Best Arm Identification with Biased Contexts

James Cheshire, Stephan Clémençon

LTCI, Telecom Paris
james.cheshire@telecom-paris.fr, stephan.clemencon@telecom-paris.fr

Abstract

We study active mitigation of selection bias in statistical learning. That is sequential maximization over a set $\mathcal{A}$ of the expectation of a reward function $R(a, X)$ w.r.t. a r.v. X drawn from a target distribution P_T possibly different from the (supposedly dominating) source distribution P_S under which rewards are observed. The importance function $dP_T/dP_S(x)$ with which the sequentially observed biased rewards should be ideally weighted being unknown in practice, auxiliary information is assumed to be available in the form of known moments of the target distribution P_T for debiasing purposes. In the batch setting, this problem has already been studied and can be solved under certain conditions in two successive steps: 1) identify a weight function so as to approximate the moments 2) maximize the resulting (empirical version of the) weighted reward. In the active setting, if the problem boils down to identifying the best arm in a stochastic multi-armed bandit (MAB) model, the presence of selection bias strongly affects the complexity of the sequential optimization problem and requires the development of a new algorithmic approach, as we show here. In a fixed confidence setting, we introduce a novel notion of complexity, which accounts for the balance between arm evaluation and (parametric) weight function estimation, establish lower bounds and propose an algorithm proved to be near optimal. Theoretical guarantees are backed up by numerical results.

Introduction

The contextual stochastic multi-armed bandit (MAB) model has proved very useful to describe and analyze various sequential decision-making problems under uncertainty. It has applications in fields as varied as personalized medicine, online advertising retargeting or recommendation systems, for example. We consider the case of *post action context*, the classic formulation of which stipulates that, at each round $t \geq 1$, after choosing an action or an arm a_t from a finite arm set $\mathcal{A}$, one observes some information x_t, the *context*, a random variable taking its values in a measurable space $\mathcal{X}$ with distribution P. The reward then depends both on the context x and the chosen arm a and is written $R(a, x)$ where $R : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function. Several versions of this generic model have been studied in the literature, also in the alternate setting where the context is observed before the learner chooses an arm, in order to understand the achievable bounds of regret minimization in this setting, see (Lattimore and Szepesvári 2020) for an overview. More recently, the problem of best arm identification (BAI) for the contextual bandit has also been considered, see (Kato and Ariu 2021),(Guan and Jiang 2018), (Deshmukh et al. 2020) and (Qin and Russo 2023). In the case of BAI, the objective is to identify the arms that are solutions to the maximization problem

$$\max_{a \in \mathcal{A}} \mathbb{E}_{X \sim P}[R(a, X)] \, , \tag{1}$$

in a fixed-confidence or fixed-budget setting. In this paper, we study best arm identification, in a fixed-confidence framework, by relaxing the assumption that the (source) distribution P_S of the observed contexts X is the same as that under which the learner is evaluated, the target distribution $P = P_T$. This is the classical covariate shift setting, in which the marginal distributions on $\mathcal{X}$ change from source to target while the conditional distributions, i.e. the reward of an action given a context, remain the same. This problem is motivated by the fact that the data collected by modern sensors (*e.g.* smartphone applications, web, social networks, IoT) are often not acquired in a controlled way and in many cases are not representative of the phenomenon under study. In the applications mentioned above, it is often the case that certain patient or user profiles, or certain conditions under which a system's performance is evaluated, may be over- or under-represented in the data observed. In medical datasets, for instance, certain sub populations are often over represented, see e.g. (Sudlow et al. 2015). In facial recognition this problem is also well documented, see *e.g.* (Wang et al. 2019).

Ignoring such a selection bias issue would naturally jeopardize the optimization problem (1). Off-line statistical learning based on biased training data has received much attention in the machine-learning literature these last few years, the representativeness of the examples used to learn decision/prediction rules being at the heart of the guarantees for the generalization of methods based on regret or risk minimization, see *e.g.* (Quionero-Candela et al. 2009). Of course, this problem cannot be solved without auxiliary information. This may concern the underlying bias mechanism at work. Related works are far too numerous to be listed exhaustively but one may refer to *e.g.* (Ausset, Clémençon, and Portier

2022) or (Li and Bradic 2020) when training data are subject to random censorship, to (Clémençon, Bertail, and Papa 2016) or (Clémençon, Bertail, and Chautru 2017) in the case where examples are collected by means of a survey plan, to (Laforgue and Clémençon 2022) in the context of (known) biasing models.

Alternatively, the auxiliary information may relate to the target distribution P_T itself, assumed to be absolutely continuous w.r.t the distribution P_S of the training examples, so that $\mathbb{E}_{X\sim P_T}[R(a,X)] = \mathbb{E}_{X\sim P_S}[g^*(X)\cdot R(a,X)]$ for all $a \in \mathcal{A}$ where $g^*(x) = (dP_T/dP_S)(x)$. It can take the form of a small sample composed of unbiased observations enabling direct estimation of the importance function g^*, as in (Huang et al. 2007; Sugiyama et al. 2007, 2008). Or it may be of macroscopic nature, in the spirit of calibration techniques (post-stratification) in survey sampling, see *e.g.* (Deville 2000), consisting of supposedly known characteristics, $M \geq 1$ generalized moments typically, of the target population P_T:

$M_l = \mathbb{E}_{P_T}[m_l(X)]$ for $l = 1, \ldots, M$, where the m_l's are known real-valued P_T-integrable functions on $\mathcal{X}$. Because it covers many situations in practice, this is the framework we consider here. It has been studied in the off-line setting in (Bertail et al. 2021) from a semi-parametric perspective and, to the best of our knowledge, the present paper is the first to address statistical learning in presence of selection bias in an active setting, formulated here as a best arm identification problem with biased contexts, in which absolutely no observations distributed according to the target distribution P_T are available.

As we show in this article, the problem significantly differs from its off-line counterpart in that it cannot be optimally solved in two successive phases: 1) estimating first a weight function $\hat{g}(x)$ in a parametric class supposed to be rich enough to contain g^* (or a reasonable approximant of the latter) by aligning with the known moments of P_T with high probability: $\mathbb{E}_{P_S}[\hat{g}(X)\cdot m_l(X)] \approx M_l$ for $l = 1, \ldots, M$, 2) solving an empirical version of the plug-in maximization problem $\max_{a\in\mathcal{A}} \mathbb{E}_{P_S}[\hat{g}(X)\cdot R(a,X)]$. To be optimal, such an approach would wish to balance the number of samples in each phase of the algorithm. However, to do this with the length of each phase fixed in advance, would require knowledge of both the suboptimality gaps of the arms and also how the expectation $\mathbb{E}_{P_S}[g(X)\cdot m_l(X)]$ changes with g across the parametric class. A naive approach to the active setting would be to run such a two stage method, with equal number of samples for phases 1 and 2, successive times, doubling the total number of samples each time. Equipped with a suitable stopping time, this approach could potentially return the optimal arm with high probability and finite expected sampling time, while not requiring additional information. However, our algorithm, as well as being significantly simpler, does not suffer additional log terms in its upper bound on expected sampling time, that would be incurred by the aforementioned approach.

Related Works

The contextual armed bandit, in the case where the source distribution of the contexts P_S differs to the one on which the learner is evaluated, P_T, has been recently studied under a *transfer exponent assumption*. Such an assumption, first introduced in (Kpotufe and Martinet 2020), is considered in the case where $\mathcal{X} = \mathbb{R}^d$ and is as follows,

$$\mathbb{P}_S(B(x,r)) \gtrsim r^\gamma \mathbb{P}_T(B(x,r)) \ ,$$

where $\gamma > 0$, $B(x,r)$ is the ℓ_∞ ball of radius $r \in (0,1]$ centered at $x \in \mathcal{X}$. Essentially, when the transfer exponent γ is small, areas with a large weighting under the target distribution P_T will be sufficiently covered in the source distribution P_S. Contextual armed bandits, for regret minimisation have been studied under covariate shift with a transfer exponent assumption. Specific settings have been considered where the learner interacts with the environment in two phases, where the first phase contexts are drawn from the source distribution and then from the target distribution in the second stage. In (Suk and Kpotufe 2021) the second phase, i.e. the covariate shift, is assumed to occur at some unknown change point, whereas in (Cai, Cai, and Li 2024) the learner is given a set of prelabeled samples with contexts drawn from the source distribution at the start of the game, interacting solely with the target distribution from then on. The above works differ from ours in two key points, that they consider regret minisation as opposed to BAI and that the learner observes both samples from the source and target distribution. The later point puts the above works in the field of transfer learning.

The article is organized as follows. The probabilistic framework for active learning under selection bias we consider here is detailed in the section back ground and preliminaries, together with the main assumptions involved in the subsequent analysis. The TACTIC algorithm for besT Arm identifiCaTion with bIased Contexts we propose and the specific notion of complexity we introduce to capture the nature of the sequential learning problem under study, are presented in the section problem complexity and algorithm. The main theoretical results, revealing that the algorithm we propose is PAC(δ), providing an upper bound for its expected sampling time, as well as a lower bound showing its near optimality, are stated in the section main theoretical results. In section numerical experiments, we illustrate the empirical performance of the TACTIC algorithm and then in the conclusion, outline some avenues for future research. Due to space constraints, certain technical details are deferred to the Supplementary Material.

Background and Preliminaries

Here we describe in detail the contextually biased bandit framework that we analyze, theoretically and empirically, in the following sections. Here and throughout, by $|E|$ is meant the cardinality of any finite set E

Contextual finite-armed bandit model. Let $\mathcal{A}$ be the finite arm set, $\mathcal{X}$ be the context set. For each arm $a \in \mathcal{A}$ and context $x \in \mathcal{X}$ we define $R(a,x) \in [0,1]$ as the reward associated to arm a under context x. The learner plays a game in several rounds: at the beginning of each round t, the learner chooses a_t, the learner then observes a context $x_t \in \mathcal{X}$, drawn from an unknown probability distribution

P_S on $\mathcal{X}$. The learner then receives a noisy evaluation of the reward $R(a_t, x_t)$, namely

$$Y_t := R(a_t, x_t) + \omega_t \ , \tag{2}$$

where the noise ω_t is assumed to be a centered 1-sub-Gaussian real valued random variable (*i.e.* $\mathbb{E}[\exp(u\omega_t)] \leq \exp(u/2)$ for all $u \in \mathbb{R}$), independent from the past contexts and arm choices. The (ω_t)'s are supposed to form an i.i.d. sequence, just like the x_t's.

Biased contexts. While contexts are drawn at each round according to the *source distribution* P_S, the learner aims to maximise performance according to some unknown distribution P_T, different from P_S and referred to as the *target distribution*. Specifically, one defines the (supposedly unique) optimal arm as

$$a^* := \sup_{a \in \mathcal{A}} \mathbb{E}_{X \sim P_T}[R(a, X)] \ , \tag{3}$$

where $\mathbb{E}_{P_T}$ is the expectation under the target distribution. Once a sufficient amount of time has elapsed, chosen at the learner's discretion, the aim is to produce a prediction, $\hat{a}$, of the optimal arm.

Auxiliary macro-information. Of course, the problem can only be solved if some auxiliary information on the target distribution P_T is available. We assume that the latter is dominated by the source distribution P_S and that the Radon-Nikodym derivative $g^*(x) := (dP_T/dP_S)(x)$ between P_S and P_T belongs to a parameterised class of link functions.

Assumption 1. *There exists a measurable function $g : \mathcal{X} \times [0,1]^d \to \mathbb{R}_+$ with $\sup_{\alpha \in [0,1]^d, x \in \mathcal{X}} |g(\alpha, x)| < \infty$ such that $g^*(x) \in \{g(\,\cdot\,, \alpha) : \alpha \in [0,1]^d\}$.*

For $\alpha \in [0,1]^d$ and $a \in \mathcal{A}$, we write $\mu_{a,\alpha}$ for the mean reward of arm a, assuming contexts are drawn according to the distribution with p.m.f $g(\alpha, x)dP_S(x)$, that is,

$$\mu_{a,\alpha} := \mathbb{E}_{X \sim P_S}[g(\alpha, X)R(a, X)] \ .$$

We also write α^* for the optimal parameter providing the link function, that is, for all $x \in \mathcal{X}$,

$$g(\alpha^*, x) = g^*(x) \ .$$

We assume that the auxiliary information available for debiasing consists of known generalized moments of the target distribution.

Assumption 2. *There exist $L \geq 1$ real-valued P_T-integrable functions $m_1, \ \ldots, \ m_L$ such that the following quantities are known: $M_l := \mathbb{E}_{P_T}[m_l(X)]$ and for any $x \in \mathcal{X}$ we write the vector $\underline{m}(x) = (m_1(x), ..., m_L(x))$. We assume that for all $x \in \mathcal{X}$, $l \in L$, $|m_l(x)| < 1$, and furthermore, there exists $\beta > 0$ such that for all $\alpha_1, \alpha_2 \in [0,1]^d$,*

$$|g(\alpha_1, x) - g(\alpha_2, x)| \leq \|\alpha_1 - \alpha_2\|_\infty^\beta \ .$$

Remark 1. (ON ASSUMPTION 2) *Observe that Assumption 2 covers many practical situations. While data collected via certain modern modalities (e.g. web applications) are notoriously subject to selection bias with possibly significant over/under-representation of certain segments of the population of interest, population-level statistics (e.g. socio-demographic features such as average salary, household composition, health status or life expectancy) are often available, see e.g. the portal of the Office for National Statistics in the UK http://infuse.ukdataservice.ac.uk/InFuse or the interface of the US Census Bureau https://www.census.gov/quickfacts/Quickfacts. Such auxiliary information is commonly used to adjust estimates in surveys (calibration), and we propose here to use it for active learning purposes.*

Class of problems. Throughout the rest of the paper, we consider the arm and context sets $\mathcal{A}, \mathcal{X}$, the constants L and the i.d.d. sequence of sub gaussian noise ω_t as fixed. A biased contextual bandit problem ν can then be defined as a tuple of the matrix of rewards $(R(a,x))_{a \in \mathcal{A}, x \in \mathcal{X}}$ where $R(a,x) \in [0,1]$, distributions P_T, P_S on the context set $\mathcal{X}$, and $L \geq 1$ real-valued P_T-integrable functions. We write $\mathcal{B}$ for the set of all problems $\nu = ((R(a,x))_{a \in \mathcal{A}, x \in \mathcal{X}}, P_T, P_S, \underline{\mathrm{m}})$, such that Assumptions 1 and 2 are satisfied.

Policies and fixed confidence regime. The way the learner interacts with the environment - i.e. their choice of arms and contexts to query, how many samples to draw in total and their estimated best arm, we term the *policy* of the learner. We write $\mathcal{C}$ for the set of all possible policies of the learner. For a policy $\pi \in \mathcal{C}$ and problem $\nu \in \mathcal{B}$ we denote random variable τ_ν^π as the stopping time of policy π. We write $\hat{a}_\nu^\pi$ for the arm outputted by policy π on problem ν. Where obvious we may drop the dependency on π, ν in the notation, referring to the arm outputted by the learner as simply $\hat{a}$. We write $\mathbb{P}_{\nu,\pi}$ as the distribution on all samples gathered by a policy π on problem ν. We similarly define $\mathbb{E}_{\nu,\pi}$, again, where obvious we may drop the dependency on π, ν in this notation.

For the duration of this paper we will work in the *fixed confidence regime*. For a confidence level δ, a policy π is said to be PAC(δ), on the class of problems $\mathcal{B}$, if, $\forall \nu \in \mathcal{B}, \mathbb{P}_{\nu,\pi}[\hat{a}_\nu^\pi = a^*] \geq 1 - \delta$. The goal of the learner is to then obtain a PAC(δ) policy π, such that the expected total number of samples drawn, in the worst case, $\sup_{\nu \in \mathcal{B}} \mathbb{E}_{\nu,\pi}[\tau_\nu^\pi]$, is minimised.

Toy example satisfying Assumptions 1 and 2. Before continuing to our proposed algorithm and theoretical guarantees, let us consider a simple example of a problem satisfying Assumptions 1 and 2, that can be easily related to many practical applications. Let us assume $\mathcal{X} = \mathbb{R}$, and let P_T, P_S be variance 1 gaussian distributions with means μ_T, μ_S respectively. Set $L = 1$ and let $m(x) = x$ with $\mathbb{E}_{P_T}[m(x)] = \mu_T$, that is, the expectation of the contexts, under the target distribution are known to the learner. The expectation of the contexts under the source distribution, however, remains unknown to the learner. The class of functions $\{g(\,\cdot\,, \alpha) : \alpha \in [0,1]^d\}$ is then simply the set of Radon–Nikodym derivatives between a standard gaussian of mean μ_1 and arbitrary mean α, i.e. for $\alpha \in \mathbb{R}$,

$$g(x, \alpha) = \exp\left(\frac{(\mu_T - x)^2 + (\alpha - x)^2}{2}\right) .$$

In this example we then have that $\alpha^* = \mu_S$. We see that to select the optimal arm under the target distribution, the

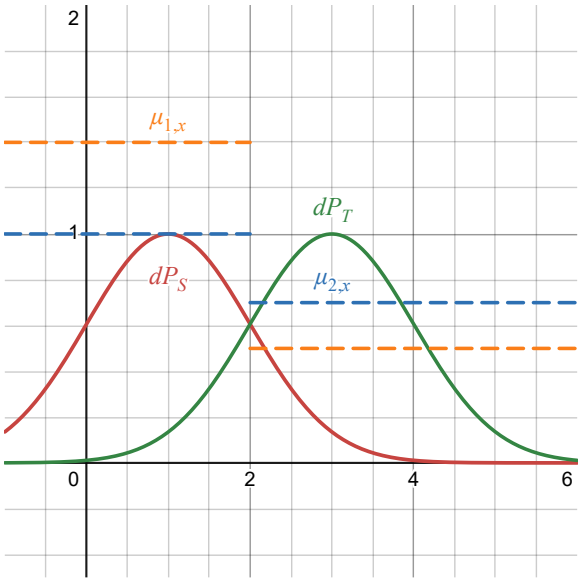

Figure 1: The expected reward of the first and second arms across the contexts, are given in orange and blue respectively. The p.m.f. of the distribution of the contexts under the source and target distribution are in red and green respectively.

learner must first estimate the expectation of the contexts under the source distribution. The number of samples required to do so optimally, i.e. maintaining a PAC(δ) guarantee in as few samples as possible, will depend upon the quantity $|\mu_T - \mu_S|$, unknown to the learner and necessitating the need for an adaptive approach.

An illustration of the above example can be seen in Figure 1. In Figure 1 we consider a simple case with only two arms, where the expected reward of both arms across the context set are given by step functions. We see that, on sections of the context space more likely to come up under the source distribution, the first arm has a higher expected reward, whereas on sections of the context space more likely to come up under the target distribution, the second arm has higher expected reward. Thus to correctly identify the optimal arm under the target distribution, the learner must first de-bias the contexts.

Problem Complexity and Algorithm

Here we introduce the quantities specific to the solving of the problem formulated in the previous section and describe the algorithm we propose based on them.

Capturing the complexity of best arm identification with biased contexts

We work in the problem dependent regime in that our results will depend upon features of the problem, e.g. the rewards $R(a, x)$ of the arms under the contexts and the distributions P_S, P_T. With this in mind, given $\alpha \in [0, 1]^d$ and arm $a \in \mathcal{A}$ we define the following optimality gap,

$$\Delta_{\alpha,a} := \sup_{b \in \mathcal{A}} \mu_{b,\alpha} - \mu_{a,\alpha} \, .$$

We see $\Delta_{\alpha,a}$ is the suboptimality gap of arm a, when the contexts are drawn according to the distribution with p.m.f $g(\alpha, x)dP_S(x)$. For a $\alpha \in [0, 1]^d$ We also define the gap,

$$\nabla_\alpha := \frac{1}{L} \sum |\eta_{\alpha,l} - M_l| \, ,$$

where $\eta_{\alpha,l} = \mathbb{E}_{X \sim P_S}[g(\alpha, X)m_l(X)]$. We see that for $\alpha \in [0, 1]^d$, the gap ∇_α is the average difference in the expectation of the functions m_l under the distribution with p.m.f $g(\alpha, x)dP_S(x)$, and their expectation under the target distribution. The gap ∇_α will essentially quantify how hard it is to distinguish α from the optimal parameter α^*

An optimal learner will balance estimation of the arm means against estimation of the parameter α^*, such that $g(x, \alpha^*)$ is the link function between the distributions P_S and P_T. Indeed, perfect recovery of α^* may not be necessary for perfect recovery of a^*. If the learner is confident that for some set $C \subset \mathcal{X}$ and arm $\tilde{a}$, the following two properties hold

$$\alpha^* \in C \, ,$$

and

$$\forall \alpha \in C, \mu_{\tilde{a},\alpha} \geq \mu_{a,\alpha} \, .$$

they can be confident $\tilde{a} = a^*$. With this in mind, for some problem $\nu \in \mathcal{B}$ we now define the complexity term H_ν which essentially captures the balance between estimating α^* and a^*

$$H_\nu := \max_{\varepsilon > 0} \left\{ \frac{1}{\varepsilon^2} : \sum_{a \in \mathcal{A}} \max_{\alpha : \nabla_\alpha \leq \varepsilon} \left(\frac{1}{\Delta_{a,\alpha}^2} \right) \leq \frac{1}{\varepsilon^2} \right\} .$$

The quantity H_ν can be seen as the minimum number of samples required to identify the optimal arm, with the condition that one can first use the same number of samples to narrow down the context set $\mathcal{C}$, thereby reducing the complexity of finding the optimal arm. If H_ν were known, one could employ a two stage approach, to first estimate α^* and then a^*, using roughly H_ν samples in each stage. However, H_ν is unknown, necessitating our adaptive approach. We then also define the global gap of arm a as,

$$\Delta_a := \operatorname*{arg\,min}_{\alpha : \nabla_\alpha \leq H_\nu^{-1/2}} \Delta_{a,\alpha} \, .$$

Performance of naive two stage approach Consider a fixed budget two stage algorithm i.e. Algorithm 3, which first estimates the optimal $\alpha : g(\alpha, x) = g^*(x)$ with half it's budget, and then estimates the optimal arm with the remaining half. If the budget in said approach is of the order H_ν then we can expect such an algorithm to match the performance of our own, `TACTIC` . However, the value H_ν is dependent upon the $\mu_{a,\alpha}$ and η_α, which are assumed to be unknown to the learner. As an alternative, one could run such a two stage approach iteratively, doubling the allocated budget at each time step. Theoretically, such a strategy can expect to iterate over the two stage approach roughly $\log(H_\nu)$ times. To ensure the algorithm does not stop prematurely one would need to increase the total number of samples drawn, leading to an additional $\log(H_\nu)$ term in the upper bound on expected sampling time. Another draw back of such an approach would be poor experimental performance due to repeated splitting of budget.

The TACTIC algorithm

The algorithm runs across several rounds. For each round t we maintain an active set of parameters $R_t \subset [0,1]^d$ and an active set of actions $A_t \subset \mathcal{A}$. During phase 1 of each round t, each arm remaining in A_t is sampled $\log(t\varepsilon_t^{-d/\beta}/\delta)2^t$ times with contexts drawn at random from the reservoir, for each arm $a \in A_t$ we write $(x_{a,s}^t)_{s<2^t\log(t/\delta)}$ for the contexts drawn while sampling arm a in round phase 1 of t. Similarly we write $(Y_{a,s}^t)_{s<2^t\log(t/\delta)}$ for the samples drawn from arm a, when sampling it in phase 2 of round t. For $\alpha \in [0,1]^d$, we now let $\widehat{\mu}_{a,\alpha}^t$ denote the empirical mean of arm a, calculated on samples drawn in phase 1 of round t, transformed via the contexts through the function $g(\alpha, x)$, i.e.

$$\widehat{\mu}_{a,\alpha}^t = \frac{1}{2^t \log(t\varepsilon_t^{-d/\beta}/\delta)} \sum_s g(\alpha, x_{a,s}^t) Y_{a,s}^t \, .$$

We also write $\hat{\Delta}_{a,t}$ for it's empirical gap, defined as follows,

$$\hat{\Delta}_{a,t} := \min_{\alpha \in R_t} |\widehat{\mu}_{a,\alpha}^t - \max_{b \in A_t} \widehat{\mu}_{b,\alpha}^t| \, .$$

In Phase 2 of round t, the TACTIC algorithm draws an additional $\log(t\varepsilon_t^{-d/\beta}/\delta)|A_t|2^t$ contexts from the reservoir, from which it estimates, for each $\alpha \in R_t$, η_α. We write $(\tilde{x}^t)_{s<\log(t\varepsilon_t^{d/\beta}/\delta)|A_t|2^t}$ for the contexts drawn from the reservoir in phase 2. Let $\hat{\eta}_{\alpha,l}^t$ denote the corresponding empirical mean for each $\alpha \in R_t$, i.e.

$$\hat{\eta}_{\alpha,l}^t = \frac{1}{\log\left(t\varepsilon_t^{-d/\beta/\delta}\right)|A|2^t} \sum_s g(\alpha, \tilde{x}_s^t) m_l(x) \, .$$

We then write $\widehat{\nabla}_{a,\alpha}$ for the empirical gap, calculated as follows,

$$\widehat{\nabla}_{\alpha,t} = \frac{1}{L} \sum_{l=1}^{L} |\hat{\eta}_{\alpha,l}^t - M_l| \, .$$

The TACTIC algorithm is an elimination algorithm, in that round by round we eliminate both arms and parameters from the active sets A_t and R_t respectively. We eliminate arms and parameters when their empirical gaps, $\hat{\Delta}_{a,t}$ and $\widehat{\nabla}_{\alpha,t}$ respectively, are above a certain tolerance. Our tolerance level for eliminating both arms and parameters decreases with time t, we set $\varepsilon_t = 2^{-t/2}$. At the end of round t an arm $a \in A_t$ is eliminated if $\hat{\Delta}_{a,t} \geq \varepsilon_t$. A parameter α is eliminated if $\widehat{\nabla}_{\alpha,t} \geq \frac{\varepsilon_t}{\sqrt{|A|}}$. By having the tolerance ε_t start high and then decrease over time, we ensure that TACTIC only eliminates arms from the active set when it is confident they are sub optimal on all remaining alphas in the active parameter set R_t. Similarly, alphas are eliminated from R_t only when we are confident that they are not the optimal parameter α^*, such that $g(\alpha^*, x)$, provides the link function between the source and target distributions. The reason we more readily eliminated parameters is that the gaps $\widehat{\nabla}_{\alpha,t} \geq \frac{\varepsilon_t}{\sqrt{|A|}}$ have been estimated with more samples, and we are thus more confident in our estimation.

Algorithm 1: TACTIC

Initialise: $A_1 = \mathcal{A}$, $R_1 = [0,1]^d$, $t = 1$
while $|A_t| > 1$ **do**
  Phase 1: Sample each arm in A_t, $\log(t\varepsilon_t^{-d/\beta}/\delta)2^t$ times, with contexts drawn from $\mathcal{X}$.
  Phase 2: Draw $\log(t\varepsilon_t^{-d/\beta}/\delta)|A_t|2^t$ contexts from $\mathcal{X}$.
  Remove all arms $a : \hat{\Delta}_{a,t} \geq \varepsilon_t$ from A_t
  Remove all $\alpha : \widehat{\nabla}_{\alpha,t} \geq \frac{\varepsilon_t}{\sqrt{|A_t|}}$ from R_t
  $t = t + 1$
end while
return $\hat{a} \in A_t$

Note that the TACTIC takes the smoothness parameter β as input, and thus we must assume β is known to the learner. However, as β only appears in the log term of our upper bound of 2, in practice a loose lower bound on β can be used.

Computational feasibility of TACTIC The reader will note that our algorithm TACTIC requires at each round t to calculate $\widehat{\nabla}_{\alpha,t}$ for all $\alpha \in [0,1]^d$. While this step does not require the learner to draw new samples, thereby having no effect on the total sampling time, it is potentially computationally unfeasible in practice. With this in mind we propose the algorithm TACTIC+ , which mirrors the algorithm TACTIC aside from one key difference. Instead of calculating $\widehat{\nabla}_{\alpha,t}$ for all $\alpha \in R_t$ at each round t, we only calculate $\widehat{\nabla}_{\alpha,t}$ for $\alpha \in \mathcal{E}_d(\varepsilon_t^{1/\beta})$, where we denote $\mathcal{E}_d(\varepsilon)$ the ε-net on $[0,1]^d$.

Algorithm 2: TACTIC+

Initialise: $A_1 = \mathcal{A}$, $R_1 = [0,1]^d$, $t = 1$
while $|A_t| > 1$ **do**
  Phase 1: Sample each arm in A_t, $\log(t\varepsilon_t^{-d/\beta}/\delta)2^t$ times, with contexts drawn from $\mathcal{X}$.
  Phase 2: Draw $\log(t\varepsilon_t^{-d/\beta}/\delta)|A_t|2^t$ contexts from $\mathcal{X}$.
  Remove all arms $a : \hat{\Delta}_{a,t} \geq \varepsilon_t$ from A_t
  For all $\alpha \in \mathcal{E}_d(\varepsilon_t^{1/\beta})$, if $\alpha : \widehat{\nabla}_{\alpha,t} \geq \frac{\varepsilon_t}{\sqrt{|A_t|}}$, remove $\{\tilde{\alpha} : \|\alpha - \tilde{\alpha}\| \leq \varepsilon_t\}$ from R_t
  $t = t + 1$
end while
return $\hat{a} \in A_t$

Main Theoretical Results

In the following theorem we demonstrate that, in the case where Assumptions 1 and 2 hold, both our algorithms, TACTIC and TACTIC+ , are PAC(δ), that is, they will return the optimal arm with probability greater than $1-\delta$. The proof can be found in the Appendix.

Theorem 1. *Consider a problem $\nu \in \mathcal{B}$, upon execution of both* TACTIC and TACTIC+ , we have that,

$$\mathbb{P}_\nu(\hat{a} = a^*) \geq 1 - \delta \, .$$

In the following theorem we provide an upper bound on the expected sampling time of our algorithm. The proof can be found in the Appendix.

Theorem 2. *Consider a problem $\nu \in \mathcal{B}$, upon execution of* TACTIC *we have that,*

$$\mathbb{E}[\tau_\nu] \leq c \sum_{a \in \mathcal{A}} \Delta_a^{-2} \log\left(\frac{c_a \log(c_a)}{\delta}\right) \log(c_\nu) \ ,$$

where $c_\nu = K \vee H_\nu^{d/2\beta}$, $c_a = c'(\Delta_a^{-2} \vee H_\nu/|\mathcal{A}|)\log(c_\nu)$ and $c, c' > 0$ are absolute constants.

Comparison to classic fixed confidence best arm identification In classical best arm identification, without the presence of contextual information, the sub optimality gap of an arm a is simply $\tilde{\Delta}_a = \mu^* - \mu_a$, where μ_a is the mean reward of the ath arm and μ^* is the maximum of the arm means. The state of the art rate for fixed confidence best arm identification in the standard multi armed bandit is then of the order $\sum(1/\tilde{\Delta}_a^2)\log(\sum 1/\tilde{\Delta}_a^2)$, achieved by the LUCB algorithm of (Kalyanakrishnan et al. 2012) - see (Jourdan, Degenne, and Kaufmann 2023),(Garivier and Kaufmann 2016) for state of the art bounds in the asymptotic regime. In our setting, in the case where the order of the number of samples needed to recover α^* is equal to the order of the number of samples needed to solve the maximisation problem $\max_{a \in \mathcal{A}} \mathbb{E}_{X \sim P_T}[R(a, X)]$, i.e. $\sum_{a \in \mathcal{A}} \Delta_{a,\alpha^*}^{-2} \approx H_\nu$, our upper bound becomes approximately $\sum \Delta_{a,\alpha^*}^{-2} \log(\Delta_{a,\alpha^*}^{-2})$. This would be of the same order as the state of the art bound for the classical best arm identification setting, in the case where the learner is given access to the distribution P_T directly. Essentially, this means that in the case where the contexts can be very easily de biased, our approach does not pay any significant additional cost compared to classical strategies.

We know provide an upper bound on the expected sampling time of our TACTIC+ algorithm.

Theorem 3. *Consider a problem $\nu \in \mathcal{B}$, upon execution of* TACTIC+ *we have that,*

$$\mathbb{E}[\tau_\nu] \leq c \sum_{a \in \mathcal{A}} \Delta_a^{-2} \log\left(\frac{c_a \log(c_a)}{\delta}\right) \log(c_\nu) \ ,$$

where $c_\nu = K \vee H_\nu^{d/2\beta}$, $c_a = c'(\Delta_a^{-2} \vee H_\nu/|\mathcal{A}|)\log(c_\nu)$ and $c, c' > 0$ are absolute constants.

The bound matches that of Theorem 2, up to constant terms, showing the learner essentially pays no cost for ensuring their approach is computationally feasible.

Lower bound In the following theorem we provide a lower bound on the expected sampling time of any PAC(δ) algorithm. The proof can be found in the Appendix.

Theorem 4. *Let $0 < \delta < 1/8$ and $\nu \in \mathcal{B}$. For any algorithm π such that, for all $\nu \in \mathcal{B}$,*

$$\mathbb{P}_\nu^\pi(\hat{a} \neq a^*) \geq 1 - \delta \ ,$$

there exist a problem $\tilde{\nu} \in \mathcal{B}$, such that the expected sampling time of said algorithm is upper bounded by, $cH_{\tilde{\nu}}$ where $c > 0$ is an absolute constant.

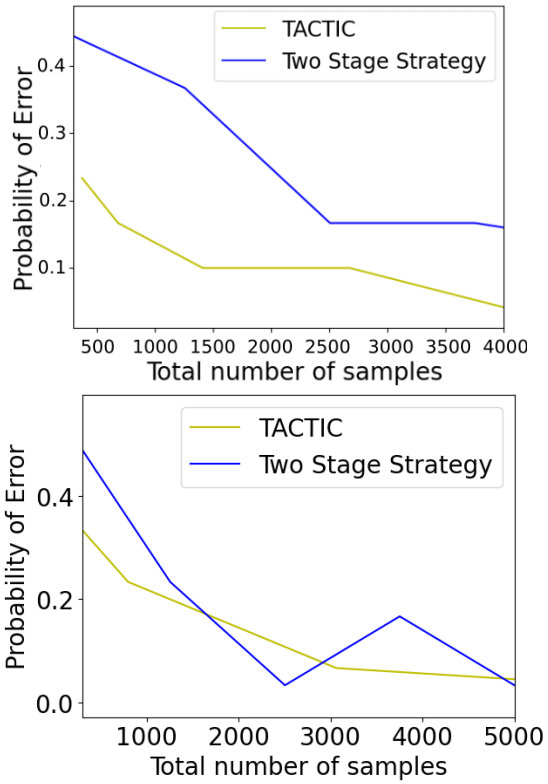

Figure 2: Empirical performance of TACTIC and Algorithm 3, on scenarios 1 and 2, top and bottom respectively, estimated via 50 Monte Carlo simulations

Note that our upper bound of Theorem 2 matches that of the above lower bound, Theorem 4, up to log terms.

Illustrative Numerical Experiments

In this section we provide illustrative numerical experiments to evaluate the empirical performance of the TACTIC algorithm on a toy problem. As a competitor we will consider a fixed budget two stage algorithm, Algorithm 3, which first estimates the optimal $\alpha : g(\alpha, x) = g^*(x)$ with half it's budget, and then estimates the optimal arm with the remaining half.

The setting we consider is as follows: we set the number of arms equal to 10, i.e. $|\mathcal{A}| = [10]$. The space of contexts $\mathcal{X}$ is set equal to the interval $[0, 1]$ with the source distribution P_S being uniform across the context set. Whereas in the theoretical analysis the parameter space for the function $g(\alpha, \cdot)$ is assumed to be $[0, 1]^d$, in the interest of ease of computation we restrict to a finite parameter space where in scenario 1 α takes values in $\{0.1, 0.9\}$ and in scenario 2 α takes values $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. In each scenario, the optimal parameter is selected at random from the range of possible values for α. We then define $g(\alpha, x) = \frac{\psi(x-\alpha)}{\Phi(1-\alpha)-\Phi(-\alpha)}$, where ψ and Φ are the p.m.f and c.d.f of a standard normal distribution respectively, that is $g(\alpha, x)$ is the p.m.f. of a normal distribution, truncated to the $[0, 1]$ interval. We see that, in

this setting, the distribution P_T will be a standard normal distribution, with mean α^*, truncated to the interval. In both scenario 1 and 2, the rewards are such that for arm a,

$$R(a,x) = \begin{cases} 5 & \text{if } x \in [(a-1)/10, a/10) \\ 1 & \text{else} \end{cases}$$

In this setup we see that for each arm a, $\arg\max(\mu_{a,\alpha})$ returns a different α and to uncover the optimal arm the learner must first identify $\alpha : g(\alpha, x) = g^*(x)$. We can expect the two phase strategy of Algorithm 3, to perform poorly in the case where there is a disproportionate difficulty in identifying the optimal α compared to the optimal arm, in the case where the optimal α is known. We see this dynamic in scenario 1, where, identifying the optimal α should be relatively easy compared to identifying the optimal arm. Our experimental results appear to back up our intuition, `TACTIC` clearly outperforms the two stage strategy on scenario 1. On the other hand, we see for scenario 2 the performance of the algorithms is much closer, with neither of the two strategies showing a clear improvement in performance over the other.

We now explicitly state the two stage strategy, used as a competitor for our TACTIC algorithm, we define $\widehat{\nabla}_{\alpha,t}$ and $\widehat{\mu}^t_{a,\alpha}$ as previously.

Algorithm 3: `Two Stage Strategy`

1: **Input:** Budget T
2: Phase 1: Sample each arm in $\mathcal{A}$, $T/2K$ times, with contexts drawn at random from the reservoir.
3: Phase 2: Draw $T/2$ contexts from the reservoir. Let $\hat{\alpha} = \arg\min \widehat{\nabla}_{\alpha,t}$
4: **return** $\hat{a} = \arg\min \widehat{\mu}^t_{a,\hat{\alpha}}$

Conclusion and Perspectives

To the best of our knowledge, this work is the first to consider best arm identification in a contextual bandit model where the learner receives no samples from the target distribution P_T, against which they will be evaluated. We approach the problem in a general setting where the learner has access to auxiliary information to be used in debiasing the source distribution. Specifically the learner has access to known generalised moments of the target distribution. In this setting we classify the problem complexity as a tradeoff between estimating the optimal link function between the source and target distribution and estimating the arm means themselves. We provide an algorithm `TACTIC` , shown to return the optimal arm with probability greater than $1 - \delta$ in all instances, with an upper bound on its expected sampling time. We also demonstrate a lower bound matching up to log terms.

This initial work opens up several interesting questions. Firstly, it is more common in contextual bandits, including contextual bandit setups for transfer learning, to consider the problem of cumulative regret minimisation as opposed to best arm identification. A natural question is to then consider the problem of regret minimisation in our setting. However, we believe a slight reformulation of the problem may be necessary, as otherwise the learner may inevitably pay linear regret due to the time taken to estimate the contexts.

In this paper we make no attempt to be adaptive to the functions $m_1, \ldots, m_L$, essentially assuming a worse case situation where the most informative approach is to simply average over $m_1, \ldots, m_L$ when estimating the optimal parameter function α^*. In practical situations however, it may often be the case that some functions within the set are far more informative than others, or indeed some composite of the m_ls is the most efficient way to estimate α^*.

It is also worth considering the necessity of the smoothness constraint of Assumption 2. It may be feasible to consider a more natural assumption on the class of functions $g(\alpha, x)m(x)$, i.e. a unimodal or log concave assumption.

Our experimental section is purely illustrative, a more thorough evaluation of the empirical performance of the `TACTIC` algorithm, relative to natural competitors remains a task for future work.

Acknowledgments

The work of J. Cheshire is supported by the FMJH, ANR-22-EXES-0013.

References

Ausset, G.; Clémençon, S.; and Portier, F. 2022. Empirical Risk Minimization under Random Censorship. *Journal of Machine Learning Research*, 23(5): 1–59.

Bertail, P.; Clémençon, S.; Guyonvarch, Y.; and Noiry, N. 2021. Learning from Biased Data: A Semi-Parametric Approach. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 803–812. PMLR.

Cai, C.; Cai, T. T.; and Li, H. 2024. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1): 207–232.

Clémençon, S.; Bertail, P.; and Papa, G. 2016. Learning from Survey Training Samples: Rate Bounds for Horvitz-Thompson Risk Minimizers. In Durrant, R. J.; and Kim, K.-E., eds., *Proceedings of The 8th Asian Conference on Machine Learning*, volume 63 of *Proceedings of Machine Learning Research*, 142–157. The University of Waikato, Hamilton, New Zealand: PMLR.

Clémençon, S.; Bertail, P.; and Chautru, E. 2017. Sampling and empirical risk minimization. *Statistics*, 51(1): 30–42.

Deshmukh, A. A.; Sharma, S.; Cutler, J. W.; Moldwin, M.; and Scott, C. 2020. Simple Regret Minimization for Contextual Bandits. arXiv:1810.07371.

Deville, J.-C. 2000. Generalized calibration and application to weighting for non-response. In Bethlehem, J. G.; and van der Heijden, P. G. M., eds., *COMPSTAT*, 65–76. Heidelberg: Physica-Verlag HD.

Garivier, A.; and Kaufmann, E. 2016. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, 998–1027. PMLR.

Guan, M. Y.; and Jiang, H. 2018. Nonparametric Stochastic Contextual Bandits. arXiv:1801.01750.

Huang, J.; Gretton, A.; Borgwardt, K. M.; Schölkopf, B.; and Smola, A. J. 2007. Correcting sample selection bias by unlabeled data. In *NIPS*, 601–608.

Jourdan, M.; Degenne, R.; and Kaufmann, E. 2023. An \varepsilon-Best-Arm Identification Algorithm for Fixed-Confidence and Beyond. *Advances in Neural Information Processing Systems*, 36: 16578–16649.

Kalyanakrishnan, S.; Tewari, A.; Auer, P.; and Stone, P. 2012. PAC subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, 655–662.

Kato, M.; and Ariu, K. 2021. The role of contextual information in best arm identification. *arXiv preprint arXiv:2106.14077*.

Kpotufe, S.; and Martinet, G. 2020. Marginal Singularity, and the Benefits of Labels in Covariate-Shift. arXiv:1803.01833.

Laforgue, P.; and Clémençon, S. 2022. Statistical Learning from Biased Training Samples. *Electronic Journal of Statistics*, 16(2): 6086 – 6134.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Li, A. H.; and Bradic, J. 2020. Censored Quantile Regression Forest. In Chiappa, S.; and Calandra, R., eds., *Proceedings of AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, 2109–2119. PMLR.

Qin, C.; and Russo, D. 2023. Adaptive Experimentation in the Presence of Exogenous Nonstationary Variation. arXiv:2202.09036.

Quionero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. *Dataset Shift in Machine Learning*. The MIT Press.

Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779.

Sugiyama, M.; Nakajima, S.; Kashima, H.; Von Buenau, P.; and Kawanabe, M. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *NIPS*, volume 7, 1433–1440. Citeseer.

Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; von Bünau, P.; Motoaki; and Kawanabe. 2008. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 8(35): 985–1005.

Suk, J.; and Kpotufe, S. 2021. Self-Tuning Bandits over Unknown Covariate-Shifts. In Feldman, V.; Ligett, K.; and Sabato, S., eds., *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, 1114–1156. PMLR.

Wang, M.; Deng, W.; Hu, J.; Tao, X.; and Huang, Y. 2019. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, 692–702.