Bias-Robust Bayesian Optimization via Dueling Bandits

Johannes Kirschner [1]   Andreas Krause [1]

Abstract

We consider Bayesian optimization in settings where observations can be *adversarially biased*, for example by an uncontrolled hidden confounder. Our first contribution is a reduction of the confounded setting to the *dueling bandit* model. Then we propose a novel approach for dueling bandits based on *information-directed sampling (IDS)*. Thereby, we obtain the first efficient kernelized algorithm for dueling bandits that comes with cumulative regret guarantees. Our analysis further generalizes a previously proposed semi-parametric linear bandit model to non-linear reward functions, and uncovers interesting links to doubly-robust estimation.

1. Introduction

Bayesian optimization (Mockus, 1982) is a model-based approach for zero-order global optimization with noisy feedback. It has been successfully applied to many applications such as hyper-parameter tuning of machine learning models, robotics and chemical design. Some variants such as Expected-Improvement (Bull, 2011) or the GP-UCB algorithm (Srinivas et al., 2010) come with theoretical guarantees, ensuring convergence to the global optimum in finite time under suitable regularity assumptions. Closely related is the field of bandit algorithms (Lattimore & Szepesvári, 2020), in particular the linear bandit model (Abe & Long, 1999; Dani et al., 2008; Abbasi-Yadkori et al., 2011).

Most linear bandit algorithms and Bayesian optimization approaches require that the true function is realized in a known reproducing kernel Hilbert space (RKHS) and rely on *unbiased* evaluations of the objective. The regularity assumptions raise questions of robustness to miss-specification and adversarial attacks, and addressing these limitations has been the content of several recent works (Lykouris et al., 2018; Bogunovic et al., 2020b;a).

[1]Department of Computer Science, ETH Zurich. Correspondence to: Johannes Kirschner <jkirschner@inf.ethz.ch>.

We study a setting where the learner's objective is to maximize an unknown function $f : \mathcal{X} \to \mathbb{R}$ with *additive confounded feedback*,

$$y_t = f(x_t) + b_t + \epsilon_t, \tag{1}$$

where x_t is the evaluation point chosen by the learner at time t, ϵ_t is σ^2-sub-Gaussian (zero-mean) observation noise, and b_t is an additive confounding term. We assume that b_t is chosen by an adversary, but does *not* depend on the input x_t. The bias term allows to model the influence of an unobserved and uncontrolled covariate, or a perturbation of the feedback signal imposed by an adversary. One can also interpret (1) as a *contextual* model, where b_t captures the effect of a changing (unobserved) context on the reward. We discuss further examples and applications in Section 1.1 below.

The proposed feedback model (1) generalizes the semi-parametric contextual bandit model studied by Krishnamurthy et al. (2018), where $f(x_t) = \langle x_t, \theta \rangle$ is a linear function defined by a parameter $\theta \in \mathbb{R}^d$. They show that a doubly-robust least-squares estimator allows to recover reward differences $\langle x - x', \theta \rangle$ for inputs $x, x' \in \mathbb{R}^d$ despite the confounding. They further propose an elimination-style algorithm, *bandit orthogonalized semiparametric estimation* (BOSE), which is based on sampling actions from a distribution that minimizes the variance of the estimator. However, finding low variance distributions requires solving a convex-quadratic feasibility problem, which is computationally demanding and in general leads to sampling distributions with support that spans $\mathbb{R}^d$.

Contributions   Our first contribution is two reductions of the confounded feedback model (1) to the *dueling bandit* setting. This allows us to leverage existing algorithms for dueling bandits in the confounded observation setting. We then propose the first efficient algorithm for kernelized dueling bandits that comes with theoretical guarantees on the cumulative regret. The approach is based on *information-directed sampling (IDS)*, which was recently studied in the context of linear partial monitoring by Kirschner et al. (2020b). In particular, we propose an efficient approximation of IDS, that reduces the computation complexity from $\mathcal{O}(|\mathcal{X}|^4)$ to $\mathcal{O}(|\mathcal{X}|)$ on finite action sets $\mathcal{X}$. For continuous action sets, the proposed algorithm requires to optimize a (non-convex) acquisition function over the input space, akin to standard Bayesian optimization.

1.1. Motivating Examples

We start by motivating our problem through applications.

Range-Adjusting Measurement Devices In many real-world optimization tasks, the observed feedback arises from a physical sensing device. Such measurement devices can be subject to calibration errors or might automatically adjust the output range for better sensitivity. For example, in optimization of free electron lasers (Kirschner et al., 2019; Duris et al., 2020), the target signal is measured with a gas-detector, which exploits a physical law to amplify the signal. An input voltage is used to control the amplification factor, which requires re-adjustment[1] with increasing target signal such that the physical relationship between the target and the measured output stays approximately linear[2] (Sorokin et al., 2019; Juranić et al., 2018). Our feedback model allows optimization that is robust to absolute changes in the target signal occurring at any time.

Distributed Optimization of Additive Functions In high-dimensional settings, previous work on Bayesian optimization often uses structural assumptions to reduce the dependence of the sample complexity on the dimension. One popular choice is additivity (Kandasamy et al., 2015), for example coordinate-wise $f(x) = \sum_{i \in [d]} f_i(x^i)$. To optimize additive functions, we can apply d individual learners to optimize each 1-dimensional component f_i separately. Note that the learners cannot directly evaluate $f_i(x_i)$, but only obtain the global noisy feedback $f(x) + \epsilon$, which depends on the choices $x = (x^1, \ldots, x^d)$ of *all* learners. When the learners act in parallel and there is *no communication* possible, the feedback of each learner is confounded as in (1) by the other learner's choices. Our robust approach guarantees that the learners are able to optimize each component successfully, despite the confounding.

Adversarial Attacks Robustness to adversarial attacks was studied recently in the context of bandit algorithms (Lykouris et al., 2018; Bogunovic et al., 2020b) and Bayesian optimization (Bogunovic et al., 2020a). In all previous work that we are aware of, the corruption of the feedback is allowed to depend on the actions, with varying assumptions of whether or not the adversary observes the action choice of the learner. Note that our feedback model is more stringent, as it does not allow for action dependence. On the other hand, our model allows for sublinear regret even with a constant corruption in *every* round, whereas in previous work, the regret scales with the total amount of corruption.

[1] Translating the range-adjusted signal into an absolute value is possible, but not straightforward and comes with other limitations.

[2] More precisely, the relationship depends on the photon energy, pulse intensity, and physical properties of the gas involved.

1.2. Related Work

There is a vast amount of literature on bandit algorithms (Lattimore & Szepesvári, 2020) and Bayesian optimization (Mockus, 1982; Srinivas et al., 2010; Shahriari et al., 2016; Frazier, 2018). Our feedback model is a generalization of the semi-parametric linear bandit setting proposed by Krishnamurthy et al. (2018) with applications for example in mobile health (Tewari & Murphy, 2017). A variant of Thompson sampling was analyzed in the same setting by Kim & Paik (2019), which they show outperforms the BOSE approach by Krishnamurthy et al. (2018), but the frequentist regret bound they derive has an extra factor in the dimension. The Thompson sampling variant is also computationally more efficient, but requires to explicitly compute the probabilities that each action is optimal under the posterior distribution. For Bayesian optimization, robust variants have been considered recently, for example with adversarial perturbations of the input (Bogunovic et al., 2018), corruption of the output (Bogunovic et al., 2020a) and distributionally robust optimization (Kirschner et al., 2020a). In the context of adversarial attacks, there is an increasing body of work (Lykouris et al., 2018; Li et al., 2019; Liu & Lai, 2020; Gupta et al., 2019; Bogunovic et al., 2020a;b), however the feedback model differs from ours, see also the discussion in Section 1.1.

Of particular relevance to our work is the (stochastic) dueling bandit setting (Yue et al., 2012; Sui et al., 2018; Bengs et al., 2021) and kernelized variants (Sui et al., 2017a;b). Early work by Yue & Joachims (2009) applied the dueling bandit model to the optimization setting with continuous action sets and concave reward functions, and established a connection to gradient-based optimization. However, to the best of our knowledge, none of the previous works provide bounds on the cumulative regret in the kernelized (non-concave) setting. A kernelized algorithm with theoretical guarantees that requires point evaluations *and* dueling feedback is by Xu et al. (2020). Closely related is also the work by Saha & Gopalan (2020) on linear dueling bandits with possibly infinite input spaces. They establish a connection to the generalized linear bandit model (GLM) and propose an algorithm which, similar to ours, relies on finding an informative action pair with low regret. However, gap-dependent bounds and a kernelized variant was not provided, and for finite action sets $\mathcal{X}$, their algorithm requires $\mathcal{O}(|\mathcal{X}|^2)$ computation steps per round. Recent work by Agarwal et al. (2021) considers the finite-armed dueling bandit setting with adversarial corruptions of the feedback. Lastly, we remark that most previous work on dueling bandits considers binary feedback, whereas here we are interested in quantitative feedback on the reward-difference between the chosen action pair. Formally, we consider sub-Gaussian dueling feedback which includes the Bernoulli likelihood, but does not exploit the heteroscedasticity of binary observations.

2. Setting

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact input space and $f : \mathcal{X} \to \mathbb{R}$ a fixed and unknown objective function. In each round $t = 1, \ldots, n$, the learner chooses an action $x_t \in \mathcal{X}$ and observes the confounded outcome

$$y_t = f(x_t) + b_t + \epsilon_t \,,$$

where ϵ_t is σ^2-sub-Gaussian, conditionally independent noise, and $b_t \in \mathbb{R}^d$ is an unobserved and possibly time-dependent confounding term. We assume that b_t does not depend on the current action x_t chosen by the learner, and satisfies one of the following assumptions:

a) The bias b_t is bounded, $|b_t| \leq C_{max}$ and fixed at the beginning of round t, but can otherwise arbitrarily depend on $(x_s, y_s)_{s=1}^{t-1}$.

b) The difference between two consecutive bias terms is bounded, $|b_t - b_{t-1}| \leq D_{max}$ and b_t is fixed at the beginning of round $t - 1$, but can otherwise arbitrarily depend on $(x_s, y_s)_{s=1}^{t-2}$.

Which assumption is used is specified in the relevant context. Let $x^* \in \arg\max_{x \in \mathcal{X}} f(x)$ be the optimal action. The suboptimality gap is $\Delta(x) = f(x^*) - f(x)$. The learner's objective is to maximize the cumulative reward $\sum_{t=1}^{n} f(x_t)$, or equivalently minimize the *regret*,

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t) = \sum_{t=1}^{n} \Delta(x_t) \,.$$

For the analysis, we assume that the function f is in a known *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ with associated kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and bounded Hilbert norm $\|f\|_{\mathcal{H}} \leq B$. This is a standard assumption in Bayesian optimization (Srinivas et al., 2010; Chowdhury & Gopalan, 2017) and justifies the use of kernelized least-squares regression, formally introduced in Section 4. We further require that the kernel function is bounded, $k(x, x) \leq 1$.

3. Reduction to Dueling Bandits

It is clear from the observation model (1) that any additive shift of the objective, i.e. $\tilde{f}(x) = f(x) + c$ for $c \in \mathbb{R}$, can be absorbed in the unobserved confounding terms b_t, hence rendering the observation sequences for $f(x)$ and $\tilde{f}(x)$ indistinguishable. In particular, the learner can only hope to recover the true function up to an additive constant. Fortunately, to determine the best action $x^* = \arg\max_{x \in \mathcal{X}} f(x)$, it suffices to estimate *reward differences* $f(x^1) - f(x^2)$ for actions $x^1, x^2 \in \mathcal{X}$, which is indeed possible. To do so, previous work in the linear setting relies on doubly-robust estimation (Krishnamurthy et al., 2018; Kim & Paik, 2019). Here

we take a different approach, and propose a generic reduction of the feedback model (1) to the *dueling bandit* model (Yue et al., 2012; Sui et al., 2018). The reduction has the advantage that we can leverage existing algorithms for dueling bandits, which also eliminates the need to find low-variance sampling distributions for doubly-robust estimation.

In dueling bandits, the learner chooses two actions $x_t^1, x_t^2 \in \mathcal{X}$ and obtains (noisy) feedback on which of the two actions has higher reward. Meanwhile, the learner suffers regret for both actions, but the reward of each action is not observed. While most previous work on dueling bandits uses a binary feedback model, here, we are concerned with a quantitative version of the same setting, which is a special case of *linear partial monitoring* (Lin et al., 2014). Specifically, for a reward function $f : \mathcal{X} \to \mathbb{R}$ and actions $x_t^1, x_t^2 \in \mathcal{X}$, we define *quantitative dueling feedback* as follows:

$$d_t = f(x_t^1) - f(x_t^2) + \xi_t \,, \tag{2}$$

where ξ_t is ρ^2-sub-Gaussian observation noise and ρ is known to the learner. The distribution of ξ_t is allowed to depend on (x_t^1, x_t^2) as long as the sub-Gaussian tail assumption is satisfied uniformly over all actions. Note that this includes binary feedback typically used for dueling bandits, and, more generally, bounded noise distributions. Next, we present two reductions schemes to generate the dueling bandit feedback (2) from confounded observations (1).

Two-Point Reduction The first scheme uses two confounded observations to construct the dueling bandit feedback. Given two inputs $x_t^1, x_t^2 \in \mathcal{X}$ in round t, we evaluate both points according to (1), where the order of evaluation is uniformly randomized. The two observations are

$$y_t^1 = f(x_t^1) + b_{2t+i_t} + \epsilon_{2t+i_t} \,,$$
$$y_t^2 = f(x_t^2) + b_{2t+1-i_t} + \epsilon_{2t+1-i_t} \,,$$

where $i_t \sim \text{Bernoulli}(0.5)$. We then define

$$d_t = y_t^1 - y_t^2 \,. \tag{3}$$

Assuming that b_{2t} and b_{2t+1} are fixed before either of x_t^1, x_t^2 is chosen by the learner and using that the observation noise ϵ_t is zero-mean, one easily confirms that $\mathbb{E}[d_t] = f(x_t^1) - f(x_t^2)$. We further use the following properties of sub-Gaussian random variables. Any random variable X such that $X \in [-B, B]$ is B^2-sub-Gaussian. For two independent random variables X_1, X_2 that are B_1^2- and B_2^2-sub-Gaussian respectively, $X_1 + X_2$ is $(B_1^2 + B_2^2)$-sub-Gaussian. Hence if $|b_{2t} - b_{2t+1}| \leq D_{\max}$, it follows that

$$\xi_t = d_t - \mathbb{E}[d_t]$$
$$= b_{2t+i_t} + \epsilon_{2t+i_t} - (b_{2t+1-i_t} + \epsilon_{2t+1-i_t})$$

is $(D_{\max}^2 + 2\sigma^2)$-sub-Gaussian.

One-Point Reduction Perhaps surprisingly, one can also construct the dueling bandit feedback from a *single* observation using randomization. For given inputs $x_t^1, x_t^2 \in \mathcal{X}$ we choose one point uniformly at random and evaluate the confounded function (1) to obtain a single observation

$$y_t = f(x_t^{(1+i_t)}) + b_t + \epsilon_t \,,$$

where $i_t \sim \text{Bernoulli}(0.5)$. The dueling bandit feedback is

$$d_t = (-1)^{i_t} 2y_t \,. \tag{4}$$

Again, we get an unbiased observation of the reward difference, $\mathbb{E}[d_t] = f(x_t^1) - f(x_t^2)$. Further, if $|b_t| \leq C_{\max}$, then $\xi_t = d_t - \mathbb{E}[d_t]$ is $4(C_{\max}^2 + \sigma^2)$-sub-Gaussian. Compared to the two-point reduction, here the sub-Gaussian variance ρ depends on the absolute value $|b_t|$ of the confounding term instead of the difference $|b_t - b_{t+1}|$. On the other hand, the one-point sampling scheme only requires b_t to be fixed before the choice of x_t, but may depend on *all* previous actions and observations.

4. Information-Directed Sampling

With the reduction to dueling feedback, we are set to readily apply any dueling bandit algorithm in the confounded setting. Furthermore, dueling bandits (as defined in (2)) are a special case of partial monitoring, for which also several algorithms exist. In the following, we adapt the *information-directed sampling* (IDS) approach (Russo & Van Roy, 2014), more specifically, the version proposed by Kirschner et al. (2020b) for linear partial monitoring. The main reason for this choice is that IDS works with *quantitative dueling feedback*, whereas most other work on dueling bandits focuses on settings with Bernoulli likelihood. Also, IDS can be formulated as a *kernelized algorithm*, and comes with *theoretical guarantees* on the regret. However, a direct adaptation of IDS to the dueling setting as proposed by Kirschner et al. (2020b) requires $\mathcal{O}(|\mathcal{X}|^4)$ computational steps per round for finite action sets. In the following, we introduce an approximation of IDS, which obtains the same theoretical guarantees and only requires to optimize a simple score function over the action set. The resulting efficient kernelized dueling bandit algorithm may be of independent interest. On a high level, IDS samples actions from a distribution that minimizes a trade-off between an estimate of the regret and an information gain, as we elaborate below.

Kernel Regression for Dueling Feedback The first step is to set up *kernelized least-squares regression* for the dueling bandit feedback. Recall that we assume that $\mathcal{H}$ is a RKHS with kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq B$. In round t, the learner has already collected data $\mathcal{D}_t = \{(x_s^1, x_s^2, d_s)\}_{s=1}^{t-1}$, where $x_s^1, x_s^2 \in \mathcal{X}$ is the input pair chosen at step s, and d_s is quantitative dueling

bandit feedback defined in Eq. (2). The kernel least-squares estimator with regularizer $\lambda > 0$ is

$$\hat{f}_t = \arg\min_{f \in \mathcal{H}} \sum_{s=1}^{t-1} \left(f(x_s^1) - f(x_s^2) - d_s \right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{5}$$

The solution corresponds to the posterior mean of the Gaussian process model with kernel k and prior variance λ^{-1} and can be computed in closed form. Let $\mathbf{d}_t = [d_1, \ldots, d_{t-1}]^{\top}$ be the vector which collects the observations and define $K_t \in \mathbb{R}^{t-1 \times t-1}$ and $k_t(x) \in \mathbb{R}^{t-1}$ for $x \in \mathcal{X}$ as follows:

$$[K_t]_{ij} \triangleq k(x_i^1, x_j^1) - k(x_i^1, x_j^2) - k(x_i^2, x_j^1) + k(x_i^2, x_j^2)$$
$$[k_t(x)]_i \triangleq k(x, x_i^1) - k(x, x_i^2)$$

The least-squares solution $\hat{f}_t$ evaluated at any $x \in \mathcal{X}$ is $\hat{f}_t(x) = k_t(x)^{\top}(K_t + \lambda \mathbf{1})^{-1} \mathbf{d}_t$, where $\mathbf{1}$ is the identity matrix in the appropriate dimension. To compute uncertainty estimates, we further define

$$k_t(x, y) \triangleq k(x, y) - k_t(x)^{\top}(K_t + \lambda \mathbf{1})^{-1} k_t(y) \,,$$
$$\psi_t(x, z) \triangleq k_t(x, x)^2 + k_t(z, z)^2 - 2k_t(x, z) \,. \tag{6}$$

For any $t \geq 1$, $x_1, x_2 \in \mathcal{X}$, the estimate $\hat{f}_t$ satisfies with probability at least $1 - \delta$,

$$\left| \hat{f}_t(x^1) - \hat{f}_t(x^2) - \left(f(x^1) - f(x^2) \right) \right|^2 \leq \beta_{t,\delta} \psi_t(x^1, x^2) \,, \tag{7}$$

where the *confidence coefficient* is chosen as follows,

$$\beta_{t,\delta}^{1/2} \triangleq \rho \sqrt{\log \det(\mathbf{1} + \lambda^{-1} K_t) + 2 \log \tfrac{1}{\delta}} + \sqrt{\lambda} B \,.$$

The confidence bound is by Abbasi-Yadkori (2012, Theorem 3.11), which improves upon earlier results by Srinivas et al. (2010).

Gap Estimate We use $\hat{f}_t$ to define an estimate $\hat{\Delta}_t(x)$ of the suboptimality gaps $\Delta(x) = f(x^*) - f(x)$ for all $x \in \mathcal{X}$. Let $\hat{x}_t^* = \arg\max_{x \in \mathcal{X}} \hat{f}_t(x)$ be the empirical estimate of the maximizer. Note that $\hat{x}_t^*$ is always defined, despite the fact that in general we can determine $\hat{f}_t$ only up to a constant shift (which does not affect the maximizer). Define

$$\delta_t \triangleq \max_{z \in \mathcal{X}} \hat{f}_t(z) - \hat{f}_t(\hat{x}_t^*) + \beta_{t,\delta}^{1/2} \psi_t(\hat{x}_t^*, z)^{1/2} \,,$$

where $\psi_t(x, z)$ is defined in (6). Intuitively, δ_t is the largest plausible regret that the learner can occur from playing $\hat{x}_t^*$ given the confidence estimates. The *gap estimate* is defined for any $x \in \mathcal{X}$ as follows:

$$\hat{\Delta}_t(x) \triangleq \delta_t + \hat{f}_t(\hat{x}_t^*) - \hat{f}_t(x) \,. \tag{8}$$

The gap estimate satisfies an upper bound on the true gaps, provided that (7) holds, as summarized in the next lemma.

Lemma 1. *With probability $1 - \delta$, for all $x \in \mathcal{X}$ and $t \geq 1$,*

$$\Delta(x) \leq 2\hat{\Delta}_t(x).$$

Proof. With probability at least $1 - \delta$,

$$
\begin{aligned}
\Delta(x) &= f(x^*) - f(x) \\
&= \max_{z \in \mathcal{X}} f(z) - f(\hat{x}_t^*) + f(\hat{x}_t^*) - f(x) \\
&\stackrel{(i)}{\leq} \max_{z \in \mathcal{X}} \hat{f}_t(z) - \hat{f}_t(\hat{x}_t^*) + \beta_{t,\delta}^{1/2} \psi_t(z, \hat{x}_t^*)^{1/2} \\
&\quad + \hat{f}_t(\hat{x}_t^*) - \hat{f}_t(x) + \beta_{t,\delta}^{1/2} \psi_t(\hat{x}_t^*, x)^{1/2} \\
&\stackrel{(ii)}{=} \hat{\Delta}_t(x) + \beta_{t,\delta}^{1/2} \psi_t(\hat{x}_t^*, x)^{1/2}.
\end{aligned}
$$

Here, (i) uses (7) twice and (ii) is by definition of $\hat{\Delta}_t(x)$. On the other hand,

$$
\begin{aligned}
\hat{\Delta}_t(x) &= \max_{z \in \mathcal{X}} \hat{f}_t(z) - \hat{f}_t(x) + \beta_{t,\delta}^{1/2} \psi_t(z, \hat{x}_t^*)^{1/2} \\
&\geq \beta_{t,\delta}^{1/2} \psi_t(\hat{x}_t^*, x)^{1/2}.
\end{aligned}
$$

We used that $\psi_t(\hat{x}_t^*, x) = \psi_t(x, \hat{x}_t^*)$ in the last step. The claim follows with the last two displays combined. $\qquad \square$

Note that this gap estimate is different from the choice proposed by Kirschner et al. (2020b), and importantly, allows us to reduce computational complexity.

Information Gain IDS further requires an information gain function $I_t(x^1, x^2)$ defined for each action, which in our case consists of an input pair, $x^1, x^2 \in \mathcal{X}$. Several choices were proposed by Kirschner et al. (2020b). Here we use the differential of the log-determinant potential,

$$
\begin{aligned}
I_t(x^1, x^2) &\triangleq \log(1 + \psi_t(x^1, x^2)), \qquad (9) \\
&= \log \frac{\det(\mathbf{1} + K_{t+1})}{\det(\mathbf{1} + K_t)}
\end{aligned}
$$

This choice of information gain corresponds to the Bayesian mutual information $\mathbb{I}_t\big(d_t; f | (x_t^1, x_t^2) = (x^1, x^2)\big)$ in the Gaussian process model. In particular, the *total information gain* for (9) resembles the Gaussian entropy

$$\gamma_n = \sum_{t=1}^n I_t(x_t^1, x_t^2) = \log \det(\mathbf{1}_n + K_{n+1}). \qquad (10)$$

The log-determinant depends on the kernel function and upper bounds are known for many popular choices. For example, the linear kernel, $k(x, x') = \langle x, x' \rangle$, satisfies $\gamma_n \leq \mathcal{O}(d \log(n))$ and the RBF kernel, $k(x, x') = \exp(-\|x - x'\|_2^2/2)$, satisfies $\gamma_n \leq \log(n)^{d+1}$ (Srinivas et al., 2010, Theorem 5).

Algorithm 1: Approx. IDS for Dueling Feedback

Input: Action set $\mathcal{X}$, confidence coefficient $\beta_{t,\delta}(\delta)$, `DuelingFeedback`(x_t^1, x_t^2), e.g. via (3) or (4)

1   **for** $t = 1, 2, 3, \dots$ **do**
2     $\hat{x}_t^* \leftarrow \arg\max_{x \in \mathcal{X}} \hat{f}_t(x)$
3     $\hat{\Delta}_t(x) \leftarrow \delta_t + \hat{f}_t(\hat{x}_t^*) - \hat{f}_t(x)$ // Eq. (8)
4     $I_t(x) \leftarrow \log(1 + \psi_t(\hat{x}_t^*, x))$ // Eq. (9)
5     $x_t, p \leftarrow \arg\min_{x \in \mathcal{X}, p \in [0,1]} \dfrac{\big((1-p)\delta_t + p\hat{\Delta}_t(x)\big)^2}{pI_t(x)}$
6     $B_t \sim \text{Bernoulli}(p)$
7   **if** $B_t == 1$, **then**
8       $(x_t^1, x_t^2) \leftarrow (\hat{x}_t^*, x_t)$
9       $d_t \leftarrow$ `DuelingFeedback`(x_t^1, x_t^2)
10   **else**
11       $(x_t^1, x_t^2) \leftarrow (\hat{x}_t^*, \hat{x}_t^*)$ // No feedback

---

(Approximate) Information-Directed Sampling Given the gap estimate $\hat{\Delta}_t(x)$ and information gain $I_t(x^1, x^2)$, we optimize the following trade-off jointly over $\mathcal{X} \times [0, 1]$,

$$z_t, p_t \triangleq \arg\min_{z \in \mathcal{X}, p \in [0,1]} \frac{\big((1-p)\delta_t + p\hat{\Delta}_t(z)\big)^2}{pI_t(\hat{x}_t^*, z)}. \qquad (11)$$

Note that for fixed z, the optimal trade-off probability is $p(z) = \min\left(\frac{\delta_t}{\hat{\Delta}_t(z) - \delta_t}, 1\right)$. Consequently, IDS samples the pair $(\hat{x}_t^*, z_t)$ with probability p_t, and otherwise, with probability $1 - p_t$, the greedy pair $(\hat{x}_t^*, \hat{x}_t^*)$. Note that choosing the same action for dueling feedback provides no information, and in particular the least-squares estimate remains unchanged. The approach is summarized in Algorithm 1.

Computational Complexity As usual, computing the kernel estimates requires to invert the kernel matrix K_t. With incremental updates, the exact kernel estimate and all related quantities can be computed in $\mathcal{O}(d^2 n^3)$ steps in total over n rounds. Note that $\hat{x}_t^*$ and δ_t can be computed in $\mathcal{O}(|\mathcal{X}|)$ steps assuming that all kernel quantities have been pre-computed. The trade-off (11) can be computed in $\mathcal{O}(|\mathcal{X}|)$ steps, since it only requires to evaluate the gap estimates $\hat{\Delta}(x)$ and information gain $I_t(\hat{x}_t^*, x)$ for all $x \in \mathcal{X}$. Therefore the overall complexity is $\mathcal{O}(d^2 n^3 |\mathcal{X}|)$. Of course, Algorithm 1 can also be applied in the linear setting without kernelization, in which case the overall complexity is $\mathcal{O}(d^2 n |\mathcal{X}|)$.

We remark that (11) is an approximation of the IDS trade-off proposed by Kirschner et al. (2020b), which requires to optimize a similar quantity over distributions $\mathscr{P}(\mathcal{X} \times \mathcal{X})$. This is also possible, but the direct implementation suggested in (Kirschner et al., 2020b) requires $\mathcal{O}(|\mathcal{X}|^4)$ compute steps per round to calculate the gap estimates and to find the IDS distribution.

4.1. Regret Bounds

In the language of linear partial monitoring, the dueling feedback (2) is a so-called *locally observable game*, which informally means that any reward difference $f(x) - f(x')$ can be estimated from playing actions which have no more regret than playing either x or x' alone (Kirschner et al., 2020b, Appendix C.5). For dueling bandits, IDS (without the approximation and sampling scheme that we introduce here) has regret at most $R_n \leq \mathcal{O}(\sqrt{n\beta_n\gamma_n})$, see Kirschner et al. (2020b, Corollary 18). Here we show that Algorithm 1 satisfies a similar result. Note that the regret guarantee applies generally to settings with quantitative dueling feedback, where we define regret as follows:

$$R_n^{duel} = \sum_{t=1}^{n} \Delta(x_t^1) + \Delta(x_t^2).$$

Theorem 1. *For ρ^2-sub-Gaussian dueling bandit feedback (2), Algorithm 1 satisfies with probability at least $1 - \delta$,*

$$R_n^{duel} \leq \mathcal{O}\left(\sqrt{n\beta_{n,\delta}(\gamma_n + \log\frac{1}{\delta})}\right).$$

Note that the learner requires knowledge of the sub-Gaussian variance ρ^2 and the Hilbert norm bound $\|f\|_{\mathcal{H}} \leq B$, which appear in the definition of the confidence coefficient $\beta_{n,\delta}$. The regret guarantee has the same scaling as the best known bound for GP-UCB in standard Bayesian optimization (Srinivas et al., 2010; Chowdhury & Gopalan, 2017). Note that $\beta_{n,\delta} = \mathcal{O}\left(\gamma_n + \log\frac{1}{\delta}\right)$, hence combined with bounds for the total information gain γ_n, we can derive bounds for specific choices of the kernel function. For example in the linear setting, we get $R_n^{duel} \leq \mathcal{O}\left(\rho\sqrt{n}(d\log(n) + \log\frac{1}{\delta})\right)$, which is the same as for LinUCB (Abbasi-Yadkori et al., 2011). For the RBF kernel, we get $R_n^{duel} \leq \mathcal{O}\left(\rho\sqrt{n}(\log(n)^{2d+2} + \log\frac{1}{\delta})\right)$.

Applied to the confounded setting with either the one- or two-point reduction, we get the following result. Note that depending on whether the learner requires one or two evaluations per round, the timescale differs by a factor of two.

Corollary 1. *In the confounded setting (1) with σ^2-sub-Gaussian observation noise and dueling feedback obtained via the one-point reduction (4), the regret of Algorithm 1 satisfies with probability at least $1 - \delta$,*

$$R_n \leq \mathcal{O}\left((C_{\max} + \sigma)\sqrt{n\beta_{n,\delta}(\gamma_n + \log\frac{1}{\delta})}\right),$$

assuming that $\max_{t \in [n]} b_t \leq C_{\max}$ and the adversary is allowed to choose b_t depending on all previous actions and observations, $\{x_s, y_s\}_{s=1}^{t-1}$.

With the two-point reduction (3), Algorithm 1 satisfies with probability at least $1 - \delta$,

$$R_n \leq \mathcal{O}\left((D_{\max} + \sigma)\sqrt{n\beta_{n,\delta}(\gamma_n + \log\frac{1}{\delta})}\right),$$

assuming that $\max_{t \in [n]} |b_{2t} - b_{2t+1}| \leq D_{\max}$ and the adversary is allowed to choose b_t depending on all but the last two actions and observations, $\{x_s, y_s\}_{s=1}^{t-2}$.

Note that the algorithm requires knowledge of the bound $C_{\max}$ or $D_{\max}$ respectively, which is needed to compute $\beta_{t,\delta}$. This is in line with the previous work in the linear setting (Krishnamurthy et al., 2018; Kim & Paik, 2019). Removing or weakening this assumption is an interesting direction for future work. Note that in the linear case with the one-point reduction method, our regret bound is $R_n \leq \tilde{\mathcal{O}}(C_{\max}d\sqrt{n})$, which matches the result by Krishnamurthy et al. (2018). The dependence on d and n cannot be improved even in the un-confounded linear bandit setting for general $\mathcal{X}$ (Lattimore & Szepesvári, 2020, Theorem 24.1). The results for the two-point reduction requires a stronger assumption on the sequence $(b_t)_{t=1}^{2n}$, but the regret only depends on the differences $|b_{2t} - b_{2t+1}|$. Hence, the result assures sub-linear regret even for settings where the confounding terms are unbounded, for example when the objective function is subject to drift.

Proof of Theorem 1. First, by Lemma 1 with probability at last $1 - \delta$, $\Delta(x_t) \leq 2\hat{\Delta}_t(x_t)$. We extend the definition of the gap estimate to two points, $\hat{\Delta}_t(x^1, x^2) = \hat{\Delta}_t(x^1) + \hat{\Delta}_t(x^2)$. For a sampling distribution $\mu \in \mathscr{P}(\mathcal{X} \times \mathcal{X})$, denote the expected gap $\hat{\Delta}_t(\mu) \triangleq \mathbb{E}_{x^1, x^2 \sim \mu}[\hat{\Delta}_t(x^1, x^2)]$ and the expected information gain $I_t(\mu) = \mathbb{E}_{x^1, x^2 \sim \mu}[I_t(x^1, x^2)]$. The *information ratio* is defined as follows:

$$\Psi_t(\mu) \triangleq \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)}. \tag{12}$$

Let $(\mu_t)_{t=1}^n$ be the sequence of sampling distributions $\mu_t \in \mathscr{P}(\mathcal{X} \times \mathcal{X})$ defined by Algorithm 1,

$$\mu_t = (1 - p_t)e_{(\hat{x}_t^*, \hat{x}_t^*)} + p_t e_{(\hat{x}_t^*, z_t)},$$

where e_x denotes a Dirac on $x \in \mathcal{X}$. By (Lemma 1, Kirschner et al., 2020b), with probability $1 - \delta$,

$$R_n^{duel} \leq \sqrt{\sum_{t=1}^{n} \Psi_t(\mu_t)\left(\gamma_n + \mathcal{O}\left(\log\frac{1}{\delta}\right)\right)} + \mathcal{O}\left(\log\frac{n}{\delta}\right).$$

The claim in the theorem follows if we show that the information ratio is bounded such that $\Psi_t(\mu_t) \leq \mathcal{O}(\beta_{n,\delta})$. To this end, note that

$$\begin{aligned}
\Psi_t(\mu_t) &= \frac{\left((1-p_t)2\delta_t + p_t(\hat{\Delta}_t(z_t) + \delta_t)\right)^2}{pI_t(\hat{x}_t^*, z_t)} \\
&\leq \min_{x \in \mathcal{X}, p \in [0,1]} \frac{4\left((1-p)\delta_t + p\hat{\Delta}_t(x)\right)^2}{pI_t(\hat{x}_t^*, x)} \\
&\leq \min_{x \in \mathcal{X}} \frac{4\hat{\Delta}_t(x)^2}{I_t(\hat{x}_t^*, x)}, \tag{13}
\end{aligned}$$

where the first inequality follows from $\delta_t \leq \hat{\Delta}_t(x)$ and the definition of z_t and p_t, and the second inequality sets $p = 1$.

On the other hand, using that the kernel is bounded, $k(x, x) \leq 1$, one easily checks that $\psi_t(x^1, x^2) \leq 4$. With $a \leq 3\log(1 + a)$ for all $a \in [0, 4]$ we find for $x^1, x^2 \in \mathcal{X}$

$$\psi_t(x^1, x^2) \leq 3\log(1 + \psi_t(x^1, x^2)) = 3I_t(x^1, x^2). \quad (14)$$

Next, define $\tilde{z}_t = \arg\max_{x \in \mathcal{X}} \hat{f}_t(x) + \beta_{t,\delta}^{1/2}\psi_t(\hat{x}_t, x)^{1/2}$ and observe that $\hat{\Delta}_t(\tilde{z}_t) = \beta_{t,\delta}^{1/2}\psi_t(\hat{x}_t, \tilde{z}_t)^{1/2}$. The claim follows with (13) from noting that

$$\Psi_t(\mu_t) \leq \frac{4\hat{\Delta}_t(\tilde{z}_t)^2}{I_t(\tilde{z}_t)} \leq 12\frac{\beta_{t,\delta}\psi_t(\hat{x}_t, \tilde{z}_t)}{\psi_t(\hat{x}_t, \tilde{z}_t)} = 12\beta_{t,\delta}. \quad (15)$$

The result follows from the fact that $\beta_{t,\delta}$ is monotonically increasing. □

Algorithm 1 also satisfies a gap-dependent bound for finite action sets $\mathcal{X}$. Let $\Delta_{\min} = \min_{x \neq x^*} \Delta(x)$ be the smallest non-zero gap and assume that x^* is unique. The theorem applies to the confounded setting via the reduction method similar to Corollary 1.

Theorem 2. *Assuming that x^* is unique, the regret of Algorithm 1 satisfies with probability at least $1 - \delta$,*

$$R_n^{duel} \leq \mathcal{O}\left(\Delta_{\min}^{-1}\beta_{n,\delta}(\gamma_n + \log\frac{n}{\delta})\right).$$

For linear bandits, the regret bound reads $R_n^{duel} \leq \mathcal{O}(\Delta_{\min}^{-1}d^2\log(n)^2)$ and for the RKHS setting with RBF kernel, the bound is $R_n^{duel} \leq \mathcal{O}(\Delta_{\min}^{-1}\log(n)^{2d+2})$.

Proof of Theorem 2. Our proof uses the strategy introduced by Kirschner et al. (2020c) that relies on finding an instance-dependent bound on the information ratio. Let $\Psi_t(\mu_t)$ be the information ratio defined in (12). We apply (Kirschner et al., 2020b, Lemma 1 with $\eta = 0$) to the scaled information gain $\tilde{I}_t = \beta_{t,\delta}I_t$, to find with probability $1 - \delta$,

$$\sum_{t=1}^{n} \hat{\Delta}_t(x_t^1, x_t^2) \leq \sqrt{\sum_{t=1}^{n} \frac{\Psi_t(\mu_t)}{\beta_{t,\delta}}\left(\beta_{n,\delta}(\gamma_n + \mathcal{O}(\log\frac{n}{\delta}))\right)}$$

We condition now on the event that the previous equation and the confidence estimate in (7) hold simultaneously. As before, let $\tilde{z}_t = \arg\max_{x \in \mathcal{X}} \hat{f}_t(x) + \beta_{t,\delta}^{1/2}\psi_t(\hat{x}_t, x)^{1/2}$. We may assume that $\tilde{z}_t \neq \hat{x}_t$, since otherwise $\delta_t = 0$ and therefore $\Psi_t(\mu_t) = 0$. On the other hand, this implies that $2\hat{\Delta}_t(\tilde{z}_t) \geq \Delta_{\min}$ by (1) and using that x^* is unique.

Next, we reuse the inequality leading to (13) to find

$$\Psi_t(\mu_t) \leq \min_{p \in [0,1]} \frac{4\left((1-p)\delta_t + p\hat{\Delta}_t(\tilde{z}_t)\right)^2}{pI_t(\hat{x}_t, \tilde{z}_t)}.$$

First, consider the case where $2\delta \leq \hat{\Delta}_t(\tilde{z}_t)$. Computing the minimizer of the previous display and using (14), we find

$$\Psi_t(\mu_t) \leq \frac{16\delta_t\hat{\Delta}_t(\tilde{z}_t)}{I_t(\hat{x}_t, \tilde{z}_t)} \leq \frac{48\delta_t\beta_{t,\delta}}{\hat{\Delta}_t(\tilde{z}_t)} \leq \frac{96\delta_t\beta_{t,\delta}}{\Delta_{\min}}.$$

For the other case where $2\delta_t > \hat{\Delta}_t(\tilde{z}_t)$, using (15) directly gives

$$\Psi_t(\mu_t) \leq 12\beta_{t,\delta} \leq \frac{24\delta_t\beta_{t,\delta}}{\hat{\Delta}_t(\tilde{z}_t)} \leq \frac{48\delta_t\beta_{t,\delta}}{\Delta_{\min}}.$$

Finally, note that $\delta_t \leq \frac{1}{2}\hat{\Delta}_t(x_t^1, x_t^2)$. Using the bound on the information ratio, and solving for the regret, we find

$$\sum_{t=1}^{n} \hat{\Delta}_t(x_t^1, x_t^2) \leq \mathcal{O}\left(\Delta_{\min}^{-1}\beta_{n,\delta}(\gamma_n + \log\frac{n}{\delta})\right).$$

The claim follows with Lemma 1 by noting that,

$$R_n^{duel} \leq \sum_{t=1}^{n} 2\hat{\Delta}_t(x_t^1, x_t^2). \qquad \square$$

4.2. A Connection to Doubly-Robust Estimation

In the finite-dimensional, linear case, the objective function is $f(x) = \langle x, \theta \rangle$ for a fixed parameter $\theta \in \mathbb{R}^d$. To obtain an estimate $\hat{\theta}_t$ of the unknown parameter θ directly from confounded data $\{(x_s, y_s = \langle x_s, \theta \rangle + b_s + \epsilon_s)\}_{s=1}^{t-1}$, Krishnamurthy et al. (2018) use a randomized policy $x_t \sim \mu_t$ and a doubly-robust estimation approach. For centered feature vectors $\bar{x}_t = \mathbb{E}_{x \sim \mu_t}[x]$ and regularizer $\lambda > 0$, they define

$$\Gamma_t \triangleq \sum_{s=1}^{t-1}(x_t - \bar{x}_t)(x_t - \bar{x}_t)^\top + \lambda\mathbf{1}_d,$$

$$\hat{\theta}_t \triangleq \Gamma_t^{-1}\sum_{s=1}^{t-1}(x_t - \bar{x}_t)y_t. \quad (16)$$

They further derive the following high-probability bound for doubly-robust estimator:

$$\|\hat{\theta}_t - \theta\|_{\Gamma_t}^2 \leq \mathcal{O}(d\log(n) + \log(n/\delta) + \lambda). \quad (17)$$

Interestingly, when $\mu_t = \text{Uniform}(\{x^1, x^2\})$ is chosen to randomize between two actions $x^1, x^2 \in \mathcal{X}$, then this estimator coincides with the least-squares estimator that we obtain for the dueling bandit feedback. This follows immediately from noting that $2(x_t - \bar{x}_t) = x^1 - x^2$. Further, the concentration bounds are on the same quantity, but the reduction avoids the detour to prove the (more general) concentration bound (17) and leads tighter bounds. We remark that in general, the BOSE algorithm by Krishnamurthy et al. (2018) requires to compute a sampling distributions μ_t supported on $d + 1$ points. We also note that Kim & Paik (2019) propose another variant of the doubly-robust estimator (16). This estimator coincides with our estimation scheme in the same way.
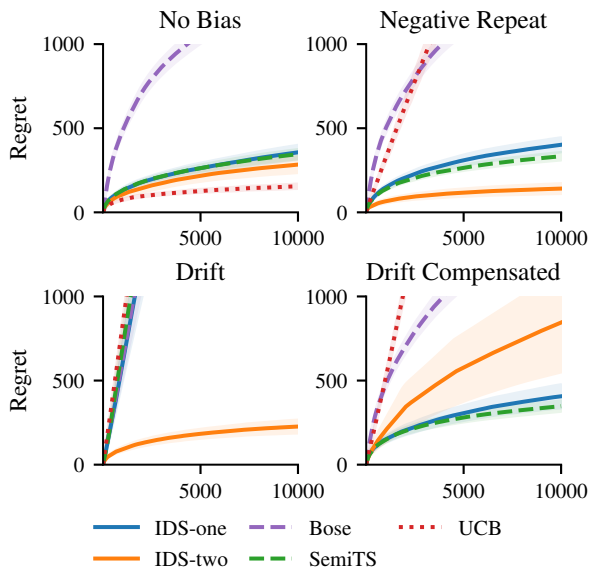
Figure 1. Benchmarks with randomly sampled action sets of size 20 and $d = 4$. The confidence region shows $2\times$ standard error over 50 repetition. Without confounding, UCB performs best, and is closely followed by IDS and Semi-TS. With additive bias the performance of UCB degrades significantly, whereas the robust methods maintain sublinear regret. Note that in the drift experiment (bottom left), the bias is unbounded, which leads to linear regret for all methods expect IDS-two. With bounded bias (right column), IDS-one and Semi-TS show similar performance, whereas the performance of IDS-two is varying. The BOSE algorithm shows sublinear behavior but is much more conservative.

5. Experiments

We evaluate the proposed method with the one-point reduction (IDS-one) and the two-point reduction (IDS-two) in two numerical experiments with confounded observations. To allow a fair comparison with the two-sample scheme, we account for the regret of both evaluations and scale the x-axis appropriately.

5.1. Baselines

UCB For the linear setting, LinUCB (Auer, 2003) is implemented as in (Abbasi-Yadkori et al., 2011, Figure 1) using a regularizer $\lambda = 1$ and confidence coefficient $\beta_{t,\delta}^{1/2} = \sqrt{\log \det V_t + 2 \log \frac{1}{\delta}} + 1$. In the kernelized setting, we use GPUCB (Srinivas et al., 2010) with an empirically tuned confidence coefficient $\beta_{n,\delta} = 1$. As shown by Bogunovic et al. (2020a), increasing the confidence coefficient to a larger value as required in the stochastic setting can lead to better robustness, although we did not see an improvement of performance in our experiments.

BOSE The BOSE algorithm (Krishnamurthy et al., 2018, Algorithm 1) uses the doubly-robust least-squares estimator
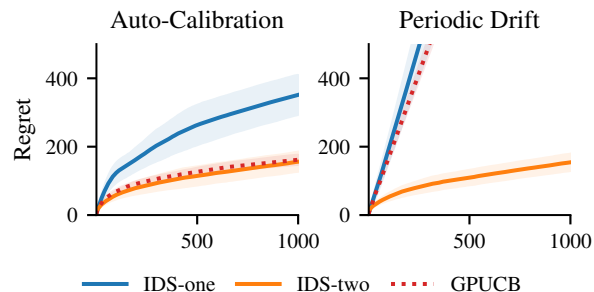

Figure 2. Performance of kernelized algorithms on the camelback benchmark. The confidence region shows $2\times$ standard error over 50 repetitions. With bounded confounding that simulates a re-calibration process (left), GP-UCB is surprisingly competitive with IDS-two, followed by IDS-one. When the feedback is confounded by an unbounded, periodic drift (right), we observe linear regret of GP-UCB and IDS-one (because the bias is unbounded), whereas IDS-two maintains good performance despite the confounding.

(16). We set the required concentration coefficient $\beta_{t,\delta}^{\mathrm{DR}}$ to

$$\beta_{t,\delta}^{\mathrm{DR}} = \sqrt{d \log(1 + t/d) + 2 \log\left(\frac{t}{\delta}\right)} + 1,$$

where we drop (conservative) constants required for the theoretical results in favor of better empirical performance. BOSE requires to solve a convex-quadratic feasibility problem on the space of sampling distributions over the remaining plausible actions, and no specific computation method was recommended by the authors. We compute the sampling distribution by solving the saddle point problem stated in (Krishnamurthy et al., 2018, Appendix D) using exponentiated gradient descent.

SemiTS The semi-parametric Thompson sampling is implemented as in (Kim & Paik, 2019, Algorithm 1), with a less conservative over-sampling parameter $v = \sqrt{2 \log(t/\delta)}$. Our choice improves performance over the theoretical value. We also remark that SemiTS requires to compute the probability of each action being optimal under a Gaussian perturbation of the mean parameter. We do so by computing the empirical sampling probabilities from 1000 random samples per round, the alternative being to compute Gaussian integrals over d-dimensional polytopes $\mathcal{C}_x = \{\nu \in \mathbb{R}^d : \langle x, \nu \rangle \geq \max_{x' \in \mathcal{X}} \langle x', \nu \rangle\}$. While SemiTS is significantly faster than BOSE in our implementation, computing the posterior probabilities accurately for larger action sets remains challenging.

In all experiments we set confidence level $\delta = 0.05$.

5.2. Environments

Linear Reward In the first experiment, we use a linear reward function $f(x) = \langle x, \theta \rangle$. For each repetition we sample $k = 20$ actions uniformly on the $d = 4$ dimensional

unit sphere. We add Gaussian observation noise with variance $\sigma^2 = 1$, that is $\epsilon_t \sim \mathcal{N}(0, 1)$ in (1). In this setting we compare to BOSE, SemiTS and LinUCB (Auer, 2003; Abbasi-Yadkori et al., 2011), where the latter does not directly deal with the confounding. We consider four different types of confounding: *a) no bias; b)* the adversary repeats the last observation with a minus sign, $b_t = -y_{t-1}$, which makes it much harder to identify the best action *(negative repeat); c)* a continues *drift*, $b_t = -0.1t$, i.e. unbounded confounding; and *d)* same as the previous, but with *compensated drift*, $b_t = -0.1t + y_{t-1}$, thereby making the bias terms bounded but dependent on the previous observation.

The result is shown in Figure 1. As expected, in the unconfounded setting UCB works best, followed by both IDS variants and SemiTS with reasonable performance. With confounding, the regret of UCB is increased by a lot, whereas BOSE shows sublinear behaviour but is relatively inefficient. In the example with unbounded bias (drift) only IDS-two performs well, as in fact the theoretical assumptions for all other methods are invalidated. With bounded bias (i.e. negative repeat and compensated drift), SemiTS and IDS-one are competitive, while IDS-two clearly outperforms the baselines in the *negative repeat* experiment.

Camelback Our second experiment is in the non-linear, kernelized setting with observation noise variance $\sigma^2 = 0.1$. As benchmark we choose the camelback function on the domain $[-2, 2] \times [-1, 1]$,

$$f(x_1, x_2) =$$
$$-\min\left(x_1^2\left(4 - 2.1x_1^2 + \frac{x_1^4}{3.}\right) + x_1 x_2 + x_2^2(4x_2^2 - 4), 2.5\right)$$

We discretize the input space using 30 points per dimension. The only direct competitor that we are aware of is the method of Bogunovic et al. (2020a). This method is, however, equivalent to GP-UCB (Srinivas et al., 2010) with an up-scaled confidence coefficient. This suggests that the UCB approach is inherently robust up to a certain degree of corruption, which is also visible in our experiment. For both algorithms, we use an RBF kernel with lengthscale 0.2 and regularizer $\lambda = 1$, and set $\beta_{n,\delta} = 1$ in favor of better empirical performance. We use two types of confounding that we expect is relevant in applications: *a)* a calibration process, which monitors a moving average over the last 10 observations and adjusts the output range to $[-0.1, 0.1]$ whenever the average is no longer in this range; and *b)* periodic drift of the objective, $b_t = \sin(0.2t) - 0.1t$. Results are shown in Figure 2. In the first variant GPUCB works surprisingly well despite the confounding and is on-par with IDS-two. With unbounded drift, both GPUCB and IDS-one obtain linear regret, whereas the performance of IDS-two in unaffected.

6. Conclusion

We introduced randomized evaluation schemes based on pair-wise comparisons that make dueling bandit algorithms applicable to robust optimization with additive confounding. Moreover, we derived a kernelized dueling bandit algorithm based on recent ideas by Kirschner et al. (2020b). The resulting algorithm satisfies worst-case and gap-dependent regret bounds on the cumulative regret and could be of broader interest in the dueling bandit setting. Our numerical experiments validate the theoretical findings.

Acknowledgements

References

Abbasi-Yadkori, Y. *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, 2012.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 3–11, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-612-2.

Agarwal, A., Agarwal, S., and Patil, P. Stochastic dueling bandits with adversarial corruption. In *Algorithmic Learning Theory*, pp. 217–248. PMLR, 2021.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, March 2003. ISSN 1532-4435.

Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A., and Hüllermeier, E. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

Bogunovic, I., Scarlett, J., Jegelka, S., and Cevher, V. Adversarially robust optimization with gaussian processes. *arXiv preprint arXiv:1810.10775*, 2018.

Bogunovic, I., Krause, A., and Scarlett, J. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1071–1081. PMLR, 2020a.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic linear bandits robust to adversarial attacks. *arXiv preprint arXiv:2007.03285*, 2020b.

Bull, A. D. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 2017.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, pp. 355–366. Omnipress, 2008.

Duris, J., Kennedy, D., Hanuka, A., Shtalenkova, J., Edelen, A., Baxevanis, P., Egger, A., Cope, T., McIntire, M., Ermon, S., et al. Bayesian optimization of a free-electron laser. *Physical review letters*, 124(12):124801, 2020.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

Juranić, P., Rehanek, J., Arrell, C. A., Pradervand, C., Ischebeck, R., Erny, C., Heimgartner, P., Gorgisyan, I., Thominet, V., Tiedtke, K., et al. Swissfel aramis beamline photon diagnostics. *Journal of synchrotron radiation*, 25(4):1238–1248, 2018.

Kandasamy, K., Schneider, J., and Póczos, B. High dimensional bayesian optimisation and bandits via additive models. In *International conference on machine learning*, pp. 295–304. PMLR, 2015.

Kim, G.-S. and Paik, M. C. Contextual multi-armed bandit algorithm for semiparametric reward model. In *International Conference on Machine Learning*, pp. 3389–3397, 2019.

Kirschner, J., Mutny, M., Hiller, N., Ischebeck, R., and Krause, A. Adaptive and safe bayesian optimization in high dimensions via one-dimensional subspaces. In *International Conference on Machine Learning*, pp. 3429–3438, 2019.

Kirschner, J., Bogunovic, I., Jegelka, S., and Krause, A. Distributionally robust bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2174–2184. PMLR, 2020a.

Kirschner, J., Lattimore, T., and Krause, A. Information directed sampling for linear partial monitoring. volume 125 of *Proceedings of Machine Learning Research*, pp. 2328–2369. PMLR, 09–12 Jul 2020b.

Kirschner, J., Lattimore, T., Vernade, C., and Szepesvári, C. Asymptotically optimal information-directed sampling. *arXiv preprint arXiv:2011.05944*, 2020c.

Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pp. 2776–2785, 2018.

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

Li, Y., Lou, E. Y., and Shan, L. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Lin, T., Abrahao, B., Kleinberg, R., Lui, J., and Chen, W. Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pp. 901–909, 2014.

Liu, G. and Lai, L. Action-manipulation attacks on stochastic bandits. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3112–3116. IEEE, 2020.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Mockus, J. The bayesian approach to global optimization. *System Modeling and Optimization*, pp. 473–481, 1982.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pp. 1583–1591, 2014.

Saha, A. and Gopalan, A. Regret minimization in stochastic contextual dueling bandits. *arXiv preprint arXiv:2002.08583*, 2020.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2016.

Sorokin, A. A., Bican, Y., Bonfigt, S., Brachmanski, M., Braune, M., Jastrow, U. F., Gottwald, A., Kaser, H., Richter, M., and Tiedtke, K. An x-ray gas monitor for free-electron lasers. *Journal of synchrotron radiation*, 26 (4):1092–1100, 2019.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *International Conference on Machine Learning*, 2010.

Sui, Y., Yue, Y., and Burdick, J. W. Correlational dueling bandits with application to clinical treatment in large decision spaces. *arXiv preprint arXiv:1707.02375*, 2017a.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017b.

Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. Advancements in dueling bandits. In *IJCAI*, pp. 5502–5510, 2018.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pp. 495–517. Springer, 2017.

Xu, Y., Joshi, A., Singh, A., and Dubrawski, A. Zeroth order non-convex optimization with dueling-choice bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 899–908. PMLR, 2020.

Yue, Y. and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208, 2009.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.