical diagnosis, a clinician or a clinical decision system (CDS) can have access to a wealth of information regarding the patient's health which can inform what treatments should be selected. Existing methods which learn online what decisions to make based on the available information assume that all this information is available at the time when the decision needs to be made: they utilize all the information to learn how to select actions that lead to high rewards. However, in many applications, it may be infeasible for the decision maker to access all the information. For instance, in the medical scenario outline before, a clinician or CDS cannot perform enormous numbers of tests on the patients to inform their treatment decisions. Moreover, it may be unnecessary to gather all the information since some of it may not be informative for specific decisions. (For instance, some information about the patient's health may be irrelevant to the decision of prescribing a specific treatment.)

In this paper, we consider the problem of simultaneously learning what sources of information to consider/what information to gather and what decisions to make on the basis of the obtained information to maximize the reward. We formalize this problem as a budgeted Multi Armed Bandit (MAB) problem, where the decision-maker can select at most R information sources. Learning in settings in which there is a hard constraint on the number of sources that the decision-maker can consult is known as *budgeted learning* [Cesa-Bianchi et al., 2011, Hazan and Koren, 2012].

The challenge in our setting is that the decision-maker does not observe additional information nor does it receive an additional reward informing it on how relevant specific sources would have been if they would have been selected to inform the current decision. The decision maker only observes the reward obtained from choosing a specific action given the selected sources and hence, it must learn which specific sources are relevant (informative) to which specific decisions (actions) solely from this feedback.

We provide an algorithm, which we refer to as Relevant Source Upper Confidence Bound (RS-UCB), which consists of two parts : (i) learning what sources to select to inform specific decisions (actions) and (ii) learning what decisions (actions) to take. The learning is performed based on standard bandit feedback.

The main contributions of the paper are :

- We first develop an algorithm (which we refer to as RS-UCB) that learns online which sources to observe and which actions to take, for the special case in which the same subset of information sources is relevant for all actions.
- We show that RS-UCB achieves logarithmic regret in time. Moreover, regret of RS-UCB scales with number of relevant sources R instead of number of all sources D.
- We derive a lower bound for the number of times that any policy will necessarily elect the wrong source (i.e. an irrelevant source). This lower bound is achieved by RS-UCB on specific cases.
- Finally, we provide a generalization of RS-UCB for the general case in which different sources are relevant for different actions. We prove that this generalized algorithm also enjoys logarithmic

regret.

II. Related Work

The work most related to ours is that on contextual bandits - a class of MABs where actions are selected on the basis of the available information, which is organized as ”contexts”. The traditional contextual bandit problems assume that an information vector arrives to a decision maker which utilizes all the available information to learn the actions that maximize the expected rewards. In contrast, in this work we learn what information sources should a decision maker consult to make an optimal decision, i.e. it should ”learn what to learn” to inform its decisions.

Contextual bandits have been studied extensively in the past 5 years [Chu et al., 2011, Slivkins, 2011, Lu et al., 2010, Dudik et al., 2011, Langford and Zhang, 2007, Tekin and Van Der Schaar, 2014]. For example, [Chu et al., 2011, Slivkins, 2011] study Lipschitz contextual bandits where a known similarity metric on expected rewards of the actions is assumed for different contexts. The lower bound on the regret for the Lipschitz contextual bandits is known to be exponential in the context dimension D [Chu et al., 2011]. Our work differs from the existing works on contextual bandits in multiple ways. First, to best of our knowledge, all past works assumed that all the information/contexts are observed when making decisions. Instead, we only observe the sources (contexts) that are selected. Second, we assume a relevance relation between the rewards of the arms and the context types and learn this relationship online. This enables us to achieve regret bounds which scales with number of relevant sources R instead of number of sources D. This is a huge performance improvement if $D \gg R$.

The work most related to ours is [Tekin and Van Der Schaar, 2014], which assumes an unknown relevance relation between the context types and actions. They study traditional contextual bandit problem with an unknown relevance relation between the source types and actions. In contrast, in our setting the decision maker cannot observe the entire context vector but it has only access to the sources (context types) that are selected. Hence, it also needs to learn which sources (context types) to select.

Another closely related line of work tackles the issue of budget limited learning [Cesa-Bianchi et al., 2011, Hazan and Koren, 2012, Szepesvari and Zolghadr, 2013], where the learner is restricted to access only a subset of the features from the set of training examples. Faced with this constraint, the learner's goal is to adaptively choose which features of the next training example it wants to observe, given the observations of the features of the past training examples. The goal is to train a linear predictor with a low expected loss given the distribution of the data instances. The problem is solved by generating adaptive versions of well-known linear predictors such as Online Gradient Descent [Zinkevich, 2003], EG [Kivinen and Warmuth, 1997] and Pegasos [Shalev-Shwartz et al., 2011]. The solution constructs an unbiased estimate of the feature vector or its gradient using random sampling techniques. Such random sampling techniques have been used for a long time to solve least-squares and low-rank matrix approximation problems involving very large matrices (see [Mahoney, 2011] and the references therein). Unlike the above line of the work, which aims to find out the best linear predictor given observed features, the goal of this paper is to identify: (i) relevant source(s) of information and (ii) the optimal action to be selected given the information provided by observing/consulting/measuring the relevant source(s).

III. Problem Formulation

A. Preliminaries

Let $\mathcal{X}$ denote the space of D-dimensional information vectors and $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ be an element in that space, with $x_d \in \mathcal{X}_d$ being a source-d information. Let $\mathcal{A} = \{1, 2, \ldots, K\}$ denote the set of actions. An instance of a problem is specified by an unknown distribution $\mathbb{D}$ over tuples $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathcal{X}$ is the information vector and $\boldsymbol{y} = (y_a)_{a \in \mathcal{A}} \in \mathcal{Y}^K$ is the vector of the rewards where $\mathcal{Y}$ is the reward space and y_a is the reward associated with action a. Throughout the paper, we assume that the information space is finite, i.e, $|\mathcal{X}_d| \leq M < \infty$ and rewards are bounded, i.e., $\mathcal{Y} = [0, 1]$. The first assumption is not restrictive since any bounded space can be discretized using techniques similar to the ones in [Slivkins, 2011]. The second assumption is standard in the multi-armed bandit literature.

Given a vector $\boldsymbol{x}$ and a set $\mathcal{S} \subseteq \mathcal{D}$, we write $\boldsymbol{x}_{\mathcal{S}}$ for the restriction of $\boldsymbol{x}$ to $\mathcal{S}$ (i.e.; a context vector containing components in $\mathcal{S}$). Let $\bar{y}_a^{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})$ be the marginal expected reward of action a when the information vector contains $\boldsymbol{x}_{\mathcal{S}}$, i.e., $\bar{y}_a^{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) = \mathbb{E}(y_a|\boldsymbol{x}_{\mathcal{S}})$ where the expectation is taken with respect to the distribution of $\mathbb{D}(y_a|\boldsymbol{x}_{\mathcal{S}})$. Let $\bar{y}_a(\boldsymbol{x})$ denote the expected reward of action a given full information vector $\boldsymbol{x}$. Similarly, let $\bar{y}_a$ be the marginal expected reward of action a, i.e., $\bar{y}_a = \mathbb{E}(y_a)$ where expectation is taken with respect to the distribution $\mathbb{D}(y_a)$.

The decision maker works in a discrete time setting $n = 1, 2, \ldots$, where the following events happen sequentially in each time slot : (i) the decision maker starts with no information (both the information vector $\boldsymbol{x}_n$ and reward vector $\boldsymbol{y}_n$ are latent) and selects the sources $\mathcal{O}_n \subset \mathcal{D}$ with $|\mathcal{O}_n| = R$ with R the observation constraint; ii) a partial information vector $\hat{\boldsymbol{x}}_n$ containing only the information from the selected sources is revealed to decision maker; (iii) the decision maker chooses an action a_n from $\mathcal{A}$; (iv) the reward associated with selected action $y_{a,n}$ is revealed to decision-maker. (We use $y_{a,n}$ for $y_{a_n,n}$ and $\hat{\boldsymbol{x}}_n$ for $\boldsymbol{x}_{\mathcal{O}_n,n}$ to improve readability.)

As mentioned previously, this model is different from contextual bandit problems [Chu et al., 2011, Slivkins, 2011] where all sources/contexts are observed at each time and can inform decisions. In this model, the decision maker needs to determine the source it wants to observe and has only access to sources that it selects. For instance, in medical informatics, $\mathcal{O}_n$ corresponds to the medical tests (specific blood tests, imaging tests etc.) which a patient needs to undertake, a_n is the treatment recommended to the patient (which is informed by the selected medical tests) and $y_{a,n}$ represents the effectiveness of the selected treatment.

B. Assumptions

We start by assuming that the rewards of actions only depend on the sources in $\mathcal{R} \subset \mathcal{D}$, which are referred to as the *set of relevant (informative) source types*. All the other sources are assumed to be irrelevant and hence, they are assumed to not convey any additional information about the reward. (In Section 6, we will relax this assumption and allow different contexts to be relevant to different actions.) The decision maker does not know which of these sources are relevant, i.e. which sources are in $\mathcal{R}$ and hence, it needs to learn this. The assumptions we make are formalized next.

Assumption 1. *i) There exists a minimal set $\mathcal{R} \subset \mathcal{D}$ with $|\mathcal{R}| = R$ (the observability constraint) such that for all $a \in \mathcal{A}$, $\boldsymbol{x} \in \mathcal{X}$ any sub-vector $\tilde{\boldsymbol{x}}$ of $\boldsymbol{x}$, we have*

$$\mathbb{E}(y_a|\boldsymbol{x}_{\mathcal{R}}, \tilde{\boldsymbol{x}}) = \mathbb{E}(y_a|\boldsymbol{x}_{\mathcal{R}}) \tag{1}$$

ii) The sources are statistically independent of each other, i.e., for all $\boldsymbol{x} \in \mathcal{X}$, $\mathbb{D}(\boldsymbol{x}) = \Pi_{d \in \mathcal{D}} \mathbb{D}(x_d)$.

The first part of the above assumption states that the expected reward of any action a depends only on the sources in the set $\mathcal{R}$; the second part states that the sources are statistically independent of each other. This is a limitation of our work. We discuss the challenges of source selection with bandit feedback for general information distributions in Section 8.

C. Problem Definition

Optimal actions can be computed based on the knowledge of $\mathbb{D}$. Let $a^*(\boldsymbol{x}_{\mathcal{R}}) = \arg\max_a \bar{y}_a^{\mathcal{R}}(\boldsymbol{x}_{\mathcal{R}})$ be the action that has the maximum expected reward conditional on the information $\boldsymbol{x}_{\mathcal{R}}$. Uncertainty about $\mathbb{D}$ and the set of relevant information sources $\mathcal{R}$ induces uncertainty about the true optimal actions $a^*(\boldsymbol{x}_{\mathcal{R}})$.

Remark 1. *Note that since $\bar{y}_a(\boldsymbol{x}) = \bar{y}_a^{\mathcal{R}}(\boldsymbol{x}_{\mathcal{R}})$ by assumption 1, we have $a^*(\boldsymbol{x}) = a^*(\boldsymbol{x}_{\mathcal{R}})$.*

Let $h_n = (\mathcal{O}_n, \hat{\boldsymbol{x}}_n, a_n, y_{a,n})$ be the observation tuple at the n^{th} instance and $\mathcal{H}_n = \{h_\tau\}_{\tau=1}^n$ be the history of observations up to the n^{th} instance. Let Ω_R denote the set of subsets of source set $\mathcal{D}$ with size R, i.e., $\Omega_R = \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{D}, |\mathcal{S}| \leq R\}$ and $\mathcal{X}_R$ be the set of all possible information vectors containing R sources, i.e., $\mathcal{X}_R = \{\boldsymbol{x}_\mathcal{S} : \mathcal{S} \in \Omega_R, \boldsymbol{x} \in \mathcal{X}\}$. A source selection policy π^{source} is a mapping from history up to the $(n-1)$th instance to Ω_R, i.e., $\pi^{\text{source}} : \mathcal{H}_{n-1} \rightarrow \Omega_R$. An action selection policy for information $\hat{\boldsymbol{x}}$ from set of sources $\mathcal{S} \subseteq \mathcal{D}$ is a mapping from history up to the $(n-1)^{\text{th}}$ instance and information $\hat{\boldsymbol{x}}$ to action set $\mathcal{A}$, i.e., $\pi^{\text{action}}(\hat{\boldsymbol{x}}) : \mathcal{H}_{n-1} \times \hat{\boldsymbol{x}} \rightarrow \mathcal{A}$. Let $\boldsymbol{\pi}^{\text{action}}$ be the all possible policies with history $\mathcal{H}_{n-1}$, i.e., $\boldsymbol{\pi}^{\text{action}} = \left[\pi^{\text{action}}(\hat{\boldsymbol{x}})\right]_{\hat{\boldsymbol{x}} \in \mathcal{X}_R}$. A policy $\boldsymbol{\pi}$ is the combination of source and action selection policies, i.e.,

$$\boldsymbol{\pi} = \left(\pi^{\text{source}}, \boldsymbol{\pi}^{\text{action}}\right).$$

Let $\boldsymbol{x}^n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{y}^n = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ and π_n denote the action (arm) selected at n^{th} instance by following policy π.[1] The N-period distribution-independent regret of policy π is the random variable,

$$\begin{aligned}&\text{Reg}_N^\pi(\boldsymbol{x}^N)\\ &\triangleq \mathbb{E}_{\boldsymbol{y}^N \sim \mathbb{D}(\boldsymbol{y}^N|\boldsymbol{x}^N)} \sum_{n=1}^{N} \left[\bar{y}^{\mathcal{R}}_{a^*(\boldsymbol{x})}(\boldsymbol{x}_{\mathcal{R},n}) - \bar{y}^{\mathcal{R}}_{\pi_n}(\boldsymbol{x}_{\mathcal{R},n})\right]\end{aligned}$$

where expectation is taken with respect to the distribution of the reward history. The regret defined above is a random variable because it depends on the information vector history $\boldsymbol{x}^N$. The distribution-dependent regret is

$$\overline{\text{Reg}}_N^\pi \triangleq \mathbb{E}_{(\boldsymbol{x}^N, \boldsymbol{y}^N) \sim \mathbb{D}^N} \sum_{n=1}^{N} \left[\bar{y}^{\mathcal{R}}_{a^*(\boldsymbol{x})}(\boldsymbol{x}_{\mathcal{R},n}) - \bar{y}^{\mathcal{R}}_{\pi_n}(\boldsymbol{x}_{\mathcal{R},n})\right]$$

The goal of this work is to learn an optimal policy that is able to minimize the distribution-dependent regret subject to source observability constraints, i.e.,

$$\begin{aligned}&\underset{\pi}{\text{minimize}} \quad \overline{\text{Reg}}_N^\pi\\ &\text{subject to} \quad |\mathcal{O}_n| \leq R, \; n = 1, \ldots, N.\end{aligned}$$

In the next section, we propose a method for solving this problem.

IV. Upper Confidence Bounds for Relevance Discovery

In this section, we provide a variant of UCB policy [Auer et al., 2002], which we call as RS-UCB. This policy works by associating an upper confidence index to each information source d and to each

[1] We drop the dependence of $\mathcal{H}_n$ and to the history for sake of brevity.

$T_a^d(x_d)$	Number of times action a is selected when source d with information x_d is observed, i.e., $\sum_{\tau=1}^n \mathbb{I}(d \in \mathcal{O}_\tau, x_{d,\tau} = x, a_\tau = a)$.
T_a	Number of times action a is selected, i.e., $\sum_{\tau=1}^n \mathbb{I}(a_\tau = a)$.
T^d	Number of times source d is observed, i.e., $\sum_{\tau=1}^n \mathbb{I}(d \in \mathcal{O}_\tau)$.
$T^d(x_d)$	Number of times source d with information x is observed, i.e., $\sum_{\tau=1}^n \mathbb{I}(d_\tau = d, x_{d,\tau} = x)$.
$\hat{y}_a^d(x_d)$	Reward estimate for action a when source d with information x_d is observed, i.e., $\frac{1}{T_a^d(x_d)} \sum_{\tau=1}^n y_{a,\tau} \mathbb{I}(d \in \mathcal{O}_\tau, x_{d,\tau} = x, a_\tau = a)$.
$\hat{y}_a^d$	Reward estimate for action a from source d, i.e., $\frac{1}{T^d} \sum_{x_d} T^d(x_d) \hat{y}_a^d(x_d)$.
$T_a^{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})$	Number of times action a is selected based on observing information $\boldsymbol{x}_{\mathcal{S}}$ from source $\mathcal{S}$, i.e., $\sum_{\tau=1}^n \mathbb{I}(\mathcal{O}_\tau = \mathcal{S}, \boldsymbol{x}_{\mathcal{S},n} = \boldsymbol{x}_{\mathcal{S}}, a_\tau = a)$.

TABLE I: Counter and Reward Estimate Definitions

action a.

In our context, the algorithm must accomplish two tasks: find the relevant sources and, given the information from the relevant sources, choose the best action. While the second task can be accomplished using UCB, the first task requires a new algorithm. Based on the history $\mathcal{H}_n$, Table I defines the counters and reward estimates which are used throughout the paper. Note that we drop n from the notations to improve readability.

Based on these counters and estimates, we define the following relevance metric

$$\hat{\text{Rel}}_d \triangleq \sum_{a \in \mathcal{A}} \sum_{x_d \in \mathcal{X}_d} \frac{T_a^d(x_d)}{T^d} |\hat{y}_a^d(x_d) - \hat{y}_a^d|. \tag{2}$$

The relevance metric estimates the difference in reward obtained if source d would have been observed.

```
for $n \geq 0$ do
  if $\mathcal{F}_n^1 \neq \emptyset$ then
    Randomly select at most $R$ from set $\mathcal{F}_n^1$
  else
    $\mathcal{O}_n = \arg\max_{\mathcal{S} \in \Omega_R} \sum_{d \in \mathcal{S}} \text{ind}_d$
  end if
  Observe $\hat{\boldsymbol{x}}_n$
  if $\mathcal{F}_n^2 \neq \emptyset$ then
    Select randomly from $\mathcal{F}_n^2(\hat{\boldsymbol{x}}_n)$
  else
    Estimate $\hat{y}_a(\hat{\boldsymbol{x}}_n)$
    $a_n = \arg\max_{a \in \mathcal{A}} \text{conf}_a(\hat{\boldsymbol{x}}_n)$.
  end if
  Observe $y_{a,n}$.
end for
```

Fig. 1: Pseudocode of RS-UCB

Let $\mathcal{F}_n^1 = \{d \in \mathcal{D} : T^d = 0\}$ be the set of sources that have not been selected until the nth instance. If this set is non-empty, RS-UCB randomly selects $\min(R, |\mathcal{F}_n^1|)$ sources from the set. Otherwise, RS-UCB assigns an index for each source based on their relevance metric and the confidence, i.e.,

$$\text{ind}_d \triangleq \hat{\text{Rel}}_d + \sqrt{\frac{8KM \ln n}{T^d}}.$$

Then, RS-UCB selects the sources with the highest indices, i.e.,

$$\mathcal{O}_n = \arg\max_{\mathcal{S} \in \Omega_R} \sum_{d \in \mathcal{S}} \text{ind}_d.$$

Having chosen sources $\mathcal{O}_n$, the partial information vector $\hat{\boldsymbol{x}}_n$ is observed. Then, RS-UCB constructs an estimate for the reward of every action a based on the observed information $\hat{\boldsymbol{x}}_n$ as

$$\hat{y}_a(\hat{\boldsymbol{x}}_n) = \frac{1}{T_a(\hat{\boldsymbol{x}}_n)} \sum_{\tau=1}^{n} y_{a,n} \mathbb{I}(\mathcal{O}_\tau = \mathcal{O}_n, \boldsymbol{x}_{\mathcal{O}_\tau,\tau} = \hat{\boldsymbol{x}}_n, a_\tau = a)$$

Above, we used the notation $T_a(\hat{\boldsymbol{x}}_n)$ and $\hat{y}_a(\hat{\boldsymbol{x}}_n)$ instead of $T_a^{\mathcal{O}_n}(\hat{\boldsymbol{x}}_n)$ and $\hat{y}_a^{\mathcal{O}_n}(\hat{\boldsymbol{x}}_n)$ to improve readability. Having constructed these estimates, the action is selected based on the UCB policy. Let

$$\mathcal{F}_n^2(\hat{\boldsymbol{x}}_n) = \{a \in \mathcal{A} : T_a(\hat{\boldsymbol{x}}_n) = 0\}$$

be the set of unselected actions for the observed information $\hat{\boldsymbol{x}}_n$. If this set is non-empty, RS-UCB randomly selects an action from this set. Otherwise, the action with the maximum confidence level is selected, where the confidence of an action is defined as

$$\text{conf}_a(\hat{\boldsymbol{x}}_n) \triangleq \hat{y}_a(\hat{\boldsymbol{x}}_n) + \sqrt{\frac{2 \ln n}{T_a(\hat{\boldsymbol{x}}_n)}} \tag{3}$$

The pseudo code of RS- UCB is given in Figure 1.

Note that the online source selection part of RS-UCB is different than UCB since the decision maker since the feedback is only the reward of the selected action. Hence, the relevance discovery uses solely the rewards of obtained for the selected actions. Therefore, RS-UCB involves a dual-learning procedure using the standard MAB feedback.

V. Analysis of RS-UCB

A. Preliminaries

In this section, we provide some definitions that are used for the regret analysis. The definitions for the relevance and suboptimality gaps play important roles in the theoretical analysis of RS-UCB; they are presented in Table II.

$\delta_a^d(x_d)$	relevance gap of information x_d from source d for action a, i.e., $\delta_a^d(x_d) = \lvert\bar{y}_a^d(x_d) - \bar{y}_a\rvert$
δ_{min}	minimal relevance gap of action a for source $d \in \mathcal{R}$, i.e., $\min_{a\in\mathcal{A}} \min_{d\in\mathcal{D}} \min_{x_d\in\mathcal{X}_d} \delta_a^d(x_d)$.
p_{min}	minimal information probability for $d \in \mathcal{R}$, i.e., $\min_{d\in\mathcal{D}} \min_{x_d\in\mathcal{X}_d} \mathbb{D}(x_d)$
$\underline{\delta}^d(x_d)$	minimal relevance gap for information x_d for $d \in \mathcal{R}$, i.e., $\min_a \delta_a^d(x_d)$.
$\underline{\delta}^d_{\text{exp}}$	expected minimal relevance gap for source d, i.e., $\mathbb{E}_{x_d\sim\mathbb{D}(x_d)}(\underline{\delta}^d(x_d))$.
$\underline{\delta}_{\text{exp}}$	expected minimal relevance gap, i.e., $\min_{d\in\mathcal{R}} \underline{\delta}^d_{\text{exp}}$.
$\Delta_a(\boldsymbol{x}_\mathcal{R})$	suboptimality gap of action a for information $\boldsymbol{x}_\mathcal{R}$, i.e., $\bar{y}^{\mathcal{R}}_{a^*(\boldsymbol{x})}(\boldsymbol{x}_\mathcal{R}) - \bar{y}^{\mathcal{R}}_a(\boldsymbol{x}_\mathcal{R})$.
Δ_{min}	minimal suboptimality gap, i.e., $\min_{a\in\mathcal{A}} \min_{\boldsymbol{x}_\mathcal{R}} \Delta_a(\boldsymbol{x}_\mathcal{R})$.

TABLE II: Definitions of suboptimality and relevance gaps

Lemma 1. *If $T_a^d(x_d) \to \infty$ as $n \to \infty$ for all a, d, x_d, then under Assumption 1, we have*
(i) $|\hat{y}_a^d(x_d) - \hat{y}_a^d| \xrightarrow{p} \delta_a^d(x_d)$,
ii) $\delta_a^d(x_d) = 0$ for all $x_d \in \mathcal{X}_d$ and $d \in \mathcal{D} \setminus \mathcal{R}$
where $\xrightarrow{p}$ denotes convergence in probability.

Proof. (Sketch) The first result follows from the continuous mapping theorem [Mann and Wald, 1943]. The second result can be shown using the tower property and the assumption 1. Detailed proofs of all results can be found in the Supplementary material. □

Lemma 1 presents technical results that guide the design of RS-UCB. Since $\delta_a^d(x_d) = 0$ for all irrelevant sources d, the relevance metric defined in (2) converges to 0 for all irrelevant sources. Therefore, the regret results will heavily depend on the relevance gap of the relevant sources. In the next section, we formally provide theoretical results.

B. Distribution-Independent Regret Analysis of RS-UCB

The following theorem bounds the regret independent of information source distribution.

Theorem 1. *(Distribution Independent Regret Bound) If $\delta_{\text{min}} > 0$, the regret of RS-UCB is bounded as follows :*

$$Reg_N^{\textit{RS-UCB}}(\boldsymbol{x}^N) \leq A_1 \ln N + A_2$$

with probability 1 where

$$A_1 = 8 \sum_{\boldsymbol{x}_\mathcal{R}} \sum_{a\in\mathcal{A}\setminus a^*(x)} \Delta_a^{-1}(\boldsymbol{x}_\mathcal{R}) + 32KMD(\delta_{\text{min}})^{-2}$$

and

$$A_2 = D + 2K(M+1)RD\pi^2/3 \\ + (1+\pi^2/3)\sum_{a\in\mathcal{A}}\sum_{\boldsymbol{x}_{\mathcal{R}}}\Delta_a(\boldsymbol{x}_{\mathcal{R}}).$$

Proof. (Sketch) The regret can be decomposed as

$$\text{Reg}_N^{\text{RS-UCB}} \leq \sum_{\boldsymbol{x}_{\mathcal{R}}}\Delta_a(\boldsymbol{x}_{\mathcal{R}})\mathbb{E}(T_a^{\mathcal{R}}(\boldsymbol{x}_{\mathcal{R}})) \\ + DRl + \sum_{n=1}^{N}\sum_{d\in\mathcal{D}\setminus\mathcal{R}}\mathbb{E}(\mathbb{I}(d\in\mathcal{O}_n, T^d \geq l))$$

where the first part can be bounded by using UCB-type bounds and second part can be divided into three sub events. By choosing $l = 32KM(\delta_{\min})^{-2}\ln n + 1$ for source d, we can show that

$$\mathbb{E}(\mathbb{I}(d\in\mathcal{O}_n, T^d \geq l)) \leq 4K(M+1)n^{-2}. \tag{4}$$

The final result can be achieved combining UCB-type bounds and this part. □

Theorem 1 shows a regret bound that is independent of the information vector distribution that holds for any information vector. The regret can be decomposed into two parts : selecting a wrong (irrelevant) source and selecting a suboptimal action (which is the same as in the UCB algorithm).

Corollary 1. *For the special case where $R = 1$, the regret is in the order of $O\left(DKM(\delta_{\min}^{-2} + \Delta_{\min}^{-1})\ln N\right)$.*

In special case where there is one relevant source, regret scales linearly with the number of sources D whereas the traditional contextual bandits scale exponentially with D. This leads to a significant performance difference if the total number of sources is large. Moreover, our algorithm achieves this regret using only partial information observation (only observing the selected sources). Since regret in Theorem 1 is the distribution independent, it depends quadratically on the minimal relevance gap $\delta_{\min}^{-1}$. In the next section, we provide distribution-dependent regret bounds.

C. Distribution-Dependent Regret Bounds

Before providing the distribution-independent regret bound, we provide an intermediate result on the number of times an irrelevant source is selected. Let $\mathcal{N}_{\mathcal{R}} = \{n \leq N : T^r(n) \geq 8(p_{\min})^{-2}\ln n \;\; \forall r \in \mathcal{R}\}$.

Lemma 2. *For all $n \in \mathcal{N}_{\mathcal{R}}$, if $\underline{\delta}_{exp} > 0$, the expected number of times RS-UCB selects the irrelevant source is bounded, i.e., $\forall d \in \mathcal{D} \setminus \mathcal{R}$,*

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}}(T^d(n)) \leq 128KM\underline{\delta}_{exp}^{-2} \ln n \\ + 5KDR(M+1)\pi^2/6 + 1.$$

Proof. (Sketch) We use three events used in Theorem 1 to bound the number of times RS-UCB selects irrelevant sources. For any $d \in \mathcal{D} \setminus \mathcal{R}$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}}(T^d(n)) \leq m + \sum_{n \in \mathcal{N}_{\mathcal{R}}} \mathbb{E}(\mathbb{I}(d \in \mathcal{O}_n, T^d \geq m)).$$

By choosing $m = 128KM\underline{\delta}_{\text{exp}}^{-2} \ln n + 1$, it can be shown by using union and Chernoff bounds that for $n \in \mathcal{N}_{\mathcal{R}}$,

$$\mathbb{E}(\mathbb{I}(d \in \mathcal{O}_n, T^d \geq m)) \leq 5KD(M+1)n^{-2}.$$

□

The next theorem combines Lemma 2 with the regret due to selecting a suboptimal arm even the relevant sources are selected.

Theorem 2. *(Distribution Dependent Regret Bound) If $\underline{\delta}_{exp} > 0$ and $p_{\min} > 0$, then the regret of RS-UCB is bounded as follows :*

$$\overline{Reg}_N^{\textit{RS-UCB}} \leq A_3 \ln N + A_4$$

where

$$A_3 = 8R(p_{\min})^{-2} + 128KMD(\underline{\delta}_{exp})^{-2} \\ + 8 \sum_{\boldsymbol{x}_{\mathcal{R}}} \sum_{a \in \mathcal{A} \setminus a^*(\boldsymbol{x})} \Delta_a^{-1}(\boldsymbol{x}_{\mathcal{R}})$$

and

$$A_4 = (R+1)D + 5KDR(M+1)\pi^2/6 \\ + (1 + \pi^2/3) \sum_{a \in \mathcal{A}} \sum_{\boldsymbol{x}_{\mathcal{R}}} \Delta_a(\boldsymbol{x}_{\mathcal{R}})$$

Proof. (Sketch) This result follows by combining Lemma 2 and UCB-type bounds. □

Theorem 2 shows distribution-dependent regret bound. There are 2 important differences between the regret bound in Theorem 1 and 2. Firstly, regret in Theorem 2 depends quadratically on $\underline{\delta}_{\text{exp}}^{-1}$ instead of $\delta_{\min}^{-1}$. Note that $\underline{\delta}_{\exp} \geq \delta_{\min}$ by definition. Therefore, bound in Theorem 2 has better dependence on relevance gaps. Secondly, there is an additional term which quadratically depends on $p_{\min}^{-1}$. The additional term appears in the result since we bound the number of times irrelevant source selected by RS-UCB when we have sufficient confidence in the relevant sources as seen from Lemma 2.

D. Lower Bounds

In this section, we first construct a simple problem for our setting and then prove a lower bound on the number of times any policy selects a wrong (irrelevant) source.

Simple Problem : Consider 2 types of source with $\mathcal{X}_1 = \{i_1, i_2\}$ and $\mathcal{X}_2 = \{i_3, i_4\}$, where source 1 is relevant and source 2 is irrelevant to all actions. Let $\bar{y}^1 = \left(\bar{y}_1^1(i_1), \bar{y}_2^1(i_1), \bar{y}_1^1(i_2), \bar{y}_2^1(i_2)\right)$ and $\bar{y}^2 = \left(\bar{y}_1^2(i_3), \bar{y}_2^2(i_3), \bar{y}_1^2(i_4), \bar{y}_2^2(i_4)\right)$ where $\bar{y}_a^d(x_d)$ is the expected reward of arm a for information x from source d. We assume that reward for action a with information x_d from source d is Bernoulli distributed with parameter $\bar{y}_a^d(x)$. Assume that $\mathbb{P}(x_1 = i_1) = \mathbb{P}(x_2 = i_3) = \nu_1$ and $\mathbb{P}(x_1 = i_2) = \mathbb{P}(x_2 = i_4) = \nu_2$. Note that decision-maker is restricted to select a single source in this simple problem, i.e., $R = 1$.

Theorem 3. *Suppose that* $\bar{y}^1 = (0.5+\delta_1/2, 0.5-\delta_1/2, 0.5-\delta_2/2, 0.5+\delta_2/2)$ *and* $\bar{y}^2 = (0.5, 0.5, 0.5, 0.5)$. *For any policy that aims to maximize its expected reward,* $\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}}\left[T^2(n)\right]$ *is lower bounded as follows,*

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}}\left[T^2(n)\right] \geq \frac{\ln n}{S(\delta_1, \delta_2)} \tag{5}$$

where

$$S(\delta_1, \delta_2) = \nu_1 KL\left(\frac{1}{2}, \frac{1+\delta_1}{2}\right) + \nu_2 KL\left(\frac{1}{2}, \frac{1+\delta_2}{2}\right),$$

and $KL(x_1, x_2)$ *denotes Kullback-Liebler divergence between Bernoulli distributions with mean* x_1 *and* x_2.

Proof. (Sketch) We convert this problem to a hypothesis testing problem by introducing two distributions ($\mathbb{D}_0$ and $\mathbb{D}_1$) : source 1 is relevant in the first distribution and both are irrelevant in the second distribution. We use the minimax bound on hypothesis testing [Tsyabkov, 2005] and simplify the KL divergence between two distributions as

$$\mathrm{KL}(\mathbb{D}_0^n, \mathbb{D}_1^n) \leq S(\delta_0, \delta_1)\mathbb{E}(T^2(n)). \tag{6}$$

□

The lower bound shows that every policy which aims to distinguish relevant and irrelevant source types requires $O(\ln n / S(\delta_1, \delta_2))$ samples. Assume that the information arrival process is uniform, i.e

$\nu_1 = \nu_2 = 0.5$. For this simple problem with uniform arrival process, for all actions $\delta_a^1(i_1) = \delta_1$ and $\delta_a^1(i_2) = \delta_2$. Note that by using the upper bound for KL divergence of two Bernoulli distribution [Rigollet and Zeevi, 2010], we get $\text{KL}\left(0.5, 0.5 + \frac{\delta_1}{2}\right) \leq C\delta_1^2$ for some constant $C > 0$ independent of problem parameters. Then, by using Jensen's inequality, we can show that

$$0.5\delta_1^2 + 0.5\delta_2^2 \geq (\underline{\delta}_{\text{exp}}^1)^2 = (0.5\delta_1 + 0.5\delta_2)^2 \tag{7}$$

where equality holds when $\delta_1 = \delta_2$. Therefore, we can argue that our source selection algorithm is approximately optimal in the special case. However, there is a gap between the regret bound of RS-UCB and lower bound for general case in terms of dependence to the relevance gap. Although RS-UCB is optimal in the special cases, an algorithm which matches the lower bound for the general case may be developed, which is left as a future work.

VI. Extension to Action-Dependent Relevant Sources

Our current setting assumes that the relevant sources are the same for all actions. Our framework can be adapted to learn the relevance relation for each action. Let $\mathcal{R}(a)$ denote the relevant sources for action a; all the sources $d \in \mathcal{D} \setminus \mathcal{R}(a)$ are irrelevant to action a. Let $\mathcal{R} = (\mathcal{R}(a))_{a \in \mathcal{A}}$ be the set of relevant sources, where $|\mathcal{R}| = R$(observability constraint). Next, we provide an modified version of RS-UCB, which we refer to as Generalized Relevant Source Upper Confidence Bounds (GRS-UCB).

For this extension, we define an action dependent relevance metric to select the sources, i.e.,

$$\hat{\text{Rel}}_d^a \triangleq \sum_{x_d \in \mathcal{X}_d} \frac{T_a^d(x_d)}{T_a^d} \left| \hat{y}_a^d(x_d) - \hat{y}_a \right|.$$

Based on the action dependent relevance metric, we assign an index for each tuple of source and action as

$$\text{ind}_d^a \triangleq \hat{\text{Rel}}_d^a + \sqrt{\frac{8M \ln n}{T_a^d}}$$

Based on these indices, we select sources and construct $\hat{\mathcal{R}}_n(a)$ for each action a for the n^{th} instance. Let $\mathcal{O}_n$ be all select sources, i.e., $\mathcal{O}_n = \left[\hat{\mathcal{R}}_n(a)\right]_{a \in \mathcal{A}}$ such that $|\mathcal{O}_n| = R$. The details of source selection procedure and pseudo-code for the GRS-UCB is given in supplementary material. Having constructed the relevance sets, GRS-UCB selects the action with highest confidence where the confidence of an action is defined as

$$\text{conf}_a(\boldsymbol{x}_{\hat{\mathcal{R}}_n(a)}) \triangleq \hat{y}_a^{\hat{\mathcal{R}}_n(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}_n(a)}) + \sqrt{\frac{2 \ln n}{T_a^{\hat{\mathcal{R}}_n(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}_n(a)})}}.$$

Next, we provide the distribution-independent regret analysis of GRS-UCB. Let $\delta^d_{\min}(a)$ be the relevance gap of source d for action a. Similar to Lemma 1, it can be shown that $\delta^d_{\min}(a) = 0$ for all the sources $d \in \mathcal{D} \setminus \mathcal{R}(a)$. Let $\delta_{\min} = \min_{a \in \mathcal{A}} \min_{d \in \mathcal{R}(a)} \delta^d_{\min}(a)$.

Theorem 4. *If $\delta_{\min} > 0$, then the regret of GRS-UCB is bounded as follows :*

$$Reg^{\textit{RS-UCB}}_N(\boldsymbol{x}^N) \leq A_5 \ln N + A_6$$

with probability 1 where

$$A_5 = 32K^2 DM \delta^{-2}_{\min} + 8 \sum_{\boldsymbol{x}_{\mathcal{R}}} \Delta^{-1}_a(\boldsymbol{x}_{\mathcal{R}})$$

and

$$\begin{aligned} A_6 &= 2KD + 2K^2 DR(M+1)\pi^2/3 \\ &+ (1+\pi^2/3) \sum_{a \in \mathcal{A}} \sum_{\boldsymbol{x}_{\mathcal{R}}} \Delta_a(\boldsymbol{x}_{\mathcal{R}}) \end{aligned}$$

VII. Numerical Results

A. Description of Datasets

Dataset 1 (D1): This is the breast cancer diagnosis dataset from UCI archive [Lichman, 2013]. The original dataset 569 instances and 30 clinically relevant attributes to the diagnosis (features extracted fromdifferent radiological images). The label is either malignant or benign. We create 7 actions by training Naive Bayes classifiers using (different) 10 instances from the dataset. Algorithms are tested on 50000 instances drawn uniformly from the original 569 instances (excluding 70 training instances) with using 6 attributes out of original 30 attributes. For each patient, our learning algorithm decides which radiological tests to perform and uses the features extracted from the obtained radiological to make a diagnosis decision.

Dataset 2 (D2): This is a synthetic dataset. Data is generated using a uniform and has 2 sources with 3 possible values and 3 actions, i.e., $\mathbb{P}(x_{d,n} = x) = 1/3$ for each source d and information x and instance n. Then, reward of the n^{th} instance for action a is generated through a Bernoulli process with mean $\bar{y}^1_a(x_{1,n})$. Therefore, relevant source is source 1. Total number of instances $N = 100000$.

B. Results

Results on irrelevant source selection In this result, we use (D2) to illustrate the source selection method of RS-UCB. Figure 2 shows the percentage of irrelevant source selection of RS-UCB based on

$\underline{\delta}_{\text{exp}}$. When $\underline{\delta}_{\text{exp}} = 0.25$, RS-UCB only selects irrelevant source 3% of time. The percentage of irrelevant source selection is increasing quadratically with decreasing $\underline{\delta}_{\text{exp}}$ as predicted by Theorem 2.

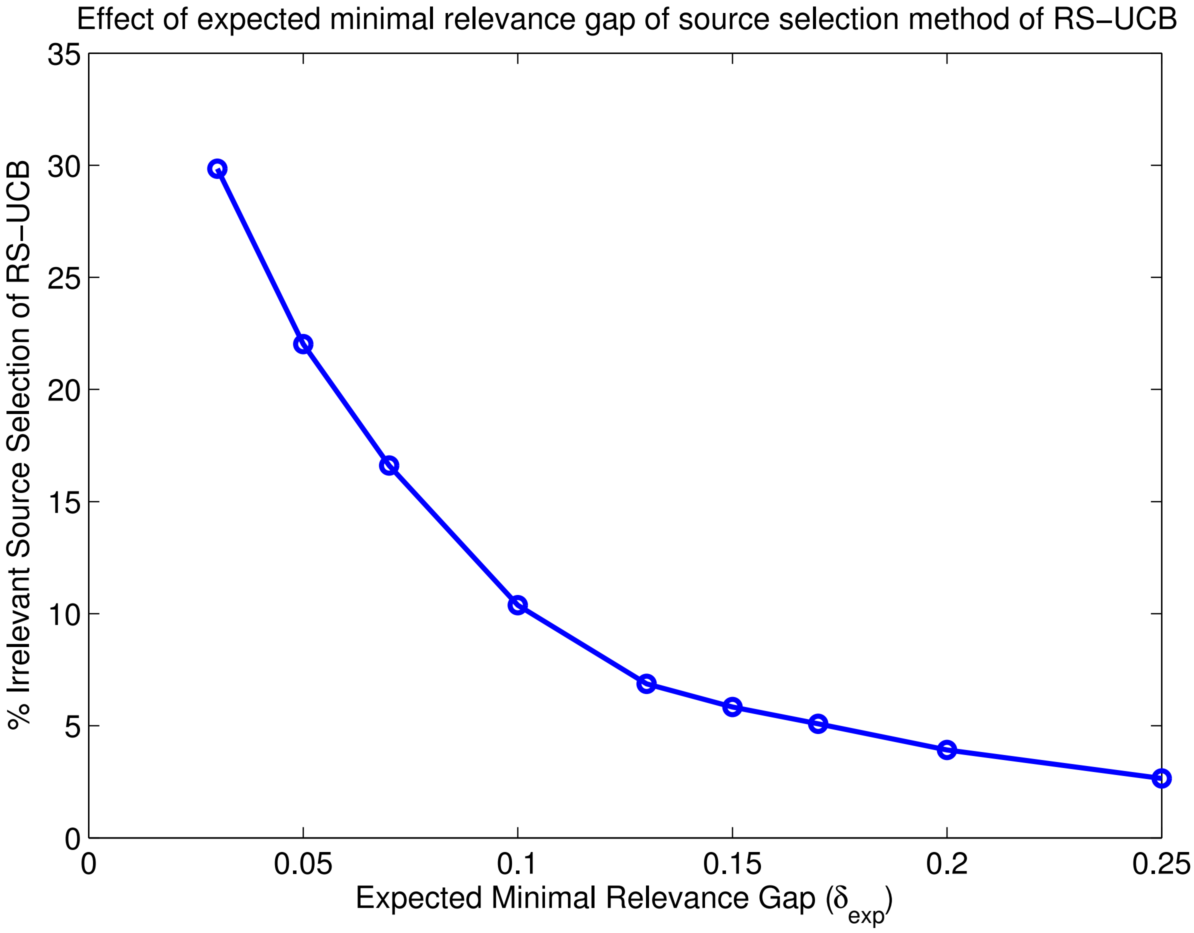

Fig. 2: The effect of $\underline{\delta}_{\text{exp}}$ on source selection method of RS-UCB

Results on prediction accuracy : In this result, we use (D1) to evaluate the performance of RS-UCB with state-of-art contextual bandit algorithms : LinUCB ([Chu et al., 2011]), Epoch-Greedy ([Langford and Zhang, 2007]) and ORL-CF [Tekin and Van Der Schaar, 2014]. Table III summarizes the performance on the breast cancer dataset.

Table III shows that RS-UCB performs slightly better than LinUCB and ORL-CF and slightly worse than Epoch-Greedy when $R = 2$. Note that RS-UCB requires less information to be observed/measured/consulted. Although LinUCB and Epoch-Greedy are observing all the information, they do not consider the relevance relationship, thereby leading to more exploration. The poor performance of ORL-CF is due to the adopted

Algorithms	# observations	Error	Relevance
RS-UCB	R=1	8.9%	Yes
RS-UCB	R=2	7%	Yes
ORL-CF	6	10%	Yes
LinUCB	6	7.4%	No
Epoch-Greedy	6	6.2 %	No

TABLE III: Comparison with state-of-the art algorithms

relevance learning metric, which scales with D^{R+1}.

VIII. Conclusion and Future Work

In this paper, we presented a novel algorithm that is able to learn what information sources should a decision maker consult to make optimal decisions, i.e. it "learns what to learn" to inform its decisions. For the proposed algorithm, we show a logarithmic regret bound and matching lower bound.

In the current framework, the information sources are assumed to be independent of each other. When the sources are dependent to each other, the relevance metric defined in (2) might not converge to relevant information source since $\delta_a^d(x_d) \neq 0$ even if source d is irrelevant to decision making problem. Therefore, new relevance metric should be used to identify the relevant source(s) for this case. This extension to correlated sources is left as a future work.

There are many interesting other directions for future work. One especially promising direction is the case in which consulting a source is costly, and consulting different sources involves different costs. In such a setting, the decision maker needs to trades-off the impact of the informativeness of the sources and the costs of consulting such sources on the total rewards.

References

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002. IV

N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Efficient learning with partially observed attributes. *The Journal of Machine Learning Research*, 12:2857–2878, 2011. I, II

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011. II, III-A, VII-B

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011. II

E. Hazan and T. Koren. Linear regression with limited observation. In *Proc. 29th Int. Conf. on Machine Learning*, pages 807–814, 2012. I, II

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. II

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems (NIPS)*, 20:1096–1103, 2007. II, VII-B

M. Lichman. UCI machine learning repository, 2013. VII-A

T. Lu, D. Pál, and M. Pál. Contextual multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 485–492, 2010. II

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011. II

H. B. Mann and A. Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14:217–226, 1943. V-A

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In *COLT 2010*, 2010. V-D

S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011. II

A. Slivkins. Contextual bandits with similarity information. In *24th Annual Conference On Learning Theory*, 2011. II, III-A

C. Szepesvari and N. Zolghadr. Onliner learning with costly features and labels. In *Advances in Neural Information Processing Systems (NIPS)*, 2013. II

C. Tekin and M. Van Der Schaar. Discovering, learning and exploiting relevance. In *Advances in Neural Information Processing Systems*, pages 1233–1241, 2014. II, VII-B

A. Tsyabkov. *Introduction to Nonparametric Estimation*. Springer, 2005. V-D

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th Int. Conf. on Machine Learning*, volume 20, pages 123–224, 2003. II