Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem

Yisong Yue YYUE@CS.CORNELL.EDU
Thorsten Joachims TJ@CS.CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY 14853 USA

Abstract

We present an on-line learning framework tailored towards real-time learning from observed user behavior in search engines and other information retrieval systems. In particular, we only require pairwise comparisons which were shown to be reliably inferred from implicit feedback (Joachims et al., 2007; Radlinski et al., 2008b). We will present an algorithm with theoretical guarantees as well as simulation results.

1. Introduction

When responding to queries, the goal of an information retrieval system – ranging from web search, to desktop search, to call center support – is to return the results that maximize user utility. So, how can a retrieval system learn to provide results that maximize utility?

The conventional approach is to optimize a proxy-measure that is hoped to correlate with utility. A wide range of measures has been proposed to this effect (e.g., average precision, precision at k, NDCG), but all have similar problems. Most obviously, they require expensive manual relevance judgments that ignore the identity of the user and the user's context. This makes it unclear whether maximization of a proxy-measure truly optimizes the search experience for the user.

We therefore take a different approach based on implicit feedback gathered directly from users. But how can a learning algorithm access the utility a user sees in a set of results? While it is unclear how to reliably derive cardinal utility values for a set of results (e.g. $U(r) = 5.6$), it was shown that interactive experiments can reliably provide ordinal judgments between two sets of results (i.e. $U(r_1) > U(r_2)$) (Joachims et al., 2007; Radlinski et al., 2008b). For example, to elicit whether a user prefers ranking r_1 over r_2, Radlinski et al. (2008b) showed how to present an interleaved ranking of r_1 and r_2 so that clicks indicate which of the two has higher utility. This leads to the following on-line learning problem addressed in this paper.

Given a space of retrieval functions and a (noisy) pairwise test for comparing any two retrieval functions, we wish to find a sequence of comparisons that has low regret (i.e., we eventually find a close to optimal retrieval function and never show clearly bad results in the process). We call this the *Dueling Bandits Problem*, since only ordinal feedback is observable, not cardinal feedback as required by conventional bandit algorithms (e.g., for optimizing web advertising revenue).

In this paper, we formalize the Dueling Bandits Problem and an appropriate notion of regret. Furthermore, we propose a gradient-descent method which builds on methods for on-line convex optimization (Zinkevich, 2003; Kleinberg, 2004; Flaxman et al., 2005). The method is compatible with many existing classes of retrieval functions, and we provide theoretical regret bounds and an experimental evaluation.

2. Related Work

Most prior works on learning from implicit feedback take an off-line approach. Usage logs (containing data such as clicks) are typically transformed into relevance judgments or integrated into the input features (e.g., Agichtein et al., 2006; Carterette & Jones, 2007; Dupret & Piwowarski, 2008). Such approaches are limited to *passive* learning from implicit feedback since they cannot control the initial results presented to users, and thus must use biased training data.

Related on-line methods use *absolute* measures of individual retrieved results (Pandey et al., 2007; Langford & Zhang, 2007; Radlinski et al., 2008a). While theoretical analyses show good regret (as formulated using absolute measures), in many settings such regret

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

formulations might not reflect real user satisfaction. For example, clicks are affected by presentation bias – users tend to click on higher results regardless of relevance (Joachims et al., 2007). Any objective based on absolute measures must use careful calibration. In contrast, the interleaving method proposed by Radlinski et al. (2008b) offers a reliable mechanism for deriving *relative* preferences between retrieval functions.

Algorithm 1 Dueling Bandit Gradient Descent

```
Input: $\gamma$, $\delta$, $w_1$
for query $q_t$ ($t = 1..T$) do
    Sample unit vector $u_t$ uniformly.
    $w'_t \leftarrow \mathbf{P}_{\mathcal{W}}(w_t + \delta u_t)$    //projected back into $\mathcal{W}$
    Compare $w_t$ and $w'_t$
    if $w'_t$ wins then
        $w_{t+1} \leftarrow \mathbf{P}_{\mathcal{W}}(w_t + \gamma u_t)$    //also projected
    else
        $w_{t+1} \leftarrow w_t$
    end if
end for
```

3. The Dueling Bandits Problem

We define a new on-line optimization problem, called the Dueling Bandits Problem, where the only actions are comparisons (or duels) between two points within a space $\mathcal{W}$ (e.g., a parameterized space of retrieval functions in a search engine). We consider the case where $\mathcal{W}$ contains the origin, is compact, convex, and contained in a d-dimensional ball of radius R[1]. Any single comparison between two points w and w' (e.g., individual retrieval functions) is determined independently of all other comparisons with probability

$$P(w \succ w') = \frac{1}{2} + \epsilon(w, w'), \tag{1}$$

where $\epsilon(w, w') \in [-1/2, 1/2]$. In the search example, $P(w \succ w')$ refers to the fraction of users who prefer the results produced by w over those of w'. One can regard $\epsilon(w, w')$ as the distinguishability between w and w'. Algorithms learn only via observing comparison results (e.g., from interleaving (Radlinski et al., 2008b)).

We quantify the performance of an on-line algorithm using the following regret formulation:

$$\Delta_T = \sum_{t=1}^{T} \epsilon(w^*, w_t) + \epsilon(w^*, w'_t), \tag{2}$$

where w_t and w'_t are the two points selected at time t, and w^* is the best point known only in hindsight. Note that the algorithm is allowed to select two identical points, so selecting $w_t = w'_t = w^*$ accumulates no additional regret. In the search example, regret corresponds to the fraction of users who would prefer the best retrieval function w^* over the selected ones w_t and w'_t. A good algorithm should achieve sublinear regret in T, which implies decreasing average regret.

3.1. Modeling Assumptions

We further assume the existence of a differentiable, strictly concave value (or utility) function $v : \mathcal{W} \rightarrow \mathcal{R}$. This function reflects the intrinsic quality of each point in $\mathcal{W}$, and is never directly observed. Since v is strictly concave, there exists a unique maximum $v(w^*)$. Probabilistic comparisons are made using a link function $\sigma : \mathcal{R} \rightarrow [0, 1]$, and are defined as

$$P(w \succ w') = \sigma(v(w) - v(w')).$$

Thus $\epsilon(w, w') = \sigma(v(w) - v(w')) - 1/2$.

Link functions behave like cumulative distribution functions (monotonic increasing, $\sigma(-\infty) = 0$, and $\sigma(\infty) = 1$). We consider only link functions which are rotation-symmetric ($\sigma(x) = 1 - \sigma(-x)$) and have a single inflection point at $\sigma(0) = 1/2$. This implies that $\sigma(x)$ is convex for $x \leq 0$ and concave for $x \geq 0$. One common link function is the logistic function $\sigma_L(x) = 1/(1 + \exp(-x))$.

We finally make two smoothness assumptions. First, σ is L_σ-Lipschitz, and v is L_v-Lipschitz. That is, $|\sigma(a) - \sigma(b)| \leq L_\sigma \|a - b\|$. Thus $\epsilon(\cdot, \cdot)$ is L-Lipschitz in both arguments, where $L = L_\sigma L_v$. We further assume that L_σ and L_v are the least possible. Second, σ is second order L_2-Lipschitz, that is, $|\sigma'(a) - \sigma'(b)| \leq L_2 \|a - b\|$. These relatively mild assumptions provide sufficient structure for showing sublinear regret.

4. Algorithm & Analysis

Our algorithm, Dueling Bandit Gradient Descent (DBGD), is described in Algorithm 1. DBGD maintains a candidate w_t and compares it with a neighboring point w'_t along a random direction u_t. If w'_t wins the comparison, then an update is taken along u_t, and then projected back into $\mathcal{W}$ (denoted by $\mathbf{P}_{\mathcal{W}}$).

DBGD requires two parameters which can be interpreted as the exploration (δ) and exploitation (γ) step sizes. The latter is required for all gradient descent algorithms. Since DBGD probes for descent directions randomly, this introduces a gradient estimation error that depends on δ (discussed Section 4.2). We will show in Theorem 2 that, for suitable δ and γ, DBGD achieves sublinear regret in T,

$$\mathbf{E}[\Delta_T] \leq 2\lambda_T T^{3/4} \sqrt{26RdL},$$

[1] An alternative setting is the K-armed bandit case where $|\mathcal{W}| = K$ (Yue et al., 2009)

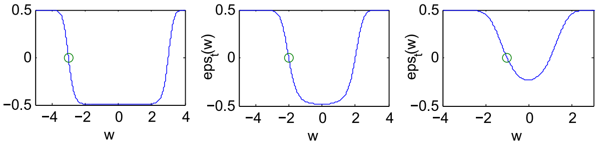

Figure 1. Example relative loss functions ($\epsilon_t(w) \equiv \epsilon(w_t, w)$) using the logistic link function, $\mathcal{W} \subseteq \mathcal{R}$, and value function $v(w) = -w^2$, for $w_t = -3, -2, -1$. Note that the functions are convex in the area around $w^* = 0$.

where λ_T approaches 1 from above as T increases. For example, when $T > \frac{64R^2d^2L_v^4L_2^4}{13^2L^2L_\sigma^4}$, then $\lambda_T < 2$.

Making an additional convexity assumption[2] described in Corollary 2 yields a much simpler result,

$$\mathbf{E}[\Delta_T] \leq 2T^{3/4}\sqrt{10RdL}.$$

To analyze DBGD, we first define relative loss as

$$\epsilon_t(w) \equiv \epsilon(w_t, w), \tag{3}$$

which is the distinguishability between w_t and any other point. We will also define $\epsilon^*(w)$ as

$$\epsilon^*(w) \equiv \epsilon(w^*, w). \tag{4}$$

This relative loss function is depicted pictorally in Figure 1 for the logistic link function and $v(w) = -w^2$.

Analysis Approach. Our analysis follows two conceptual phases. We first present basic results demonstrating the feasibility of performing gradient descent on the relative loss functions ϵ_t (3). These results include proving that ϵ_t is partially convex[3], and how pairwise comparisons can yield good gradient estimates. We then build on existing results (Zinkevich, 2003; Flaxman et al., 2005) to show that DBGD minimizes our regret formulation (2). We begin by observing that ϵ_t is partially convex.

Observation 1. *For link functions $\sigma(x)$ and value functions $v(w)$ satisfying assumptions from Section 3.1, $\epsilon_t(w)$ is partially convex for $w_t \neq w^*$.*

Proof. Define $W_t = \{w : v(w) \geq v(w_t)\}$, which has a non-empty interior for $w_t \neq w^*$. For $a, b \in W_t$ and $\beta \in [0, 1]$ we know that

$$v(\beta a + (1-\beta)b) \geq \beta v(a) + (1-\beta)v(b),$$

since v is concave. We then write $\epsilon_t(\beta a + (1-\beta)b)$ as

$$\begin{aligned}
&= \sigma(v(w_t) - v(\beta a + (1-\beta)b)) - 1/2 \\
&\leq \sigma(v(w_t) - \beta v(a) - (1-\beta)v(b)) - 1/2 \\
&\leq \beta\sigma(v(w_t) - v(a)) + (1-\beta)\sigma(v(w_t) - v(b)) - 1/2 \\
&= \beta\epsilon_t(a) + (1-\beta)\epsilon_t(b)
\end{aligned}$$

The first inequality follows from monotonicity of $\sigma(x)$. The second inequality holds since $\sigma(x)$ is convex for $x \leq 0$ (holds for $a, b \in W_t$). Since W_t is convex (due to concavity of v), we conclude that ϵ_t is partially convex. □

4.1. Estimating Gradients

We now elaborate on the update procedure used by DBGD. Flaxman et al. (2005) observed that

$$\nabla c_t(w_t) \approx \mathbf{E}_u[c_t(w_t + \delta u)u]\frac{d}{\delta}, \tag{5}$$

where $\delta > 0$, d denotes the dimensionality, and u is a uniformly random unit vector. Let $X_t(w)$ denote the event of w winning a comparison with w_t:

$$X_t(w) = \left\{ \begin{array}{ll} 1 & \text{w.p. } 1 - P(w_t \succ w) \\ 0 & \text{w.p. } P(w_t \succ w) \end{array} \right. . \tag{6}$$

We can model the update in DBGD (ignoring γ) as

$$X_t(\mathbf{P}_\mathcal{W}(w_t + \delta u_t))u_t,$$

which we now show, in expectation, matches the RHS of (5) (ignoring d/δ) with an additional projection.

Lemma 1. *Let*

$$c_t(w) = P(w_t \succ w) = \epsilon_t(w) + 1/2.$$

Then for $\delta > 0$ and uniformly random unit vector u,

$$\mathbf{E}_{X_t,u}[X_t(\boldsymbol{P}_\mathcal{W}(w_t + \delta u))u] = -\mathbf{E}_u[c_t(\boldsymbol{P}_\mathcal{W}(w_t + \delta u))u].$$

Proof. Let $\mathbb{S}$ denote the unit sphere. Then we see that $\mathbf{E}_{X_t,u}[X_t(w_t + \delta u)u]$ can be written as

$$\begin{aligned}
&= \mathbf{E}_u[\mathbf{E}_{X_t}[X_t(\mathbf{P}_\mathcal{W}(w_t + \delta u))|u]u] \\
&= \textstyle\int_\mathbb{S} \mathbf{E}_{X_t}[X_t(\mathbf{P}_\mathcal{W}(w_t + \delta u))|u]udu \\
&= \textstyle\int_\mathbb{S}(1 - c_t(\mathbf{P}_\mathcal{W}(w_t + \delta u)))udu \\
&= 0 - \textstyle\int_\mathbb{S} c_t(\mathbf{P}_\mathcal{W}(w_t + \delta u))udu \\
&= -\mathbf{E}_u[c_t(\mathbf{P}_\mathcal{W}(w_t + \delta u))u]
\end{aligned}$$

□

4.2. Gradient Quality & Function Smoothing

We now characterize the quality of the proposed gradient approximation (5). Let $\hat{c}_t$ denote a smoothed version of some function c_t,

$$\hat{c}_t(w) = \mathbf{E}_{x\in\mathbb{B}}[c_t(\mathbf{P}_\mathcal{W}(w + \delta x))],$$

where x is selected uniformly within the unit ball $\mathbb{B}$. We can show using Stokes Theorem that our sampled gradient direction is an unbiased estimate of $\nabla\hat{c}_t$.

[2] The assumption currently lacks theoretical justification, but is observed empirically in many settings.

[3] A function $f : \mathcal{W} \rightarrow \mathcal{R}$ is partially convex if there is a convex region with a non-empty interior and containing w^* where f is convex.

Lemma 2. *Fix $\delta > 0$, over random unit vectors u,*

$$\mathbf{E}_u[c_t(\boldsymbol{P}_{\mathcal{W}}(w + \delta u))u] = \frac{\delta}{d}\nabla \hat{c}_t(w),$$

where d is the dimensionality of x. (Proof analagous to Lemma 2.1 of Flaxman et al., 2005)

Combining Lemma 1 and Lemma 2 implies that DBGD is implicitly performing gradient descent over

$$\hat{\epsilon}_t(w) = \mathbf{E}_{x\in\mathbb{B}}[\epsilon_t(\mathbf{P}_{\mathcal{W}}(w + \delta x))]. \tag{7}$$

Note that $|\hat{\epsilon}_t(w) - \epsilon_t(w)| \leq \delta L$, and that $\hat{\epsilon}_t$ is parameterized by δ (suppressed for brevity). Hence, good regret bounds defined on $\hat{\epsilon}_t$ imply good bounds defined on ϵ_t, with δ controlling the difference.

One concern is that $\hat{\epsilon}_t$ might not be convex at w_t. Observation 1 showed that ϵ_t is convex at w_t, and thus satisfies $\epsilon_t(w_t) - \epsilon_t(w^*) \leq \nabla\epsilon_t(w_t)\cdot(w_t - w^*)$. We now show that $\hat{\epsilon}_t(w_t)$ is "almost convex" in a specific way.

Theorem 1. *For λ defined as*

$$\lambda = \frac{L_\sigma}{L_\sigma - \delta L_v L_2}, \tag{8}$$

and $\delta \in \left(0, \frac{L_\sigma}{L_v L_2}\right)$, then

$$\hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*) \leq \lambda\nabla\hat{\epsilon}_t(w_t)\cdot(w_t - w^*) + (3+\lambda)\delta L.$$

Proof. First define $w_{t,\delta x} \equiv \mathbf{P}_{\mathcal{W}}(w_t + \delta x)$, and also $\epsilon_{t,\delta x}(w) \equiv \epsilon(w_{t,\delta x}, w)$. We rewrite $\hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*)$ as

$$\begin{aligned}
&= \mathbf{E}_{x\in\mathbb{B}}\left[\epsilon_t(\mathbf{P}_{\mathcal{W}}(w_t + \delta x)) - \epsilon_t(\mathbf{P}_{\mathcal{W}}(w^* + \delta x))\right] \\
&\leq \mathbf{E}_{x\in\mathbb{B}}\left[\epsilon_{t,\delta x}(w_{t,\delta x}) - \epsilon_{t,\delta x}(w^*)\right] + 3\delta L \qquad (9)\\
&\leq \mathbf{E}_{x\in\mathbb{B}}\left[\nabla\epsilon_{t,\delta x}(w_{t,\delta x})\cdot(w_{t,\delta x} - w^*)\right] + 3\delta L \qquad (10)
\end{aligned}$$

where (9) follows from ϵ being L-Lipschitz, and (10) follows from $w_{t,\delta x}$ and w^* both being in the convex region of $\epsilon_{t,\delta x}$. Now define $\sigma_t(y) \equiv \sigma(v(w_t) - y)$, and $\sigma_{t,\delta x}(y) \equiv \sigma(v(w_{t,\delta x}) - y)$. We can see that

$$\nabla\epsilon_t(w_{t,\delta x}) = \sigma'_t(v(w_{t,\delta x}))\nabla v(w_{t,\delta x}).$$

and similarly

$$\nabla\epsilon_{t,\delta x}(w_{t,\delta x}) = \sigma'_{t,\delta x}(v(w_{t,\delta x}))\nabla v(w_{t,\delta x}).$$

We can then write (10) as

$$= \mathbf{E}_x\left[\sigma'_{t,\delta x}(w_{t,\delta x})\nabla v(w_{t,\delta x})\cdot(w_{t,\delta x} - w^*)\right] + 3\delta L. \tag{11}$$

We know that both $\sigma'_{t,\delta x}(y) \leq 0$ and $\sigma'_t(y) \leq 0$, and

$$\sigma'_{t,\delta x}(v(w_{t,\delta x})) = -L_\sigma,$$

since that is the inflection point. Thus

$$-L_\sigma \leq \sigma'_t(v(w_{t,\delta x})) \leq -L_\sigma + \delta L_v L_2,$$

which follows from σ being second order L_2-Lipschitz. Since $\epsilon_{t,\delta x}(w_{t,\delta x}) - \epsilon_{t,\delta x}(w^*) \geq 0$, the term inside the expectation in (11) is also non-negative. Using our definition of λ (8), we can write (11) as

$$\begin{aligned}
&\leq \mathbf{E}_x\left[\lambda\sigma'_t(w_{t,\delta x})\nabla v(w_{t,\delta x})\cdot(w_{t,\delta x} - w^*)\right] + 3\delta L \\
&= \mathbf{E}_x\left[\lambda\nabla\epsilon_t(w_{t,\delta x})\cdot(w_{t,\delta x} - w^*)\right] + 3\delta L \\
&= \mathbf{E}_x\left[\lambda\nabla\epsilon_t(w_{t,\delta x})\cdot(w_{t,\delta x} - w_t + w_t - w^*)\right] + 3\delta L \\
&\leq \mathbf{E}_x\left[\lambda\nabla\epsilon_t(w_{t,\delta x})\cdot(w_t - w^*)\right] + (3+\lambda)\delta L \qquad (12)\\
&= \lambda\nabla\hat{\epsilon}_t(w_t)\cdot(w_t - w^*) + (3+\lambda)\delta L
\end{aligned}$$

where (12) follows from observing that

$$\mathbf{E}_x\left[\nabla\epsilon_t(w_{t,\delta x})\cdot(w_{t,\delta x} - w_t)\right] \leq \mathbf{E}_x\left[\|\nabla\epsilon_t(w_{t,\delta x})\|\delta\right] \leq \delta L.$$

□

4.3. Regret Bound for DBGD

Thus far, we have focused on proving properties regarding the relative loss functions ϵ_t and $\hat{\epsilon}_t$. We can easily bound our regret formulation (2) using ϵ_t.

Lemma 3. *Fix $\delta > 0$. Expected regret is bounded by*

$$\mathbf{E}\left[\Delta_T\right] \leq -2\mathbf{E}\left[\sum_{t=1}^{T}\epsilon_t(w^*)\right] + \delta LT.$$

Proof. We can write expected regret as

$$\begin{aligned}
\mathbf{E}\left[\Delta_T\right] &\leq 2\mathbf{E}\left[\textstyle\sum_{t=1}^{T}\epsilon^*(w_t)\right] + \delta LT \\
&= -2\mathbf{E}\left[\textstyle\sum_{t=1}^{T}\epsilon_t(w^*)\right] + \delta LT
\end{aligned}$$

by noting that $|\epsilon^*(w'_t) - \epsilon^*(w_t)| \leq \delta L$, and also that $\epsilon_t(w^*) = -\epsilon^*(w_t)$. □

We now analyze the regret behavior of the smoothed loss functions $\hat{\epsilon}_t$. Lemma 4 provides a useful intermediate result. Note that the regret formulation analyzed in Lemma 4 is different from (2).

Lemma 4. *Fix $\delta \in \left(0, \frac{L_\sigma}{L_v L_2}\right)$, and define λ as in (8). Assume a sequence of smoothed relative loss functions $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_T$ ($\hat{\epsilon}_{t+1}$ depending on w_t) and $w_1, \ldots, w_T \in \mathcal{W}$ defined by $w_1 = 0$ and $w_{t+1} = \mathbf{P}_{\mathcal{W}}(w_t - \eta g_t)$, where $\eta > 0$ and $g_1, \ldots, g_T$ are vector-valued random variables with (a) $\mathbf{E}[g_t|w_t] = \nabla\hat{\epsilon}_t$, (b) $\|g_t\| \leq G$, and (c) $\mathcal{W} \subseteq R\mathbb{B}$. Then for $\eta = \frac{R}{G\sqrt{T}}$,*

$$\mathbf{E}\left[\sum_{t=1}^{T}\hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*)\right] \leq \lambda RG\sqrt{T} + (3+\lambda)\delta T. \tag{13}$$

(Adapted from Lemma 3.1 in Flaxman et al., 2005)

Proof. Theorem 1 implies the LHS of (13) to be

$$
\begin{aligned}
&= \sum_{t=1}^{T} \mathbf{E}\left[\hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*)\right] \\
&\leq \sum_{t=1}^{T} \mathbf{E}\left[\ \lambda \nabla \hat{\epsilon}_t(w_t) \cdot (w_t - w^*) + (3+\lambda)\delta L\ \right] \\
&= \lambda \sum_{t=1}^{T} \mathbf{E}\left[\mathbf{E}[g_t|w_t] \cdot (w_t - w^*)\right] + (3+\lambda)\delta L T \\
&= \lambda \sum_{t=1}^{T} \mathbf{E}[g_t \cdot (w_t - w^*)] + (3+\lambda)\delta L T \qquad (14)
\end{aligned}
$$

Following the analysis of Zinkevich (2003), we will use the potential function $\|w_t - w^*\|^2$. In particular we can rewrite $\|w_{t+1} - w^*\|^2$ as

$$
\begin{aligned}
&= \|\mathbf{P}_{\mathcal{W}}(w_t - \eta g_t) - w^*\|^2 \\
&\leq \|w_t - \eta g_t - w^*\|^2 \qquad (15) \\
&= \|w_t - w^*\|^2 + \eta^2 \|g_t\|^2 - 2\eta(w_t - w^*) \cdot g_t \\
&\leq \|w_t - w^*\|^2 + \eta^2 G^2 - 2\eta(w_t - w^*) \cdot g_t
\end{aligned}
$$

where (15) follows from the convexity of $\mathcal{W}$. Rearranging terms allows us to bound $g_t \cdot (w_t - w^*)$ as

$$
\leq \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 + \eta^2 G^2}{2\eta} \qquad (16)
$$

We can thus bound $\sum_{t=1}^{T} \mathbf{E}[g_t \cdot (w_t - w^*)]$ by

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \mathbf{E}\left[\frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 + \eta^2 G^2}{2\eta}\right] \\
&= \mathbf{E}\left[\frac{\|w_1 - w^*\|^2}{2\eta} + T\frac{\eta^2 G^2}{2\eta}\right] \leq \frac{R^2}{2\eta} + T\frac{\eta G^2}{2} \qquad (17)
\end{aligned}
$$

which follows from choosing $w_1 = 0$ and $\mathcal{W} \subseteq R\mathbb{B}$. Combining (14) and (17) bounds the LHS of (13) by

$$
\leq \lambda\left(\frac{R^2}{2\eta} + T\frac{\eta G^2}{2}\right) + (3+\lambda)\delta T.
$$

Choosing $\eta = \frac{R}{G\sqrt{T}}$ finishes the proof. □

We finally present our main result.

Theorem 2. *By setting* $w_1 = 0$,

$$
\delta = \frac{\sqrt{2Rd}}{\sqrt{13L}T^{1/4}},\ \gamma = \frac{R}{\sqrt{T}},\ T > \left(\frac{\sqrt{2Rd}L_v L_2}{\sqrt{13L}L_\sigma}\right)^4, \qquad (18)
$$

DBGD achieves expected regret (2) bounded by

$$
\mathbf{E}\left[\Delta_T\right] \leq 2\lambda_T T^{3/4}\sqrt{26RdL}
$$

where

$$
\lambda_T = \frac{L_\sigma \sqrt{13L}T^{1/4}}{L_\sigma \sqrt{13L}T^{1/4} - L_v L_2 \sqrt{2Rd}}. \qquad (19)
$$

Proof. Adapting from Flaxman et al. (2005), if we let

$$
g_t = -\frac{d}{\delta} X_t(\mathbf{P}_{\mathcal{W}}(w_t + \delta u_t)) u_t,
$$

using X_t as described in (6), then by Lemma 1 and Lemma 2 we have $\mathbf{E}[g_t|w_t] = \nabla \hat{\epsilon}_t(w_t)$. By restricting T in (18), we guarantee $\delta \in (0, L_\sigma / L_v L_2)$. We can then apply Lemma 4 using the update rule

$$
\begin{aligned}
w_{t+1} &= \mathbf{P}_{\mathcal{W}}(w_t - \eta g_t) \\
&= \mathbf{P}_{\mathcal{W}}(w_t + \eta \tfrac{d}{\delta} X_t(\mathbf{P}_{\mathcal{W}}(w_t + \delta u_t)) u_t)
\end{aligned}
$$

which is exactly the update rule of DBGD if we set $\eta = \gamma\delta/d$. Note that

$$
\|g_t\| = \left\|\frac{d}{\delta} X_t(\mathbf{P}_{\mathcal{W}}(w_t + \delta u_t)) u_t\right\| \leq \frac{d}{\delta}.
$$

Setting $G = d/\delta$ and noting our choice of $\gamma = R/\sqrt{T}$, we have $\eta = \frac{R}{G\sqrt{T}}$. Applying Lemma 4 yields

$$
\mathbf{E}\left[\sum_{t=1}^{T} \hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*)\right] \leq \frac{\lambda R d\sqrt{T}}{\delta} + (3+\lambda)\delta L T. \ (20)
$$

Combining Lemma 3 and (20) yields

$$
\begin{aligned}
\mathbf{E}[\Delta_T] &\leq -2\mathbf{E}\left[\textstyle\sum_{t=1}^{T} \epsilon_t(w^*)\right] + \delta L T \\
&= 2\mathbf{E}\left[\textstyle\sum_{t=1}^{T} \epsilon_t(w_t) - \epsilon_t(w^*)\right] + \delta L T \\
&\leq 2\mathbf{E}\left[\textstyle\sum_{t=1}^{T} \hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*)\right] + 5\delta L T \\
&\leq \tfrac{2\lambda R d\sqrt{T}}{\delta} + (11 + 2\lambda)\delta L T \\
&\leq \lambda\left(\tfrac{2Rd\sqrt{T}}{\delta} + 13\delta L T\right)
\end{aligned}
$$

Choosing $\delta = \frac{\sqrt{2Rd}}{\sqrt{13L}T^{1/4}}$ completes the proof. □

Corollary 1. *Using choices of* w_1, δ, *and* γ *as stated in Theorem 2, if*

$$
T > \left(\frac{\sqrt{2Rd}L_v L_2}{\sqrt{13L}L_\sigma}\right)^4 \left(\frac{1+\alpha}{\alpha}\right)^4,
$$

for $\alpha > 0$, *then*

$$
\mathbf{E}[\Delta_T] \leq 2(1+\alpha)T^{3/4}\sqrt{26RdL}.
$$

The potential non-convexity of $\hat{\epsilon}_t$ significantly complicates the regret bound. By additionally assuming that $\hat{\epsilon}_t$ is convex at w_t (which we have observed empirically in many settings), we arrive at a much simpler result.

Corollary 2. *Assume for all possible* w_t *that* $\hat{\epsilon}_t$ *is convex at* w_t*, which implies*

$$
\hat{\epsilon}_t(w_t) - \hat{\epsilon}_t(w^*) \leq \nabla \hat{\epsilon}_t(w_t) \cdot (w_t - w^*).
$$

Then for $w_1 = 0$, $\delta = \frac{\sqrt{2Rd}}{\sqrt{5L}T^{1/4}}$, *and* $\gamma = \frac{R}{\sqrt{T}}$, *we have*

$$
\mathbf{E}[\Delta_T] \leq 2T^{3/4}\sqrt{10RdL}.
$$

(Proof very similar to Theorem 2 and is omitted)

Table 1. Average regret of DBGD with synthetic functions.

δ_L Factor	0.6	0.8	1	2	3
P_1	0.465	0.398	0.334	**0.303**	0.415
P_2	0.803	0.767	**0.760**	0.780	0.807
P_3	0.687	0.628	**0.604**	0.637	0.663
P_4	0.500	0.378	0.325	**0.304**	0.418
P_5	0.710	**0.663**	0.674	0.798	0.887

4.4. Practical Considerations

Choosing δ to achieve the regret bound stated in Theorem 2 requires knowledge of ϵ_t (i.e., L), which is typically not known in practical settings. The regret bound is indeed robust to the choice of δ. So sublinear regret is achievable using many choices for δ, as we will verify empirically. In the analysis $w_1 = 0$ was chosen to minimize its distance to any other point in $\mathcal{W}$. In certain settings, we might choose $w_1 \neq 0$, in which case our analysis still follows with slightly worse constants.

5. Experiments

5.1. Synthetic Value Functions

We first experimented using synthetic value functions, which allows us to test the robustness of DBGD to different choices of δ. Since L is unknown, we introduced a free parameter δ_L and used $\delta = T^{-1/4}\delta_L\sqrt{0.4Rd}$. We tested on five settings P_1 to P_5. Each setting optimizes over a 50-dimensional ball of radius 10, and uses the logistic transfer function with different value functions that explore a range of curvatures (which affects the Lipschitz constant) and symmetries:

$$v_1(w) = -w^T w, \quad v_2(w) = -|w|$$

$$v_3(w) = -\sum_{i:odd}\left(w^{(i)}\right)^2 - \sum_{i:even}\left|w^{(i)}\right|$$

$$v_4(w) = -\sum_i\left[\exp\left(w^{(i)}\right) + \exp\left(-w^{(i)}\right)\right]$$

$$v_5(w) = v_3(w) - \sum_{i:(i\%3=1)} e^{\left[w^{(i)}\right]_+} - \sum_{i:(i\%3=2)} e^{\left[-w^{(i)}\right]_+}$$

The initial point is $w_1 = \vec{\mathbf{1}}\sqrt{5/d}$. Table 1 shows the regret over the interesting range of δ_L values. Performance degrades gracefully beyond this range. Note that the regret of a random point is about 1 since most points in $\mathcal{W}$ have much lower value than $v(w^*)$.

We also compared against Bandit Gradient Descent (BGD) (Flaxman et al., 2005). Like DBGD, BGD explores in random directions at each iteration. However, BGD assumes access to $P(w_t \succ w)$, whereas DBGD only observes random outcomes. Thus BGD assumes strictly more information[4]. We evaluated two versions: BGD1 using $P(w_t \succ w)$, and BGD2 using $\epsilon_t(w) = P(w_t \succ w) - 1/2$. We expect BGD2 to perform best since the sign of $\epsilon_t(w)$ reveals significant information regarding the true gradient. Figure 2 shows the average regret for problems P_1 and P_5 with $\delta_L = 1$. We observe the behaviors of DBGD and BGD being very similar for both. Interestingly, DBGD outperforms BGD1 on P_5 despite having less information. We also observe this trend for P_2 and P_3, noting that all three problems have significant linear components.

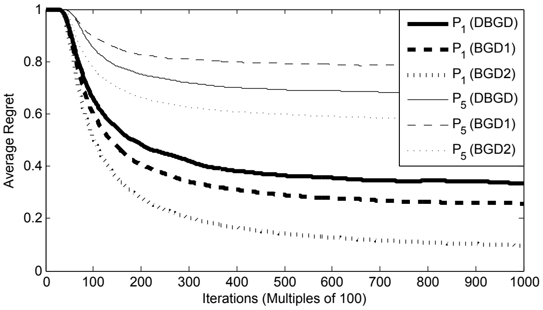

Figure 2. Average regret for $\delta_L = 1$

5.2. Web Search Dataset

For a more realistic simulation environment, we leveraged a real Web Search dataset (courtesy of Chris Burges at Microsoft Research). The idea is to simulate users issuing queries by sampling from queries in the dataset. For each query, the competing retrieval functions will produce rankings, after which the "user" will randomly prefer one ranking over the other; we used a value function based on NDCG@10 (defined below) to determine the comparison outcome probabilities.

We stress that our usage of the dataset is very different from supervised learning settings. In particular, (extensions of) our algorithm might be applied to experiments involving real users where very little is known about each user's internal value function. We leverage this dataset as a reasonable first step for simulating user behavior in an on-line learning setting.

The training, validation and test sets each consist of 1000 queries. We only simulated on the training set, although we measured performance on the other sets to check for, e.g., generalization power. There are about 50 documents per query, and documents are labeled by 5 levels of relevance from 0 (Bad) to 4 (Perfect). The compatibility between a document/query pair is

[4] Our analysis yields matching upper bounds on expected regret for all three methods, though it can be shown that the BGD gradient estimates have lower variance.

Table 2. Average (upper) and Final (lower) NDCG@10 on Web Search training set (sampling 100 queries/iteration)

$\delta \setminus \gamma$	0.001	0.005	0.01	0.05	0.1
0.5	0.524	0.570	0.580	0.569	0.557
0.8	0.533	0.575	0.582	0.576	0.566
1	0.537	0.575	0.584	0.577	0.568
3	0.529	0.565	0.573	0.575	0.571
0.5	0.559	0.591	0.592	0.569	0.565
0.8	0.564	0.593	0.593	0.574	0.559
1	0.568	0.592	0.595	0.582	0.570
3	0.557	0.581	0.582	0.577	0.576

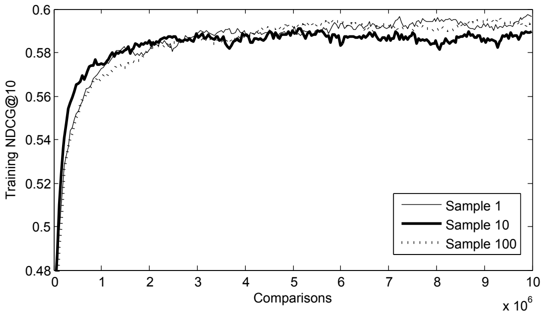

Figure 3. NDCG@10 on Web Search training set

represented using 367 features. A standard retrieval function computes a score for each document based on these features, with the final ranking resulting from sorting by the scores. For simplicity, we considered only linear functions w, so that the score for document x is $w^T x$. Since only the direction of w matters, we are thus optimizing over a 367-dimensional unit sphere.

Our value function is based on Normalized Discounted Cumulative Gain (NDCG), which is a common measure for evaluating rankings (Donmez et al., 2009). For query q, NDCG@K of a ranking for documents of q is

$$\frac{1}{N_K^{(q)}} \sum_{k=1}^{K} \frac{2^{r_k} - 1}{\log(k+1)},$$

where r_k is the relevance level of the kth ranked document, and $N_K^{(q)}$ is a normalization factor[5] such that the best ranking achieves NDCG@K=1. For our experiments, we used the logistic function and 10×NDCG@10 to make probabilistic comparisons.

We note a few properties of this setup, some going beyond the assumptions in Section 3.1. This allows us to further examine the generality of DBGD. First, the value function is now random (dependent on the query). Second, our feasible space $\mathcal{W}$ is the unit sphere and not convex, although it is a well-behaved manifold. Third, we assume a homogenous user group (i.e., all users have the same value function – NDCG@10). Fourth, rankings vary discontinuously w.r.t. document scores, and NDCG@10 is thus a discontinuous value function. We addressed this issue by comparing multiple queries (i.e., delaying multiple iterations) before an update decision, and also by using larger choices of δ and γ. Lastly, even smoothed versions of NDCG have local optima (Donmez et al., 2009), making it difficult to find w^* (which is required for computing regret). We thus used NDCG@10 to measure performance.

We tested DBGD for $T = 10^7$ and a range of γ and δ values. Table 2 shows the average (across all iterations) and final training NDCG@10 when comparing 100 queries per update. Performance peaks at $(\delta, \gamma) = (1, 0.01)$ and degrades smoothly. We found similar results when varying the number of queries compared per update. Figure 3 depicts per iteration NDCG@10 for the best models when sampling 1, 10 and 100 queries. Making multiple comparisons per update has no impact on performance (the best parameters are typically smaller when sampling fewer queries). Sampling multiple queries is very realistic, since a search system might be constrained to, e.g., making daily updates to their ranking function. Performance on the validation and test sets closely follows training set performance (so we omit their results). This implies that our method is not overfitting.

For completeness, we compared our best DBGD models with a ranking SVM, which optimizes over pairwise document preferences and is a standard baseline in supervised learning to rank settings. More sophisticated methods (e.g., Chakrabarti et al., 2008; Donmez et al., 2009) can further improve performance. Table 3 shows that DBGD approaches ranking SVM performance despite making fundamentally different assumptions (e.g., ranking SVMs have access to very specific document-level information). We caution against over-optimizing here, and advocate instead for developing more realistic experimental settings.

6. Conclusion

We have presented an on-line learning framework based on pairwise comparisons, and naturally fits with recent work on deriving reliable pairwise judgments. Our proposed algorithm, DBGD, achieves sublinear regret. As evidenced by our simulations based on web data, DBGD can be applied much more generally than suggested by our theoretical analysis. Hence, it begs for more sophisticated formulations which account for properties such as heterogenous user behavior, query dependent value functions, and the discontinuity of

[5] Note that $N_K^{(q)}$ will be different for different queries.

Table 3. Comparing Ranking SVM vs. final DBGD model using average NDCG@10 and per-query win/tie/loss counts.

Model	SVM	Sample 1	Sample 5	Sample 10	Sample 25	Sample 50	Sample 100
NDCG@10	0.612	0.596	0.593	0.589	0.593	0.596	0.595
W/T/L	–	490/121/389	489/121/390	504/118/378	489/118/393	472/119/409	490/116/394

rankings. Another interesting direction is adaptively choosing δ and γ for any-time regret analyses.

Our framework is extendable in many ways, such as integrating pairwise document preferences (Joachims et al., 2007; Carterette et al., 2008), and diversity (Yue & Joachims, 2008; Radlinski et al., 2008a). Progress in this area can lead to cost-effective systems for a variety of application domains such as personalized search, enterprise search, and also small interest groups.

Acknowledgements

The work was funded under NSF Award IIS-0713483, NSF CAREER Award 0237381, and a gift from Yahoo! Research. The first author is also partly funded by a Microsoft Research Graduate Fellowship and a Yahoo! Key Technical Challenges Grant. The authors also thank Robert Kleinberg, Josef Broder and the anonymous reviewers for their helpful comments.

References

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving Web Search Ranking by Incorporating User Behavior Information. *ACM Conference on Information Retrieval (SIGIR)* (pp. 19–26).

Carterette, B., Bennett, P., Chickering, D. M., & Dumais, S. (2008). Here or There: Preference Judgments for Relevance. *European Conference on Information Retrieval (ECIR)* (pp. 16–27).

Carterette, B., & Jones, R. (2007). Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks. *Neural Information Processing Systems (NIPS)* (pp. 217–224).

Chakrabarti, S., Khanna, R., Sawant, U., & Battacharyya, C. (2008). Structured Learning for Non-Smooth Ranking Losses. *ACM Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 88–96).

Donmez, P., Svore, K., & Burges, C. (2009). On the Local Optimality of LambdaRank. *ACM Conference on Information Retrieval (SIGIR).*

Dupret, G., & Piwowarski, B. (2008). A User Browsing Model to Predict Search Engine Click Data from Past Observations. *ACM Conference on Information Retrieval (SIGIR)* (pp. 331–338).

Flaxman, A., Kalai, A., & McMahan, H. B. (2005). Online Convex Optimization in the Bandit Setting: Gradient Descent Without a Gradient. *ACM-SIAM Symposium on Discrete Algorithms (SODA)* (pp. 385–394).

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Transactions on Information Systems (TOIS)*, *25*, 7:1–26.

Kleinberg, R. (2004). Nearly tight bounds for the continuum-armed bandit problem. *Neural Information Processing Systems (NIPS)* (pp. 697–704).

Langford, J., & Zhang, T. (2007). The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. *Neural Information Processing Systems (NIPS)* (pp. 817–824).

Pandey, S., Agarwal, D., Chakrabarti, D., & Josifovski, V. (2007). Bandits for Taxonomies: A Model-based Approach. *SIAM Conference on Data Mining (SDM)* (pp. 216–227).

Radlinski, F., Kleinberg, R., & Joachims, T. (2008a). Learning Diverse Rankings with Multi-Armed Bandits. *International Conference on Machine Learning (ICML)* (pp. 784–791).

Radlinski, F., Kurup, M., & Joachims, T. (2008b). How Does Clickthrough Data Reflect Retrieval Quality? *ACM Conference on Information and Knowledge Management (CIKM)* (pp. 43–52).

Yue, Y., Broder, J., Kleinberg, R., & Joachims, T. (2009). The K-armed Dueling Bandits Problem. *Conference on Learning Theory (COLT).*

Yue, Y., & Joachims, T. (2008). Predicting Diverse Subsets Using Structural SVMs. *International Conference on Machine Learning (ICML)* (pp. 1224–1231).

Zinkevich, M. (2003). Online Convex Programming and Generalized Infinitesimal Gradient Ascent. *International Conference on Machine Learning (ICML)* (pp. 928–936).