Correlational Dueling Bandits with Application to Clinical Treatment in Large Decision Spaces

Yanan Sui, Joel W. Burdick
California Institute of Technology
Pasadena, CA 91125
ysui@caltech.edu, jwb@robotics.caltech.edu

Abstract

We consider sequential decision making under uncertainty, where the goal is to optimize over a large decision space using noisy comparative feedback. This problem can be formulated as a K-armed Dueling Bandits problem where K is the total number of decisions. When K is very large, existing dueling bandits algorithms suffer huge cumulative regret before converging on the optimal arm. This paper studies the dueling bandits problem with a large number of arms that exhibit a low-dimensional correlation structure. Our problem is motivated by a clinical decision making process in large decision space. We propose an efficient algorithm CORRDUELwhich optimizes the exploration/exploitation tradeoff in this large decision space of clinical treatments. More broadly, our approach can be applied to other sequential decision problems with large and structured decision spaces. We derive regret bounds, and evaluate performance in simulation experiments as well as on a live clinical trial of therapeutic spinal cord stimulation. To our knowledge, this marks the first time an online learning algorithm was applied towards spinal cord injury treatments. Our experimental results show the effectiveness and efficiency of our approach.

1 Introduction

In many online learning settings, particularly those that involve human feedback, reliable feedback is often limited to pairwise preferences instead of real valued feedback. Examples include implicit or subjective feedback for information retrieval and recommender systems, such as clicks on search results, and subjective feedback on the quality of recommended care [Chapelle *et al.*, 2012; Sui and Burdick, 2014]. This setup motivates the dueling bandits problem [Yue and Joachims, 2009], which formalizes the problem of online regret minimization via preference feedback (e.g., choosing a pair of arms to be compared at each time step). Many dueling bandits algorithms [Yue and Joachims, 2009; 2011; Zoghi *et al.*, 2014; Ailon *et al.*, 2014; Komiyama *et al.*, 2015; Wu and Liu, 2016] have been developed for efficiently computing this problem with independent arms. However, these algorithms are not efficient in situations involving a large number of dependent arms. Specifically, when the time horizon T is smaller than the number of arms K, it is hopeless to achieve low regret without leveraging structure among arms.

Our problem is motivated by clinical research for recovering motor function after severe spinal cord injury. Previous research [Harkema *et al.*, 2011] has shown that electrical stimulation applied to the spinal cord via electrode arrays implanted in the epidural space over the lumbosacral area enables paralyzed patients to achieve full weight-bearing standing, improvements in stepping, and partial recovery of lost autonomic functions. Stimulation consists of electrical pulse trains applied to selected electrodes. The challenge is that the optimal stimulus pattern (the choice of active electrodes and their polarities, the pulse amplitude and width, and the pulse train frequency) varies significantly across patients. And even for the same patient, the response to the same stimulus has some variation across trials. Hence, clinicians must determine the optimal stimulus for each patient under noisy conditions, which currently a laborious and ad-hoc approach.

Figure 1 shows the clinical treatment procedure for *stand-training* of paraplegics. During a treatment/optimization session, new stimuli are recommended by the algorithm to be applied to the electrode implanted in the patient. The patient then attempts to stand using the given stimuli, and the

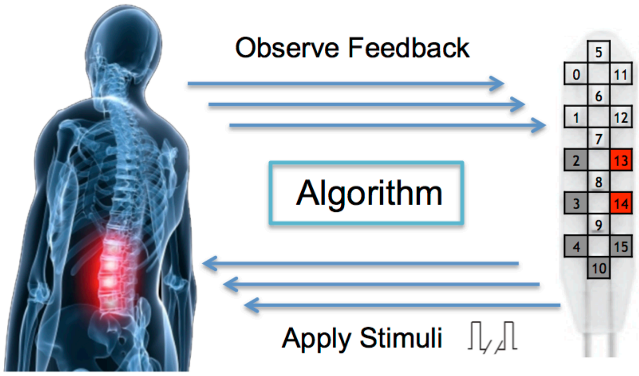

Figure 1: The Standing Experiment under spinal stimulation.

observing clinicians compare the patient's standing performance. The total number of different stimulating configurations is $\sim 4.3 \times 10^7$ due to the complexity of electrodes, and so it is not feasible to search through the whole space. The goal is to develop a algorithm that can automatically select stimuli in order to quickly converge to good treatments.

Motivated by this application, we consider the problem of finding optimal stimuli based on the general setting of the multi-armed bandit problem. The classical bandit problem trades off between exploration and exploitation among a number of different arms, each having a quantifiable but stochastic reward with an initially unknown distribution. In contrast, for our clinical problem, the patient's motor response to stimulation is hard to quantify. Neither video motion capture nor electromyographic (EMG) recordings of muscle activity can yet provide a consistent and satisfactory measure of motor skill under stimulation. One reasonably reliable measure is that of pairwise comparisons, e.g., whether one stimulus more effective than another. While the patient's performance under a specific stimulus is hard to quantify in the clinical setting, we can obtain comparisons of stimuli which are tested within the short time period of one training session.

Our Contributions. In this paper, we show how to cast the problem of online learning of personalized clinical treatment as a dueling bandits problem with a correlated action space, which we call *correlational dueling bandits* . We present an algorithm which meets the demands of such clinical settings, and can effectively model such correlation dependencies to achieve good performance. Our algorithm takes advantage of the correlations among different arms to update the whole active set of arms instead of only updating the two dueling arms. This approach achieves fast convergence to the (near) optimal decisions regardless of the large decision space. We deployed CORRDUELas the first algorithmic approach to the control of spinal cord stimulation in clinical experiments. We find that CORRDUELcan identify a group of optimal stimuli and help paraplegic human patients to achieve full-weight standing.

2 Related Work

2.1 Multi-Armed Bandits

The stochastic multi-armed bandits problem [Robbins, 1952] refers to an iterative decision making problem in which one repeatedly chooses among K options, such as pulling one of K arms of a bandit machine. In each round, we receive a reward that depends on the arm being selected. Without loss of generality, assume that every reward is bounded between $[0, 1]$. The goal then is to minimize the cumulative regret compared to the best arm.

Popular algorithms for the stochastic setting include UCB (upper confidence bound) algorithms [Auer *et al.*, 2002a; Bubeck and Cesa-Bianchi, 2012], and Thompson Sampling [Chapelle and Li, 2011; Russo and Van Roy, 2014].

In the adversarial setting, the rewards are chosen in an adversarial fashion, rather than sampled independently from some underlying distribution. In this case, regret is rephrased as the difference in the sum of rewards. The predominant algorithm for the adversarial setting is EXP3 [Auer *et al.*, 2002b].

2.2 Correlated Bandits

The set of candidate actions is very large (or even infinite) in many applications. When that is the case, one must exploit dependencies between payoffs of different decisions in order to arrive at an efficient algorithm.

In some applications, the underlying problem comes equipped with a correlational structure. Various methods of introducing dependence include bandits on trees [Kocsis and Szepesvári, 2006], bandits with linear correlations [Dani *et al.*, 2008; Abernethy *et al.*, 2008; Abbasi-Yadkori *et al.*, 2011; Gentile *et al.*, 2014] or Lipschitz continuous payoffs [Kleinberg *et al.*, 2008; Bubeck *et al.*, 2008], and Gaussian payoffs [Srinivas *et al.*, 2010].

2.3 Dueling Bandits

Dueling bandits problem [Yue *et al.*, 2012], as a variant of the multi-armed bandits, takes (noisy) comparative feedback instead of real-valued feedback. It is under the general framework of preference learning (learning with preferential feedback). The dueling bandits problem can also be viewed as a special case of partial monitoring problems [Cesa-Bianchi *et al.*, 2006]. Its problem setting naturally fits in with many applications such as information retrievals and recommender systems. The stochastic dueling bandits problem has been extensively studied in [Yue *et al.*, 2012; Ailon *et al.*, 2014; Zoghi *et al.*, 2014; Komiyama *et al.*, 2015; Wu and Liu, 2016].

Beyond the stochastic K-armed dueling bandits setting, other dueling bandit settings include multi-way preference feedback [Sui and Burdick, 2014], continuous-armed convex dueling bandits [Yue and Joachims, 2009], contextual dueling bandits which also introduces the von Neumann winner solution concept [Dudík *et al.*, 2015], sparse dueling bandits that focus on the Borda winner solution concept [Jamieson *et al.*, 2015], Copeland dueling bandits that focus on the Copeland winner solution concept [Zoghi *et al.*, 2015], and adversarial dueling bandits [Gajane *et al.*, 2015]. It would be interesting to study how to extend our analysis to these other settings as well.

3 Problem Statement

In the classical dueling bandits problem, at each iteration t, the following happens:

- The algorithm chooses a pair of actions $b^{(1)}(t)$ and $b^{(2)}(t)$ from a set of K possible actions.
- The algorithm duels $b^{(1)}(t)$ and $b^{(2)}(t)$ and receives (noisy) feedback corresponding to the winner.

Our procedure can be described as follows. There is a set of arms $\mathcal{B} = \{b_1, \cdots, b_K\}$, and a total number of T tests to be performed. At each time step, a pair of arms are chosen from the set $\mathcal{B}$ and a (noisy) comparison of them is observed. T is determined before we run the algorithm. The set of arms are correlated and $T \leq |\mathcal{B}| = K$ in general.

We follow the original notation of the dueling bandit problem. For two arms b_i and b_j sampled from $\mathcal{B}$, we write the comparison factor as

$$\epsilon(b_i, b_j) = P(b_i \succ b_j) - 1/2,$$

where $P(b_i \succ b_j)$ is the probability that b_i dominates b_j and $\epsilon(b_i, b_j) \in [-1/2, 1/2]$ represents the priority between b_i and b_j. We define $b_i \succ b_j \Leftrightarrow \epsilon(b_i, b_j) > 0$. We use the notation $\epsilon_{i,j} \equiv \epsilon(b_i, b_j)$ for convenience. Note that $\epsilon(b_i, b_j) = -\epsilon(b_j, b_i)$ and $\epsilon(b_i, b_i) = 0$. We assume the distribution of reward for each arm is stationary so that all comparison factors converge in [-1/2,1/2]. We also assume $w.l.o.g.$ that the arms are indexed in preferential order $b_1 \succ b_2 \succ \cdots \succ b_K$ so that there is one preferred arm.

Our goal is to minimize the total regret:

$$R_T = \sum_{t=1}^{T} \epsilon(b_1, b^{(1)}(t)) + \epsilon(b_1, b^{(2)}(t))$$

The total regret $R_T = 0$ if we constantly choose $b(t) = b_1$ during the experiment. $R_T = \Theta(T)$ is linear $w.r.t.\ T$ if we constantly choose $b(t) \in \mathcal{B}$.

We also inherit two properties of the comparison factors from the original dueling bandit problem:

Strong Stochastic Transitivity. For any triplet of arms $b_i \succ b_j \succ b_k$, we assume $\epsilon_{i,k} \geq \max\{\epsilon_{i,j}, \epsilon_{j,k}\}$.

Stochastic Triangle Inequality. For any triplet of arms $b_i \succ b_j \succ b_k$, we assume $\epsilon_{i,k} \leq \epsilon_{i,j} + \epsilon_{j,k}$. This can be viewed as a diminishing returns property.

Correlational Dueling Bandits. When the size of the decision set, K, is large, it is unavoidable to carry out a very large number of tests before the algorithm converges to its optimal solution. In some applications like our clinical example, each test is expensive and time-consuming. The number of tests – the time horizon of an algorithm – is often predetermined by clinical conditions. We thus augment the dueling bandits problem into correlational dueling bandits, which takes the correlations among arms into consideration. For any pair of arms $(b_i, b_j) \in \mathcal{B}^2$, we consider the dependence between them are captured by some similarity function $r_{ij} \in [0, 1]$, and it satisfies:

- $r_{ij} = r_{ji}$;
- $r_{ij} = 0 \iff b_i$ and b_j are not correlated;
- $r_{ij} = 1 \iff b_i = b_j$.

For all tuples $(b_i; b_j, b_k) \in \mathcal{B}^3$, if we play pair (b_i, b_j) once and observe $b_i \succ b_j$, we define $\kappa(b_k; b_i, b_j)$ to be the update of wins of b_k and $\tau(b_k; b_i, b_j)$ to be the update of plays of b_k. $\kappa(\cdot; \cdot, \cdot)$ and $\tau(\cdot; \cdot, \cdot)$ represent the dependent structure of the tuple arms. They could be functions of r_{ij}, r_{ik}, and r_{jk}.

In our synthetic experiments, we assume the input space (set of arms) $\mathcal{B}$ has dependent structure and there exists an underlying utility function $f(b) : \mathcal{B} \to \mathbb{R}$ over input space which we cannot observe directly. Our observations are the noisy comparisons between pairs of arms (e.g., b_i and b_j) which can be viewed as the noisy comparison of utility values (e.g., $f(b_i)$ and $f(b_j)$). The properties of strong stochastic transitivity, stochastic triangle inequality, and the dependency assumptions on r_{ij} generally hold for a wide range of applications. In the clinical experiments, we extract comparisons from physician's online judgment.

4 Algorithm

Our algorithm, CORRDUEL as shown in Algorithm 1, is a correlational dueling bandits algorithm based on the Beat-the-Mean algorithm [Yue and Joachims, 2011]. It uses observational feedback and the correlational structure to successively remove suboptimal arms, while keeping the optimal one(s) in the sample space with high probability. The inputs to CORRDUEL are the set of arms $\mathcal{B}$, the total number of iterations T, and the correlational structure (κ, τ).

Parameters-Initialization (Algorithm 2) defines the set of active arms W_ℓ, whose size shrinks as more tests are completed. For each arm b, let n_b be the total number of comparisons between b and other arms, and let w_b be the total number of wins against all other arms. Let $\hat{P}_b$ be the empirical average of $P(b \succ b')$ for all b' in W_ℓ, and let $\hat{P}_{b,n}$ be the value of $\hat{P}_b$ after n comparisons between arm b and any other arms. Set the confidence interval of $P(b \succ b')$ as:

$$\hat{C}_{b,n} = (\hat{P}_{b,n} - c_\delta(n), \hat{P}_{b,n} + c_\delta(n))$$

where $c_\delta(n) = \sqrt{(1/n)log(1/\delta)}$, and δ is the confidence that $P(b \succ b')$ lies in $\hat{C}_{b,n}$. The function $c_\delta(n)$ decreases as the number of comparisons n increases. By properly setting parameter δ, the optimal reward can be reached within the fixed time horizon as shown in Proposition 1.

Active-Elimination (Algorithm 3) is the key part of CORRDUEL. For each pair of tests, two arms are randomly chosen from W_ℓ. The randomized selection method enjoys low-variance total regret in general. For each arm b, the values of w_b, n_b and $\hat{P}_b$ are updated, as is the corresponding confidence radius c^*. An arm b dominates another arm b', if their confidence intervals do not overlap, and the inferior arm is eliminated from W_ℓ. The algorithm runs until the time horizon T is reached, or only one active arm remains.

CORRUPDATE (Algorithm 4) is the subroutine of *Active-Elimination* (Algorithm 3) which updates the weights of b_k by rules $\kappa(\cdot; \cdot, \cdot)$ and $\tau(\cdot; \cdot, \cdot)$. In the classical dueling bandits setting, we assume arms are independent. For independent arms, if we have one comparison between b_i and b_j and gets $b_i \succ b_j$, we only update the weights for arm b_i and b_j:

$$w_i \leftarrow w_i + 1, n_i \leftarrow n_i + 1 \tag{1}$$

$$w_j \leftarrow w_j, n_j \leftarrow n_j + 1 \tag{2}$$

For a large decision space, existing dueling bandits algorithms are extremely slow if one does not exploit dependencies among arms, even if they can achieve provably optimal cumulative regret (w.r.t. independent arms). When the arms are correlated and the correlation between any pair of arms b_i and b_j is measured properly by r_{ij}, we can update all active arms at each iteration.

As shown in Algorithm 4, we update every arm b_k after comparing arms b_i and b_j ($w.l.o.g.$ assume $b_i \succ b_j$) via:

$$w_k \leftarrow w_k + \kappa(b_k; b_i, b_j) \tag{3}$$

Algorithm 1 CORRDUEL

```
Input: $\mathcal{B}, T, (\kappa, \tau)$
Input: $c_\delta(n) = \sqrt{(1/n)log(1/\delta)}$
Run: [Parameters-Initialization]
Run: [Active-Elimination]
return $b^*$   $// Optimal\ arm$
```

Algorithm 2 Parameters-Initialization

```
$W_1 \leftarrow \mathcal{B}$   $// set\ of\ active\ arms$
$\ell \leftarrow 1$   $// rounds$
$\forall b \in W_\ell,\ n_b \leftarrow 0$   $// comparisons$
$\forall b \in W_\ell,\ w_b \leftarrow 0$   $// priorities$
$\forall b \in W_\ell,\ \hat{P}_b \equiv w_b/n_b$, or 1/2 if $n_b = 0$
$n^* \equiv min_{b \in W_\ell} n_b$
$c^* \equiv c_\delta(n^*)$, or 1 if $n^* = 0$   $// confidence\ radius$
$t \leftarrow 0$   $// total\ number\ of\ iterations$
return all new parameters
```

$$n_k \leftarrow n_k + \tau(b_k; b_i, b_j) \tag{4}$$

where $\kappa(\cdot;\cdot,\cdot)$ and $\tau(\cdot;\cdot,\cdot)$ represent the correlational structure, which is assumed to satisfy:

- $0 \le \kappa(b_k; b_i, b_j) \le \tau(b_k; b_i, b_j) \le 1$;
- if $b_k = b_i$, $\kappa(b_k; b_i, b_j) = \tau(b_k; b_i, b_j) = 1$;
- if $b_k = b_j$, $\kappa(b_k; b_i, b_j) = 0$, $\tau(b_k; b_i, b_j) = 1$.

These updates are based on the assumption that $\kappa(\cdot;\cdot,\cdot)$ $\tau(\cdot;\cdot,\cdot)$ is an unbiased estimation of the dependent structure. The CORRUPDATE subroutine (Algorithm 4) can efficiently update all arms at each iteration. So CORRDUEL enjoys fast convergence towards the near optimal arms.

Definition 1. *ε-optimal arm. If arm b satisfies $\epsilon(b_1, b) \le \varepsilon$, then b is an ε-optimal arm.*

Proposition 1. *If $\exists\mu > 0$ such that $\tau(b_k; b_i, b_j) \ge \mu$ for every $(b_i, b_j, b_k) \in \mathcal{B}^3$, then with high probability, the cumulative time to achieve purely ε-optimal arms $T(\varepsilon)$ is bounded by:*

$$T(\varepsilon) = \mathcal{O}\left(\frac{1}{\mu\varepsilon^2}\log\frac{1}{\delta}\right).$$

Proof. Proposition 1 holds based on the Theorem 1 of [Yue and Joachims, 2011]. After t iterations, since $\tau(b_k; b_i, b_j) \ge \mu$, we have $n^* \ge \mu t$. Then $c^* = c_\delta(n^*) = \sqrt{(1/n^*)log(1/\delta)} \le \sqrt{(1/\mu t)log(1/\delta)}$. Notice c^* is a function of time step t.

For any arm b which is not ε-optimal (satisfies $\epsilon(b_1, b) > \varepsilon$), with probability $1 - \delta$, $\hat{P}_{b_1} - \hat{P}_b > \varepsilon C_\delta$ holds for some fixed concentration parameter C_δ. Suppose arm b has not been eliminated at iteration t. Then from elimination criterion Line 7 of Algorithm 3 we have $\varepsilon C_\delta < \hat{P}_{b_1} - \hat{P}_b < 2c^* \le 2\sqrt{(1/\mu t)log(1/\delta)}$. The inequality breaks when $t \ge \frac{4}{\mu\varepsilon^2 C_\delta^2}\log\frac{1}{\delta} = \mathcal{O}\left(\frac{1}{\mu\varepsilon^2}\log\frac{1}{\delta}\right)$. □

Notice, the iteration time $T(\varepsilon)$ in Propositions 1 does not depend on $|\mathcal{B}| = K$, which suggests the fast convergence of CORRDUEL in large decision spaces.

Algorithm 3 Active-Elimination

```
 1: while $|W_\ell| > 1$ and $t \le T$ do
 2:     select $b_i, b_j \in W_\ell$ at random
 3:     compare selected arms (assume $b_i \succ b_j$)
 4:     for all $b_k \in W_\ell$ do
 5:         update $w_k, n_k$ by CORRUPDATE
 6:     end for
 7:     if $\min_{b' \in W_\ell} \hat{P}_{b'} + c^* \le \max_{b \in W_\ell} \hat{P}_b - c^*$ then
 8:         $b' \leftarrow \arg\min_{b \in W_\ell} \hat{P}_b$
 9:         $\forall b \in W_\ell$, delete comparisons with $b'$ from $w_b, n_b$
10:         $W_{\ell+1} \leftarrow W_\ell \backslash \{b'\}$   $// update\ working\ set$
11:         $\ell \leftarrow \ell + 1$   $// new\ round$
12:     end if
13: end while
14: return $b^* = \arg\max_{b \in W_\ell} \hat{P}_b$
```

Algorithm 4 CORRUPDATE

```
Input: $b_k, b_i \succ b_j$
$w_k \leftarrow w_k + \kappa(b_k; b_i, b_j)$
$n_k \leftarrow n_k + \tau(b_k; b_i, b_j)$
return $w_k, n_k$
```

In our application of CORRDUEL to selection of optimal multi-electrode stimulating parameters for paraplegic, we define the similarity of different configurations to be the correlation coefficient of electrical potential fields generated by the two different electrode stimulating configurations. Since the correlation coefficient function $r(\cdot,\cdot)$ has support on $[-1, 1]$, we only update with the CORRUPDATE rule when $r(\cdot,\cdot) \ge 0$. The existence of negative r values is based on clinical observations. The correlational property arises from analysis of electric fields applied by the array as shown in Figure **??**.

The standard notion of correlation coefficient, $r_{XY} = E[XY - E[X]E[Y]]/\sqrt{Var[X]Var[Y]}$, is used in our experiments. However, one can use any measure as a basis for r_{XY} as long as $r_{XY} \in [0, 1]$, $r_{XY} = 1$ when $X = Y$, and $r_{XY} = 0$ when X has an "irrelevant" relation to Y. The coefficient r can take negative values, but the algorithm doesn't use negative values for its updates.

For correlated arms, we perform an update for every arm k for which $r_{ik}, r_{jk} > 0$ as follows:

$$\kappa(b_k; b_i, b_j) \leftarrow \frac{\log r_{jk}}{\log r_{ik} + \log r_{jk}} \cdot \frac{r_{ik} + r_{jk}}{1 + r_{ij}} \tag{5}$$

$$\tau(b_k; b_i, b_j) \leftarrow \frac{r_{ik} + r_{jk}}{1 + r_{ij}} \tag{6}$$

Proposition 2. *If $\exists\mu > 0$ such that $r_{ij} \ge \mu$ for every pair $(b_i, b_j) \in \mathcal{B}^2$, then with high probability, the cumulative time to achieve purely ε-optimal arms $T(\varepsilon)$ satisfies:*

$$T(\varepsilon) = \mathcal{O}\left(\frac{1}{\mu\varepsilon^2}\log\frac{1}{\delta}\right).$$

Proof. If $r_{ij} \ge \mu$ for every pair $(b_i, b_j) \in \mathcal{B}^2$, since $r_{ij} \le 1$, $\tau(b_k; b_i, b_j) = \frac{r_{ik} + r_{jk}}{1 + r_{ij}} \ge \frac{2\mu}{2} = \mu$ for every tuple (b_i, b_j, b_k). The result follows from substituting it into Proposition 1. □

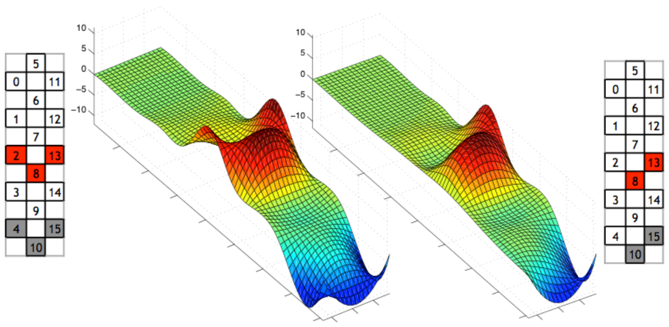

Figure 2: Depiction of the correlations between two different multi-electrode stimulating configurations.

The CORRUPDATE subroutine above updates the dueling pair b_i, b_j in the same way as if they are independent since (5) and (6) will collapse to (1) and (2) for b_i and b_j. For extreme cases, if $b_i \succ b_j$ and arm b_k is very close to b_j. We have $r_{ik} \simeq 1$ and $r_{jk} \simeq r_{ij}$, the updating rules for arm b_k will be close to the updates of arm b_j. If b_k is far from both b_i and b_j, (5) and (6) guarantees that the update for b_k is very small since we acquire little information about b_k from far away comparisons. Also, if b_i and b_j are less dependent (with smaller r_{ij}), we would expect to acquire larger updates for the points in between.

One can also consider a Bayesian version, e.g., by using Gaussian processes. In this paper, we focus on a frequentist approach, which is a better model of the clinical application.

5 Experiments

We evaluated our approach in two settings, synthetic simulations and a real clinical application of online optimization for spinal cord stimulation therapy. In our controlled synthetic experiments, we seek to address the following questions:

- How does the algorithm compare against standard dueling bandit algorithms?
- How effective is it in terms of convergence?

We compare the algorithm against Beat-the-Mean, RUCB, and Sparring algorithm with UCB1. These three algorithms are the representative dueling bandits algorithms designed for independent arms, which do not, however, leverage the correlations between arms.

5.1 Simulation Experiments

Setup. We first evaluate the algorithm with simulation experiments. The purpose of this experiment is to validate our algorithm, and demonstrate its quick convergence when the arms are dependent. To generate the underlying utility function over correlated arms, we sampled random functions from a zero-mean Gaussian Process with squared exponential kernel over the sample space $\mathcal{B} = [0, 1] \times [0, 1]$, uniformly discretized into 50×50 points (set of arms) and used this function as the mean function for the 2500 arms. We chose $\sigma = 0.5$ as the standard deviation of arms. One evaluation of the mean functions is shown in Figure 3. The utility function is not necessarily convex or simple. Within each iteration, we sample 2 points in the active set and compare their (noisy) sampling values to get the $\{0, 1\}$ feedback of the duel. We run the duel for $T = 100$ iterations for 10000 trials for each of the 4 comparing algorithms.

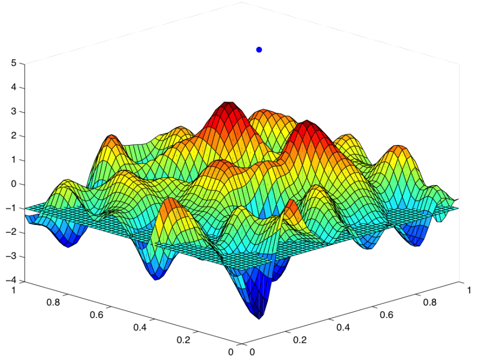

Figure 3: Mean function sampled from a Gaussian process.

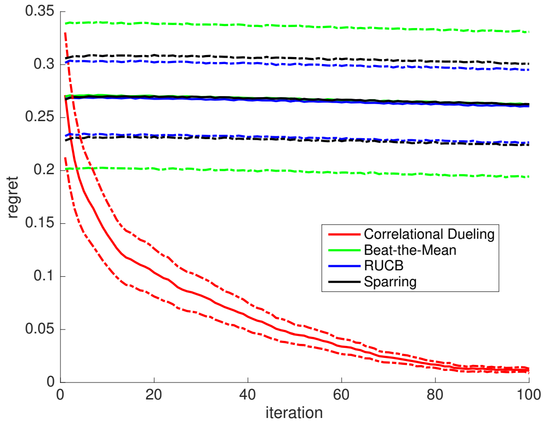

Figure 4: Regret versus iteration. The dashed lines represent one standard deviation.

Results. We report a notion of *regret* as the stepwise regret instead of the cumulative regret. It converges to zero as iteration number goes to infinity for every no-regret algorithm. As seen in Figure 4, CORRDUEL converges much faster than the other three algorithms since it takes the advantage of the dependent arms. The independent-armed dueling bandits algorithms require an exhaustive searching period which is significantly larger than the time horizon we use here before concentrating on the (near) optimal arms.

5.2 Human Experiments

Background. As depicted in Figure 1 from before, our human clinical experiments involve optimizing a system for stand training under spinal cord stimulation with spinal cord injury patients. The subject practices standing under spinal stimulation using a stand frame for assistant in balance. The training processes largely follow the procedures in [Rejc *et al.*, 2015]. Two trainers on the subject's left and right protect and assist the subject. Within each experiment, a specific stimulating pattern (a combination of active electrode selections, the polarity of the actively selected electrodes, and the stimulation amplitude and frequency) is applied through the implanted electrode array and its controlling circuitry. An anonymous short video[1] shows the standing quality under different stimuli. The first part shows a low quality bipedal standing and the second part shows a better standing, both with electrical spinal cord stimulation. Different standings could look similar for the non-specialist.

The participants are under stable medical condition and have no musculoskeletal dysfunction that might interfere with stand training. They have no motor response present in leg muscles during transcranial magnetic stimulation, indicating that there are no strongly active neural pathways connecting cortex and lower limb muscles. No volitional control can be achieved during voluntary movement attempts in leg muscles as measured by EMG activity.

Setup. We use clinical knowledge to restrict the decision space from around 4.3×10^7 to be on the order of $10^3 \sim 10^4$. It is still a very large decision space considering the number of trials, or arm pulls, are on the order of 10^2.

A total of 414 experimental comparisons were done with two patients under the CorrDuel algorithm. Each trial lasted for about 5 minutes. Within each trial, one stimulating pattern was generated by the 16-channel electrode. The patterns were unchanged within each trial. For a fixed electrode configuration, the stimulation frequency and amplitude were modulated synergistically in order to find the best values for effective weight-bearing standing. We optimized the electrode patterns with CorrDuel and performed exhaustive search for stimulation frequency and amplitude over a narrow range.

Stimulation began while the patient was seated. Then the participant initiated the sit to stand transition by positioning his feet shoulder width apart and shifting his weight forward to begin loading the legs.

Results. For the clinical experiments, we cannot create a direct plot for regrets since the ground truth optimal stimulation is unknown. In the experiments, we observed the convergence of CorrDuel, which is not possible for independent-armed dueling bandits algorithms. The set of (near) optimal configurations found by CorrDuel is shown in Figure 5. We compared the performance of CorrDuel to the optimal selections found heuristically for each patient by clinicians, which are shown in Figure 6. We found that the manual selection is a subset of the algorithm's selection, and there exist high performing configurations (e.g., the 2nd in Figure 5) found by the algorithm which are not in the manual selection. This shows that CorrDuel is performing no worse than specialized physicians.

6 Conclusion and Discussion

Our analysis and simulation demonstrate that CorrDuel indeed exhibits fast convergence properties compared to independent-armed dueling bandits algorithms when correlation information is available. We deployed this algorithm in clinical experiments for the control of spinal cord stimulation and showed that CorrDuel performs no worse than specialized physicians. We believe that our result provides an important step towards employing machine learning algorithms in many problems with a large volume of parameter selection and sequential decision making. These problems could be facilitated by our algorithm, which simultaneously delivers effective decisions and explores the decision space based on comparative feedback.

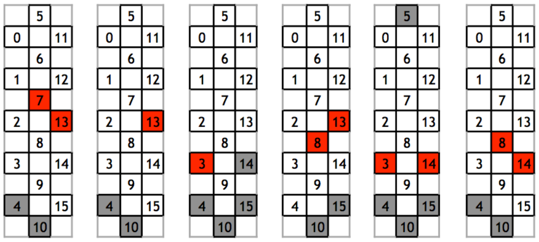

Figure 5: The set of (near) optimal configurations found by the algorithm for a specific patient (in decreasing order in terms of performances).

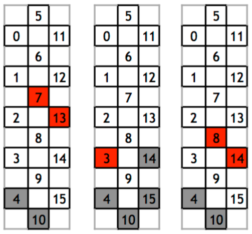

Figure 6: The set of (near) optimal configurations found by physician's manual pick for that specific patient (in decreasing order in terms of performances).

The CorrUpdate subroutine is easy to incorporate with Beat-the-Mean algorithm to achieve efficient CorrDuel. Although we developed CorrDuel specifically based on Beat-the-Mean, CorrUpdate is a more general approach which has potential to incorporate with the existing dueling bandits algorithms. For instance, it can incorporate with RUCB to realize a variant of RUCB for dependent arms by updating the wins w_{ij} with CorrUpdate.

To our knowledge, our work is the first to apply an algorithmic approach towards spinal cord injury treatments. The algorithm could find a proper set of optimal stimulating configurations within the test time horizon. We achieved good performance in both simulations and human experiments. The paraplegic human patients could achieve full-weight standing under the stimulation provided by our algorithm.

Acknowledgements

Yisong Yue (yyue@caltech.edu) is a coauthor of this paper. This research was supported in part by Caltech/JPL PDF IAMS100224, NIH-U01-EB007615-08, NIH-U01-EB015521-05, and a gift from Northrop Grumman.

[1] https://youtu.be/loJLtbcUBDM

References

[Abbasi-Yadkori *et al.*, 2011] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320, 2011.

[Abernethy *et al.*, 2008] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.

[Ailon *et al.*, 2014] Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning (ICML)*, 2014.

[Auer *et al.*, 2002a] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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

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

[Bubeck *et al.*, 2008] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[Cesa-Bianchi *et al.*, 2006] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

[Chapelle and Li, 2011] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[Chapelle *et al.*, 2012] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, 30(1):6:1–6:41, 2012.

[Dani *et al.*, 2008] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, 2008.

[Dudík *et al.*, 2015] Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory (COLT)*, 2015.

[Gajane *et al.*, 2015] Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning (ICML)*, 2015.

[Gentile *et al.*, 2014] Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *ICML*, pages 757–765, 2014.

[Harkema *et al.*, 2011] Susan Harkema, Yury Gerasimenko, Jonathan Hodes, Joel Burdick, Claudia Angeli, Yangsheng Chen, Christie Ferreira, Andrea Willhite, Enrico Rejc, Robert G Grossman, et al. Effect of epidural stimulation of the lumbosacral spinal cord on voluntary movement, standing, and assisted stepping after motor complete paraplegia: a case study. *The Lancet*, 377(9781):1938–1947, 2011.

[Jamieson *et al.*, 2015] Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[Kleinberg *et al.*, 2008] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing (STOC)*. Association for Computing Machinery, Inc., May 2008.

[Kocsis and Szepesvári, 2006] Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *Machine Learning: ECML*, pages 282–293, 2006.

[Komiyama *et al.*, 2015] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *COLT*, 2015.

[Rejc *et al.*, 2015] Enrico Rejc, Claudia Angeli, and Susan Harkema. Effects of lumbosacral spinal cord epidural stimulation for standing after chronic complete paralysis in humans. *PloS one*, 10(7):e0133998, 2015.

[Robbins, 1952] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

[Russo and Van Roy, 2014] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[Srinivas *et al.*, 2010] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, 2010.

[Sui and Burdick, 2014] Yanan Sui and Joel Burdick. Clinical online recommendation with subgroup rank feedback. In *ACM Conference on Recommender Systems (RecSys)*, 2014.

[Wu and Liu, 2016] Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pages 649–657, 2016.

[Yue and Joachims, 2009] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *International Conference on Machine Learning (ICML)*, 2009.

[Yue and Joachims, 2011] Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *International Conference on Machine Learning (ICML)*, 2011.

[Yue *et al.*, 2012] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[Zoghi *et al.*, 2014] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International Conference on Machine Learning (ICML)*, 2014.

[Zoghi *et al.*, 2015] Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pages 307–315, 2015.