

Variance Reduction in Gradient Exploration for Online Learning to Rank

Huazheng Wang, Sonwoo Kim, Eric McCord-Snook, Qingyun Wu, Hongning Wang

Department of Computer Science, University of Virginia

Charlottesville, VA 22904, USA

{hw7ww,sak2km,esm7ky,qw2ky,hw5x}@virginia.edu

ABSTRACT

Online Learning to Rank (OL2R) algorithms learn from implicit user feedback on the fly. The key to such algorithms is an unbiased estimate of gradient, which is often (trivially) achieved by uniformly sampling from the entire parameter space. Unfortunately, this leads to *high-variance* in gradient estimation, resulting in high regret during model updates, especially when the dimension of the parameter space is large.

In this work, we focus on Dueling Bandit Gradient Descent (DBGD) based OL2R algorithms, which constitute a major endeavor in this direction of research. In particular, we aim at reducing the variance of gradient estimation in DBGD-type OL2R algorithms. We project the selected updating direction (i.e., the winning direction) into a space spanned by the feature vectors from examined documents under the current query (termed the “document space” for short), *after* an interleaved test. Our key insight is that the result of an interleaved test is solely governed by a user’s relevance evaluation over the *examined* documents. Hence, the true gradient introduced by this test is only reflected in the constructed document space, and components of the proposed gradient which are orthogonal to the document space can be safely removed, for variance reduction purpose. We prove that this projected gradient is still an unbiased estimation of the true gradient, and show that this lower-variance gradient estimation results in significant regret reduction. Our proposed method is compatible with all existing DBGD-type OL2R algorithms which rank documents using a linear model. Extensive experimental comparisons with several best-performing DBGD-type OL2R algorithms have confirmed the effectiveness of our proposed method in reducing the variance of gradient estimation and improving overall ranking performance.

CCS CONCEPTS

• Information systems → Learning to rank; • Theory of computation → Online learning algorithms;

KEYWORDS

Online learning to rank; Dueling bandit; Variance Reduction

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '19, July 21–25, 2019, Paris, France

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6172-9/19/07...\$15.00

<https://doi.org/10.1145/3331184.3331264>

ACM Reference Format:

Huazheng Wang, Sonwoo Kim, Eric McCord-Snook, Qingyun Wu, Hongning Wang. 2019. Variance Reduction in Gradient Exploration for Online Learning to Rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*, July 21–25, 2019, Paris, France. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3331184.3331264>

1 INTRODUCTION

Online Learning to Rank (OL2R) [6] is a family of online learning solutions, which exploit implicit feedback from users to directly optimize parameterized rankers on the fly. It has drawn increasing attention in research community in recent years due to its advantages over classical offline learning to rank algorithms [12]. First, it avoids the expensive and time consuming process of offline result relevance annotation. Second, as it directly learns from user feedback, it optimizes the ranking results to best reflect current user preferences [17]. Third, because the model is updated on the fly, there is no need to store user click history offline, which alleviates many privacy concerns [23].

One strain of OL2R algorithms, represented by Dueling Bandit Gradient Descent (DBGD) [26], optimize a linear scoring function by exploring the parameter space via interleaved test. Algorithms of this type first propose an exploratory direction as a tentative model update direction, and then update the current ranker if the proposed direction provides better ranking utility. In practice, result utility is usually inferred from user clicks on an interleaved list of ranking results from each ranker [25]. The key technical insight of DBGD-type algorithms is that the expectation of selected directions is an *unbiased* estimate of true gradient of the unknown loss function for ranking [5]. As a result, DBGD is essentially a stochastic online gradient descent algorithm. However, because the exploration directions are uniformly sampled from the entire parameter space, when the dimensionality of the space is high (which is usually the case in practice), the *variance* in gradient estimation becomes large. This directly slows down the learning convergence of the algorithm and inevitably increases sample complexity.

Recently, several follow-up works have realized this deficiency of gradient exploration in DBGD, and propose various types of solutions to improve its learning efficiency. One type of studies explore multiple random directions in each iteration of model update. Unbiased estimate of gradient is maintained in this type of revisions of DBGD, as the directions are still uniformly sampled. Model estimation variance is expected to be reduced by testing more exploratory directions; but, in practice, as the users would only examine a finite number of documents under each query (e.g., due to position bias [9]), the sensitivity of interleaved test drops as a result of more exploratory rankers having to be tested at once.

This unfortunately introduces additional variance in model estimation. Another type of research constrains the sampling space for gradient exploration [7, 14, 22]. However, this line of solutions cannot guarantee the estimated gradient remains unbiased, and thus face high risk of converging towards a sub-optimal solution.

Although empirically effective, previous OL2R solutions neglect an important property of click-based result utility evaluation: users only perceive utility from the documents that they actually examine. As a result, the *true* gradient is only revealed by features playing an essential role in ranking those examined documents under this query. Here we define essential features in ranking a particular set of documents as those features with non-zero variance among the documents. Assume in an interleaved test, one ranking feature takes a constant value in all examined documents under this query, such that it has no effect in differentiating the quality of those documents. Then, the proposed exploratory direction’s contribution to the ranker update on this particular dimension cannot be justified by this test result. Random gradient exploration hence introduces an arbitrary update on this dimension, which inevitably leads to high estimation variance over time. This example can be generalized to situations where multiple (even correlated) features have no effect in differentiating the utility of examined documents in the result of an interleaved test. Because in practice users usually only examine a handful of documents under each query [4, 9], but each document consists of hundreds or even thousands of ranking features, the variance introduced by random exploration on those non-essential features could be considerably large.

The above analysis suggests that an interleaved test only reveals the projection of true gradient in the spanned space of examined documents under a test query (termed the “document space” in this paper). With this as our motivation, we decide to project the winning direction back into the document space so as to reduce the variance introduced by random gradient exploration. We construct the document space from inferred users’ result examinations [4], which are not observable in the user response but can be statistically modeled. Because this projection is independent from how the proposal directions are created, this solution can be directly applied to any DBGD-type OL2R algorithm. We theoretically prove that the projected direction is still an unbiased estimate of the true gradient, i.e., model convergence is guaranteed, and also prove the reduced variance directly leads to considerable regret reduction in online model update. We compare the proposed method with several best-performing DBGD-type OL2R algorithms on a collection of large-scale learning to rank datasets and confirmed the effectiveness of our proposed solution.

2 RELATED WORK

One key family of OL2R methods root in Dueling Bandit Gradient Descent (DBGD) [26], which uses online gradient descent to solve a bandit convex optimization problem [5]. In each iteration, DBGD uniformly samples a random direction from the entire parameter space to create an exploratory ranker, and uses an interleaved test [17] to compare the current ranker with the exploratory one. If the exploratory ranker is preferred, the proposed direction is used as the gradient to update the model. This procedure yields an unbiased estimate of true gradient [24]. However, the variance of DBGD’s

gradient estimation is high due to the nature of uniform exploration of the entire parameter space, which limits its learning efficiency.

Recently, attempts have been made to improve the learning efficiency of DBGD-type algorithms. Schuth et al. [19] proposed a Multileave Gradient Descent (MGD) algorithm to explore multiple stochastic directions in each iteration with multi-interleaving comparison [20]. Zhao and King [27] developed a Dual-Point Dueling Bandit Gradient Descent algorithm to sample two stochastic vectors with opposite directions as the candidate gradients. The basic idea of this line of solutions is to test more exploratory directions at once so as to obtain the true gradient estimate sooner. However, their gradient exploration is still within the entire feature space. As users often only examine a small number of documents under each query, the sensitivity of interleaved test drops due to more exploratory rankers need to be tested. In a different direction of solutions, researchers proposed to constrain the sampling space for gradient exploration. Hofmann et al. chose to filter the stochastic directions by historical comparisons before an interleaved test [7]. Oosterhuis et al. [14] proposed to explore gradients in a subspace constructed by a set of pre-selected reference documents from an offline training corpus. Wang et al. [22] proposed to use historical interactions to avoid repeatedly exploring less promising directions, which also reduces gradient exploration to a subspace. However, the variance of gradient exploration is reduced at a cost of introducing bias into gradient approximation, so that such algorithms have a risk of converging to sub-optimal results.

Our solution falls into this second category of variance reduction approaches for DBGD-type algorithms. Distinct from previous attempts to restrict gradient exploration before an interleaved test, we instead modify the selected direction after the test. As users’ result examination is affected by the ranked results, which are in turn determined by the proposed exploratory directions, restricting the exploration space before the interleaved test potentially introduces bias in the subsequent interleaved test and model update. Our solution is based on the insight that only the projected true gradient in the document space can be revealed by an interleaved test. Hence, we decide to project the selected direction after each interleaved test, and thus guarantee an unbiased estimate of true gradient. Since the document space is expected to be smaller than the entire parameter space (as it is constructed only by the examined documents), the projected gradient enjoys low variance and leads to faster model convergence in online update.

There are also other parallel lines of OL2R algorithms that do not explore the gradient space the way DBGD-type algorithms do, but directly optimize the ranking model from click feedback. Kveton et al. [11] proposed Cascading Bandits to learn from users’ click behaviour, where skipped documents are assumed to be less attractive than later clicked ones. This model is then extended to the dependent click model [10] to support multiple clicks in one query, and further studied for general stochastic click models [28]. However, these algorithms estimate a separate model for each query and do not share estimation across queries, which lead to slow convergence. Oosterhuis et al. [15] proposed a Pairwise Differentiable Gradient Descent (PDGD) algorithm that constructs gradients from pairwise result comparisons to update the model, and can be used to optimize neural network models. We should note that our solution is not compatible nor directly comparable with these non-DBGD

algorithms, as there is no gradient exploration in these algorithms and our proposed gradient project does not apply.

3 METHOD

In this section we describe our proposed document space gradient projection method for online learning to rank. We first describe the problem setup in Section 3.1. Then we present Document Space Projected Dueling Bandit Gradient Descent (DBGD-DSP) algorithm as an example of our proposed general solution in Section 3.2. Our gradient projection method is independent from how the exploratory gradient is proposed, and thus can be directly applied to any existing DBGD-type OL2R algorithm¹ to reduce its variance of gradient estimation. We rigorously prove the unbiasedness of our gradient estimation in Section 3.3 and analyze the regret of DBGD-DSP in Section 3.4. The same procedure and conclusions can be applied to any DBGD-type algorithm of interest.

3.1 Problem Setup

The estimation of OL2R models can be formalized as a dueling bandit problem [26]. In iteration t , an OL2R algorithm receives a query and associated candidate documents, which are represented as a set of d -dimensional query-document pair feature vectors $X_t = \{x_1, x_2, \dots, x_s\}$. The algorithm takes two actions: first, it proposes two rankers, whose parameters are denoted as w, w' ; second, it ranks the given documents with these two rankers accordingly. An oracle (i.e., user) compares (duels) the two rankers' results and provides feedback. In practice, an interleaving method [17] is applied to merge the ranking lists of the two rankers and display the resulting ranked list to the user. User preference is inferred from the click feedback. Thus, the ranker that contributes more clicked documents is preferred. We denote $w > w'$ for the event that w is preferred over w' . The comparison between two individual rankers is determined independently of other comparisons performed before with a probability $P(w > w' | X_t)$, such that $P(w > w' | X_t) = P_t(w > w') = f_t(w, w')$. $f_t(w, w')$ can be viewed as the distinguishability of the two rankers w and w' by an interleaving comparison under query X_t .

We quantify the performance of an online learning algorithm using cumulative regret defined as follows:

$$R(T) = \sum_{t=1}^T f_t(w^*, w_t) + f_t(w^*, w'_t), \quad (1)$$

where w_t and w'_t are rankers compared at time t , and w^* is the best ranker in ground-truth. As a result, the distinguishability measure $f_t(w^*, w)$ indicates the loss of proposing a sub-optimal ranker w . We denote $f_t(w_t, w)$ as $f_t(w)$ for simplicity. The goal of an OL2R algorithm is to optimize its parameter towards w^* according to loss $f_t(w)$. A desired OL2R algorithm should have a sublinear regret in a finite time horizon T , so that the one-step regret is quickly decreasing to zero over time.

In this work, we make the following assumptions similar to [26]. We assume an unknown utility function $v_t(w)$ that quantifies the quality of a ranker w over query X_t . The utility function v_t is

¹In the following discussions, we will use "DBGD-type OL2R algorithm" and "OL2R algorithm" interchangeably, as the focus of this paper is improving the efficiency of DBGD-type OL2R algorithms.

Algorithm 1 Document Space Projected Dueling Bandit Gradient Descent (DBGD-DSP)

```

1: Inputs:  $\delta, \alpha$ 
2: Initiate  $w_1 = \text{sample\_unit\_vector}()$ 
3: for  $t = 1$  to  $T$  do
4:   Receive query  $X_t = \{x_1, x_2, \dots, x_s\}$ 
5:    $u_t = \text{sample\_unit\_vector}()$ 
6:    $w'_t = w_t + \delta u_t$ 
7:   Generate ranked lists  $l(X_t, w_t), l(X_t, w'_t)$ 
8:   Set  $L_t = \text{Interleave}(\{l(X_t, w_t), l(X_t, w'_t)\})$ , and present  $L_t$  to user
9:   Receive click positions  $C_t$  on  $L_t$ , and infer click credits  $\{c_t, c'_t\}$ 
10:  if  $c_t \geq c'_t$  then
11:     $w_{t+1} = w_t$ 
12:  else
13:    Based on  $C_t$ , infer user examined top  $m_t$  documents in  $L_t$ .
14:    Solve the orthogonal projection matrix  $A_t$  for document space  $S_t = \text{span}(\{x_{L_t, 1}, x_{L_t, 2}, \dots, x_{L_t, m_t}\})$ .
15:    Project  $u_t$  onto  $S_t$  by  $g_t = A_t u_t$ 
16:     $w_{t+1} = w_t + \alpha g_t$ 
17:  end if
18: end for

```

assumed to be differentiable, strongly concave and L_v -Lipschitz, which means $|v_t(x) - v_t(y)| \leq L_v |x - y|$.

A link function σ describes the probabilistic comparison of utilities of two rankers as,

$$P_t(w > w') = f_t(w, w') = \sigma(v_t(w) - v_t(w')).$$

The link function should be rotation-symmetric, which means $\sigma(x) = 1 - \sigma(-x)$. We assume the link function is L_σ -Lipschitz and second order L_2 -Lipschitz. The link function behaves like a cumulative probability distribution function. For example, a common choice of link function is the standard logistic function $\sigma(x) = \frac{1}{1 + \exp(-x)}$, which satisfies all the assumptions.

3.2 Document Space Projected Dueling Bandit Gradient Descent

We describe our proposed Document Space Projected Dueling Bandit Gradient Descent (DBGD-DSP) in Algorithm 1. We should note it fits all DBGD-type OL2R algorithm settings. At the beginning of iteration t , user initiates a query X_t . We denote w_t as the parameter of the current ranker. DBGD-DSP first uniformly samples a vector u_t from d dimensional unit sphere \mathbb{S}^{d-1} (i.e., $|u_t|_2 = 1$) as an exploratory direction, and proposes a candidate ranker $w'_t = w_t + \delta u_t$, where δ is the step size of exploration. The algorithm then uses the two rankers (w_t and w'_t) to generate ranking lists $l(X_t, w_t)$ and $l(X_t, w'_t)$ accordingly, and combines them with an interleaving method, such as Team Draft Interleaving [17] or Probabilistic Interleaving [8]. The user examines the result list and provides implicit click feedback to indicate their relevance evaluation of the results. The interleaving method uses this implicit feedback to infer which ranker is preferred by the user. If the exploratory ranker is preferred

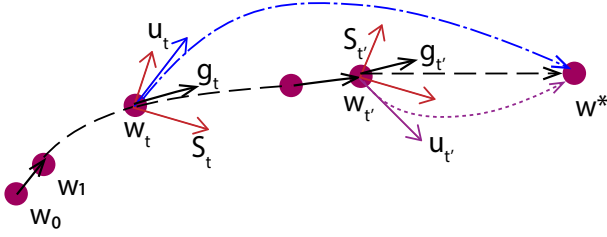


Figure 1: Illustration of model update for DBGD-DSP in a three dimensional space. Dashed lines represent the trajectory of DBGD following different update directions. u_t is the selected direction by DBGD, which is in the 3-d space. Red bases present the document space S_t on a 2-d plane. u_t is projected onto S_t to become g_t for model update.

(i.e., wins the duel), previous DBGD-style algorithms update the current ranker by $w_{t+1} = w_t + \alpha u_t$, where α is the learning rate; otherwise the current ranker stays intact. This gradient exploration strategy yields an unbiased estimate of the true gradient [5], in terms of expectation.

However, since the exploratory gradient u_t is required to be uniformly sampled from the entire d dimensional unit sphere \mathbb{S}^{d-1} , the model update suffers from high variance in its gradient estimation, especially when d is large, as in practice. Various improvements to this issue have been proposed in the past, but they still introduce other difficulties, such as variance and bias trade-off [7, 14, 22], and test sensitivity and efficiency [20, 27].

Unlike previous works that reduce the sampling space of gradient exploration before the interleaved test [7, 14, 22], we change the winning direction after the test. The key insight is that only the projected true gradient in the spanned space of *examined* documents under query X_t (denoted as document space S_t) can be revealed by an interleaved test. For example, as shown in Figure 1, a DBGD-style algorithm is comparing the current ranker w_t and $w'_t = w_t + \delta u_t$ with a uniformly sampled exploration direction u_t . The user examines top m documents, e.g., $\{x_1, \dots, x_m\}$, of the interleaved ranking list (of course m is unknown to the algorithm) and w'_t wins the duel. The estimated gradient u_t can therefore be separated into two components, one component g_t that belongs to the document space $S_t = \text{span}\{x_1, \dots, x_m\}$ and the other component $u_t - g_t$ that is orthogonal to document space S_t . The orthogonal component $u_t - g_t$ does not affect the ranking among the examined documents, i.e. $(w_t + \delta u_t)^T x_i = (w_t + \delta g_t)^T x_i$, and thus does not contribute to the loss function and true gradient estimation. Intuitively, $u_t - g_t$ is not supported by the observed interleaved test, as anything sampled from the complement of S_t cannot be verified by the examined documents. As a result, it is safe to exclude the direction $u_t - g_t$ from model update, which we later prove maintains the unbiasedness of the original DBGD-type gradient estimation, and reduces the variance. As illustrated in Figure 1, although u_t will eventually lead to the same model estimation, as it is unbiased, this guarantee is only obtained in expectation. The variance could potentially be large: for example, the blue and purple updating traces slow down model convergence, when the number of observations is finite.

As shown in line 14 to 16 of Algorithm 1, we solve for the orthogonal projection matrix A_t of document space S_t , and project the selected direction u_t onto the document space S_t after each interleaved test. We leave the detailed design of constructing document space and solving projection matrix A_t in Section 3.5. Before that, we first rigorously prove the projection maintains an unbiased estimate of true gradient in Section 3.3. Since the document space is constructed only by the examined documents, the rank of document space is expected to be smaller than the entire parameter space. This directly leads to lower variance and faster model convergence. We show that our document space projection reduces the variance of gradient estimation from d to $\text{Rank}(A_t)$ in Section 3.4, and then analyze its benefit for regret reduction from a low-variance gradient estimation.

3.3 Unbiasedness of Gradient Estimation

We now prove that our document space projected gradient is an unbiased estimate of true gradient in the sense of expectation [26]. We define $Z_t(w)$ as the event of w winning the duel with w_t ,

$$Z_t(w) = \begin{cases} 1 & \text{w.p. } 1 - P_t(w_t > w) \\ 0 & \text{w.p. } P_t(w_t > w) \end{cases}$$

Then the gradient used for model update in DBGD-DSP (as described in Algorithm 1) can be described as,

$$h_t = -Z_t(w_t + \delta u_t)g_t. \quad (2)$$

Note that by adding a negative sign we view our model update as online gradient descent $w_{t+1} = w_t - \alpha g_t$.

We now show in the following theorem that this is an unbiased gradient estimation of true gradient. By defining a smoothed version of f_t as $\hat{f}_t(w) = \mathbb{E}_{u \in \mathbb{B}}[f_t(w + \delta u)]$, we have:

THEOREM 3.1. *The projected gradient g_t in DBGD-DSP is an unbiased estimate of true gradient, i.e.,*

$$\mathbb{E}[h_t] = \frac{\delta}{d} \nabla \hat{f}_t(w) \quad (3)$$

over random unit vector u_t .

PROOF. Based on the Lemma 1 of [26], we have

$$\mathbb{E}[h_t] = \mathbb{E}[-Z_t(w_t + \delta u_t)A_t u_t] = \mathbb{E}_{u_t \in \mathbb{S}^{d-1}}[f_t(w + \delta A_t u_t)u_t]$$

Define $F_t(w) = f_t(A_t w)$, we have

$$\begin{aligned} \mathbb{E}[h_t] &= \mathbb{E}_{u_t \in \mathbb{S}^{d-1}}[f_t(w_t + \delta A_t u_t)u_t] \\ &= \mathbb{E}_{u_t \in \mathbb{S}^{d-1}}[F_t(A_t^{-1} w_t + \delta u_t)u_t] \\ &= \frac{\delta}{d} \nabla \mathbb{E}_{u_t \in \mathbb{B}^d}[F_t(A_t^{-1} w_t + \delta u_t)u_t] \\ &= \frac{\delta}{d} \nabla \hat{F}_t(A_t^{-1} w_t) \\ &= \frac{\delta}{d} A_t \nabla \hat{f}_t(w_t) \\ &= \frac{\delta}{d} \nabla \hat{f}_t(w_t) \end{aligned}$$

where the third equality is based on Stokes' Theorem. The last equality holds because gradient $\nabla \hat{f}_t(w_t)$ belongs to document space S_t , and thus projecting it by A_t maps back to itself. \square

The guarantee of unbiased gradient estimation is a major advantage of our proposed document space gradient projection method, compared with previous attempts to reduce the gradient exploration space, such as Oosterhuis et. al [14] and Wang et al. [22]. Our method enjoys reduced variance of gradient estimate (which will be proved next), without the risk of converging towards a sub-optimal solution. We should note that the above is independent from the mechanism of how the proposal directions are generated, as shown in the first four steps of proof above. As a result, if the input direction to our projection procedure is unbiased, the resulting update direction is also unbiased. This enables our solution's generalization to other types of DBGD algorithms.

3.4 Regret Analysis of DBGD-DSP

We now analyze the regret of our proposed DBGD-DSP algorithm, starting with its variance of gradient update.

LEMMA 3.2. *The variance of gradient update in DBGD-DSP is bounded by*

$$\mathbb{E}[|h_t|^2] = \mathbb{E}_{u_t \in \mathbb{S}^{d-1}} [| -Z_t(w_t + \delta u_t) \mathbf{A}_t u_t |^2] \leq \frac{\text{Rank}(\mathbf{A}_t)}{d}.$$

PROOF.

$$\begin{aligned} \mathbb{E}[|h_t|^2] &= \mathbb{E}_{u_t} [| -Z_t(w_t + \delta u_t) \mathbf{A}_t u_t |^2] \\ &\leq \mathbb{E}_{u_t} [| \mathbf{A}_t u_t |^2] \\ &= \mathbb{E}_{u_t} [(\mathbf{A}_t u_t)^\top (\mathbf{A}_t u_t)] \\ &= \text{tr}(\mathbb{E}_{u_t} [\mathbf{A}_t u_t u_t^\top \mathbf{A}_t^\top]) \quad // \text{apply the trace trick} \\ &= \text{tr}(\mathbf{A}_t \mathbb{E}_{u_t} [u_t u_t^\top] \mathbf{A}_t^\top) \\ &= \text{tr}\left(\mathbf{A}_t \frac{1}{d} \mathbf{I} \mathbf{A}_t^\top\right) \\ &= \frac{1}{d} \text{tr}(\mathbf{A}_t \mathbf{A}_t^\top) \\ &= \frac{1}{d} \text{tr}(\mathbf{A}_t) \quad // \text{a projection matrix is idempotent} \\ &= \frac{\text{Rank}(\mathbf{A}_t)}{d} \end{aligned}$$

where $\text{tr}(\cdot)$ denotes the matrix trace operation. The sixth equality holds because u_t is uniformly sampled from a unit sphere, and its covariance matrix $\mathbb{E}_{u_t} [u_t u_t^\top]$ is $\frac{1}{d} \mathbf{I}$. Since \mathbf{A}_t is an orthogonal projection matrix, the eighth equality holds for $\mathbf{A}_t \mathbf{A}_t^\top = \mathbf{A}_t$. \square

Remark. The variance of gradient update in DBGD [26] is bounded by $\mathbb{E}_{u_t} [| -Z_t(w_t + \delta u_t) u_t |^2] \leq 1$.

Comparing the variance of gradient update in DBGD-DSP with DBGD, our method reduces the variance from 1 to $\frac{\text{Rank}(\mathbf{A}_t)}{d}$. Since the dimension of projection matrix \mathbf{A}_t is d -by- d , we have $\text{Rank}(\mathbf{A}_t) \leq d$, which guarantees the reduction of variance in DBGD-DSP comparing to that in DBGD. The rank of \mathbf{A}_t is also bounded by the number of *examined* documents m_t , since document space S_t is constructed by these m_t examined documents. In practice, users would only examine a handful of documents [4, 9], while the ranking feature dimension is expected to be much larger. We argue that $m_t \ll d$, such that our document space projection achieves considerable variance reduction.

The significance of this variance reduction can be intuitively understood from Figure 1: though different traces of model update would eventually lead to the same converged model, if one has a sufficiently large amount of interactions with users, the one with lower variance would always require less observations. A faster converging algorithm leads to user satisfaction earlier. Next, we verify this benefit by proving the reduction of regret introduced by the reduced variance in gradient estimation.

THEOREM 3.3. *By setting*

$$m = \max_t m_t, \delta = \frac{\sqrt{2Rm}}{\sqrt{13LT}^{1/4}}, \alpha = \frac{Rm}{\sqrt{T}\delta},$$

the expected regret of DBGD-DSP as defined in Eq (1) is upper bounded by,

$$\mathbb{E}[\text{Reg}] \leq 2\lambda_T T^{3/4} \sqrt{26RmL}, \quad (4)$$

where

$$\lambda_T = \frac{L_\sigma \sqrt{13LT}^{1/4}}{L_\sigma \sqrt{13LT}^{1/4} - L_\sigma L_2 \sqrt{2Rm}}$$

The proof is obtained by extending Theorem 2 in [26]. We omit the details due to space limit, and emphasize that the key difference is introduced by replacing variance of gradient estimation from $\mathbb{E}_{u_t} [| -Z_t(w_t + \delta u_t) u_t |^2]$ to $\mathbb{E}_{u_t} [| -Z_t(w_t + \delta u_t) \mathbf{A}_t u_t |^2]$. Since the variance of gradient estimation is reduced from 1 to $\frac{\text{Rank}(\mathbf{A}_t)}{d}$, the regret of DBGD can be reduced from $O(\sqrt{dT}^{3/4})$ to $O(\sqrt{mT}^{3/4})$, where m is the maximum number of documents included in a document space under a single query. Again, as the number of included ranking features is oftentimes much larger than the number of documents a user would examine under a single query, the reduction of regret is considerable. Moreover, as the reduction of variance from our project-based method is independent from the way about how the proposal directions are generated, our method can be generally applied to most existing DBGD-type OL2R algorithms to improve their learning convergence.

3.5 Practical Treatments of Document Space Projection

Now we discuss several practical treatments of our proposed Document Space Projection method, including the construction of document space and orthogonal projection matrix.

In our theoretical analysis, we have assumed the knowledge of users' examined documents and corresponding projection matrix. However, in practice, a user's result examination is unobserved. A rich body of research has been developed to perform statistical inference of it, collectively known as click modeling [3, 4]. Any of these existing click models can be plugged into our solution framework, i.e., line 13 of Algorithm 1. In this work, we simply follow [9] to infer user examination by the last clicked position: given the click position list C_t , we use the last clicked position $c_{l,t}$ to approximate the last examined position M_t by setting $M_t = c_{l,t} + k$, where k is a hyper-parameter. Based on sequential examination hypothesis of click modeling, every document before the last clicked position is examined, and we use k to approximate the number of positions following the last clicked position that were still examined. We leave more comprehensive study of click modeling in our solution as future work.

The above treatment provides a reasonable inference of examined documents. However, it requires a careful choice of k for each query (preferably). If k is set too large, the variance of gradient estimate will increase (as proved in Lemma 3.2). If k is too small, the document space may not include all examined documents, and it is at risk of introducing bias in gradient projection. To avoid bias in constructing the document space, we also consider adding historically examined documents to the current query’s document space. Specifically, we add r recently examined documents to the current document space S_t to compensate the potentially overlooked examined documents in the current query.

In line 14 of Algorithm 1, we solve the orthogonal projection matrix A_t of document space S_t . A_t could be computed by several methods. Denote D_t as a d -by- m_t matrix where each column is the feature vector for an examined document. One can use QR decomposition or Singular Value Decomposition (SVD) to solve for its orthonormal basis V_t , and the projection matrix can then be constructed by $A_t = V_t V_t^T$. In our experiments, we chose SVD for constructing the basis of document space, because of its widely available and efficient large-scale implementations. But the choice for the construction of this project matrix does not affect the convergence nor unbiasedness of our proposed solution.

4 EXPERIMENTS

To demonstrate our proposed Document Space Projection method’s empirical efficacy, we compare the performance of several best-performing DBGD-type OL2R algorithms on five public learning to rank datasets, with and without our document space projection method applied.

4.1 Experiment Setup

• **Datasets.** We tested our algorithms and the baselines on five benchmark datasets: including MQ2007, MQ2008, NP2003 [13], MSLR-WEB10K [16], and the Yahoo! Learning to Rank Challenge dataset [2]. In each of the five datasets, each query-document pair is encoded as a vector of ranking features. These features include PageRank, TF.IDF, Okapi-BM25, URL length, language model score, and many more varied by dataset.

The MQ2007 and MQ2008 datasets are collected from the 2007 and 2008 Million Query track at TREC [21]. MQ2007 contains about 1700 queries, and MQ2008 contains about 800 queries, which represent a mix of informational and navigational search intents. They both have 46-dimensional feature vectors to represent query-document pairs, and the document relevance are labeled in three grades: 0 (not relevant), 1 (relevant), and 2 (most relevant).

The NP2003 dataset also comes from the TREC Web track, consisting of queries crawled from the .gov domain. It is comprised of about 150 navigational-focused queries, with over 1000 document relevance assessments per query. It uses 64 ranking features, and the document relevance labels are binary (0 and 1 only).

The MSLR-WEB10K dataset was released by Microsoft in 2010, and consists of 10,000 queries with relevance assessments coming from a labeling set from the Microsoft Bing search engine. It has 136 ranking features, and the relevance judgments range from 0 (not relevant) to 4 (most relevant).

Table 1: Configurations of simulation click models.

R	Click Probability					Stop Probability				
	0	1	2	3	4	0	1	2	3	4
Per	0.0	0.2	0.4	0.8	1.0	0.0	0.0	0.0	0.0	0.0
Nav	0.05	0.3	0.5	0.7	0.95	0.2	0.3	0.5	0.7	0.9
Inf	0.4	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5

The Yahoo! Learning to Rank Challenge dataset was also released in 2010, as an effort on part of Yahoo! to promote the dataset as well as research into better learning to rank algorithms. The dataset contains about 36,000 queries, 883,000 assessed documents, and 700 ranking features. Again, the relevance judgments range from 0 (not relevant) to 4 (most relevant).

This diversity in the structure of the datasets that we chose to test on helps us to evaluate our algorithms more holistically. While small, the MQ2007 and MQ2008 sets have been around for a long time and have a good mix of query types. NP2003 gives us insight into how the algorithms perform on navigational search intents specifically, which are markedly different in nature from informational search intents. MSLR-WEB10K and the Yahoo! dataset are large-scale datasets used by actual commercial search engines, which give us a better understanding of how the algorithms perform in practice. Since each dataset was split into training, testing, and validation subsets, we used the training sets for online experiments to measure cumulative performance, and used the testing sets for evaluating offline performance.

• **Simulated User Interactions.** Based on an online learning to rank framework proposed in [15], we use the standard setup to simulate user interactions. Within this framework, we used the Cascade Click Model to simulate user click behavior. This model assumes that a user interacts with a set of search results by linearly scanning the list from top and making a decision for each document as to whether or not to click. In the model, the probability of a click for a given document is conditioned on the relevance label of that document, as a user is expected to be more likely to click on relevant documents. After evaluating each document, the user must decide whether or not to continue perusing the list. This decision’s probability distribution is again conditioned on the relevance of the examined document, as a user is more likely to stop looking through the results if he/she has already satisfied their information need. These aforementioned probabilities can be altered to simulate different types of users and interactions.

As illustrated in Table 1, we use three different click model probability configurations to represent three different types of users. First, we have the *perfect* user, who clicks on all relevant documents and does not stop browsing until they have visited all of the documents. This type of users contribute the least noise, as they make no mistakes and the feedback is entirely accurate. Second, we have the *navigational* user, who is very likely to click on the first highly relevant document that he/she sees and stops there. Third, we have the *informational* user, who, in his/her search for information, sometimes clicks on irrelevant documents, and as such contributes a significant amount of noise in click feedback.

• **Evaluation Metrics.** As set forth in [18], cumulative (online) Normalized Discounted Cumulative Gain (NDCG) and offline NDCG

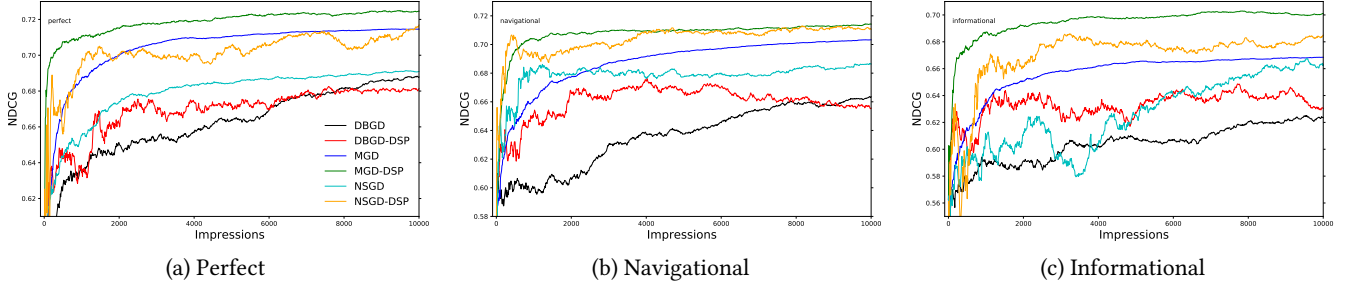


Figure 2: Offline NDCG@10 on Yahoo! dataset.

are commonly used metrics for evaluating OL2R algorithms. Cumulative NDCG is calculated by summing NDCG scores from successive iterations with a discount factor γ set to 0.995. We assess our model’s estimation convergence via cosine similarity between the current weight vector and a reference weight vector (considered to be the optimal vector) as estimated by an offline learning-to-rank algorithm trained with the complete true relevance judgment labels. Due to its superior empirical performance, we used LambdaRank [1] with no hidden layer in our experiments to estimate this reference weight vector. In each experiment, the number of iterations T was set to 10,000, and the current query X_t was randomly sampled from the dataset in each iteration. We execute all the experiments 15 times with different random seeds, and report and compare the average performance in all experiments.

• **Evaluation Questions.** To better understand the advantages of our proposed algorithms, we aim to answer the following evaluation questions through the course of our experiments.

- Q1: Can our proposed Document Space Projection method consistently improve the performance of best-performing DBGD-type OL2R algorithms?
- Q2: Do gradients rectified by our document space projection explore the gradient space more efficiently?
- Q3: How do different hyper-parameter settings alter the performance of our document space projection?

• **Baseline Algorithms.** We choose the following three best-performing DBGD-type OL2R algorithms as our baselines for comparison:

- **DBGD** [26]: A single direction uniformly sampled from the whole parameter space is explored.
- **MGD** [19]: Multiple directions are explored in one iteration to reduce the gradient estimation variance. Multileaving is used to compare multiple rankers. The model updates towards the mean of all rankers that beat the current model.
- **NSGD** [22]: Multiple directions are sampled from the null space of previously poorly performing gradients. Ties are broken by evaluating the tied candidate rankers on a recent set of difficult queries.

We apply our proposed Document Space Projection to the baseline algorithms, and compare them with DBGD-DSP, MGD-DSP and NSGD-DSP, respectively.

4.2 Performance of Document Space Projection

We begin our experimental analysis by answering our first evaluation question. We compared all algorithms over 3 click models and

5 datasets. We set the hyper-parameters of DBGD, MGD and NSGD according to their original papers. Following [19, 26], we set the exploration step size δ to 1 and learning rate α to 0.1. Both MGD and NSGD explore 9 proposal directions in one iteration. For our document space projection method, we consider $k = 3$ documents following the last clicked position as examined documents, and add $r = 10$ recently examined documents into document space S_t . We use SVD to solve for orthonormal basis V_t of the document space S_t , and compute the projection matrix by $A_t = V_t V_t^T$.

We reported the offline NDCG@10 and online cumulative NDCG@10 after 10,000 iterations in Table 2 and Table 3. Due to space limit, we only reported the offline performance during the 10,000 iterations over 3 click models on Yahoo dataset, a large-scale real-world L2R dataset with 700 ranking features, in Figure 2. MGD improves the online performance over DBGD by exploring multiple rankers simultaneously, and NSGD further improves over MGD by exploring gradients in a constrained subspace, as shown in Table 2. We observe that our proposed document space projection method consistently improves the online performance of all baseline algorithms. Recall that in Section 3.4 our theoretical analysis suggested that document space projection reduces both the gradient estimation variance and the regret (online performance) with respect to the ratio between the rank of document space and feature dimension. Correspondingly, we observe that indeed we improved the OL2R models’ ranking performance significantly over MSLR-WEB10K and Yahoo datasets, which are collected from real-world commercial search engines and have much higher feature dimensions (130 and 700 respectively). This result demonstrates the potential of document space projection to improve large-scale real-world DBGD-type OL2R applications with high-dimensional ranking features, as our algorithm attains satisfactory performance earlier than other baseline OL2R algorithms measured by online NDCG@10. We also notice that the standard deviation of those models’ ranking performance is reduced when applying document space projection, which confirms our analysis of variance reduction in Lemma 3.2.

From Figure 2 and Table 3 we notice that document space projection mostly improves offline performance over baseline algorithms. Figure 2 shows that document space projection significantly accelerates the convergence rate over the baseline algorithms, because of the reduced variance in gradient estimation. We also observe that applying document space projection under the perfect click model may lead to degraded performance, for example DBGD on MQ2007

Table 2: Online NDCG@10, standard deviation and relative improvement of document space projection of each algorithm after 10,000 queries.

Click Model	Algorithm	MQ2007	MQ2008	MSLR-WEB10K	NP2003	Yahoo
Perfect	DBGD	679.3 (21.6)	847.1 (38.4)	532.2 (15.3)	1130.2 (43.3)	1165.5 (22.6)
	DBGD-DSP	689.1 (19.5)(+1.44%)	858.0 (39.2)(+1.29%)	553.6 (13.1)(+4.02%)	1198.8 (40.0)(+6.07%)	1198.8 (33.5)(+2.86%)
	MGD	689.1 (14.6)	859.4 (38.1)	558.3 (7.0)	1192.9 (44.6)	1201.9 (16.3)
	MGD-DSP	757.3 (16.2)(+9.90%)	919.5 (42.2)(+6.99%)	626.4 (9.6)(+12.20%)	1335.3 (39.1)(+11.94%)	1309.4 (10.6)(+8.94%)
	NSGD	684.4 (20.5)	867.5 (40.3)	589.5 (14.2)	1274.9 (47.4)	1162.3 (12.9)
	NSGD-DSP	732.5 (20.0)(+7.03%)	904.3 (38.0)(+4.24%)	635.6 (12.8)(+7.82%)	1368.5 (41.1)(+7.34%)	1270.1 (2.5)(+9.27%)
Navigational	DBGD	646.1 (23.4)	817.9 (45.5)	517.5 (20.9)	1062.3 (55.4)	1133.3 (40.8)
	DBGD-DSP	664.9 (26.9)(+2.91%)	830.3 (44.1)(+1.52%)	543.1 (14.8)(+4.95%)	1140.1 (52.5)(+7.32%)	1199.4 (34.6)(+5.83%)
	MGD	632.7 (15.5)	827.5 (35.5)	538.2 (7.2)	1115.4 (44.6)	1171.3 (20.4)
	MGD-DSP	694.5 (15.7)(+9.77%)	882.3 (40.0)(+6.62%)	586.9 (9.5)(+9.05%)	1300.9 (39.6)(+16.63%)	1290.2 (15.3)(+10.15%)
	NSGD	660.1 (24.5)	849.1 (36.6)	562.1 (18.8)	1211.1 (66.5)	1186.2 (16.8)
	NSGD-DSP	724.6 (24.5)(+9.77%)	895.8 (34.2)(+5.50%)	608.3 (12.1)(+8.22%)	1296.2 (24.3)(+7.03%)	1283.4 (7.2)(+8.19%)
Informational	DBGD	583.4 (46.0)	763.9 (55.1)	472.4 (34.6)	849.8 (144.5)	1107.3 (46.6)
	DBGD-DSP	620.1 (40.8)(+6.29%)	782.4 (51.8)(+2.42%)	522.1 (18.6)(+10.52%)	992.5 (81.1)(+16.79%)	1158.5 (22.0)(+4.62%)
	MGD	621.2 (18.2)	817.5 (45.3)	538.3 (10.8)	1107.9 (46.2)	1146.6 (37.5)
	MGD-DSP	671.4 (18.9)(+8.08%)	865.9 (37.7)(+5.92%)	580.5 (10.4)(+7.84%)	1274.5 (42.9)(+15.04%)	1268.1 (16.4)(+10.60%)
	NSGD	629.7 (25.3)	814.9 (37.1)	532.9 (15.2)	1123.5 (59.8)	1110.5 (10.9)
	NSGD-DSP	703.6 (29.2)(+11.74%)	871.3 (48.3)(+6.92%)	597.9 (14.1)(+12.20%)	1222.8 (43.8)(+9.03%)	1204.7 (9.6)(+8.48%)

Table 3: Offline NDCG@10, standard deviation and relative improvement of document space projection of each algorithm after 10,000 queries.

Click Model	Algorithm	MQ2007	MQ2008	MSLR-WEB10K	NP2003	Yahoo
Perfect	DBGD	0.484 (0.023)	0.683 (0.023)	0.331 (0.009)	0.737 (0.056)	0.688 (0.011)
	DBGD-DSP	0.480 (0.020)(-0.83%)	0.685 (0.024)(+0.29%)	0.333 (0.011)(+0.6%)	0.738 (0.059)(+0.14%)	0.681 (0.013)(-1.02%)
	MGD	0.495 (0.022)	0.691 (0.020)	0.334 (0.003)	0.746 (0.048)	0.715 (0.002)
	MGD-DSP	0.501 (0.021)(+1.21%)	0.695 (0.022)(+0.58%)	0.409 (0.006)(+22.46%)	0.748 (0.055)(+0.27%)	0.725 (0.003)(+1.40%)
	NSGD	0.488 (0.019)	0.689 (0.024)	0.397 (0.012)	0.743 (0.050)	0.691 (0.005)
	NSGD-DSP	0.491 (0.022)(+0.61%)	0.691 (0.025)(+0.29%)	0.398 (0.008)(+0.25%)	0.750 (0.042)(+0.94%)	0.717 (0.004)(+3.76%)
Navigational	DBGD	0.463 (0.028)	0.667 (0.021)	0.320 (0.012)	0.728 (0.054)	0.663 (0.020)
	DBGD-DSP	0.465 (0.024)(+0.43%)	0.668 (0.023)(+0.15%)	0.327 (0.011)(+2.19%)	0.734 (0.052)(+0.82%)	0.656 (0.013)(-1.06%)
	MGD	0.426 (0.019)	0.664 (0.016)	0.321 (0.003)	0.740 (0.048)	0.703 (0.010)
	MGD-DSP	0.467 (0.021)(+9.62%)	0.684 (0.017)(+3.01%)	0.331 (0.005)(+3.12%)	0.744 (0.053)(+0.54%)	0.714 (0.006)(+1.56%)
	NSGD	0.473 (0.022)	0.676 (0.024)	0.389 (0.013)	0.732 (0.053)	0.686 (0.008)
	NSGD-DSP	0.478 (0.020)(+1.06%)	0.683 (0.026)(+1.04%)	0.376 (0.014)(-3.34%)	0.788 (0.006)(+7.65%)	0.711 (0.001)(+3.64%)
Informational	DBGD	0.410 (0.034)	0.641 (0.031)	0.294 (0.022)	0.699 (0.063)	0.623 (0.037)
	DBGD-DSP	0.427 (0.027)(+4.15%)	0.632 (0.031)(-1.4%)	0.309 (0.011)(+32.65%)	0.692 (0.062)(-1.00%)	0.63 (0.030)(1.12%)
	MGD	0.406 (0.020)	0.651 (0.020)	0.317 (0.003)	0.726 (0.050)	0.668 (0.044)
	MGD-DSP	0.444 (0.025)(+0.44%)	0.669 (0.018)(+0.67%)	0.325 (0.004)(+0.33%)	0.738 (0.054)(+0.74%)	0.701 (0.005)(+4.94%)
	NSGD	0.469 (0.018)	0.674 (0.023)	0.360 (0.013)	0.733 (0.056)	0.663 (0.015)
	NSGD-DSP	0.466 (0.019)(-0.64%)	0.668 (0.026)(-0.89%)	0.340 (0.018)(-5.56%)	0.789 (0.013)(+7.64%)	0.685 (0.004)(+3.32%)

and Yahoo dataset. This is because document space projection guarantees an unbiased gradient estimation under the assumption of known result examinations, as discussed in Section 3.3. However, since in practice a user’s result examination is unobserved, we approximated the examined documents by including all documents before the last clicked position and k additional documents after the last clicked position. The perfect click model is an ideal case

that users’ stop probability is set to 0.0 (see Table 1) and every document is examined. Here, the document space needs to include all displayed documents to guarantee the unbiasedness, which requires a significantly larger k compared to the k used for navigational and informational click models. We argue that in practice since users only examine a handful of documents, we could well approximate

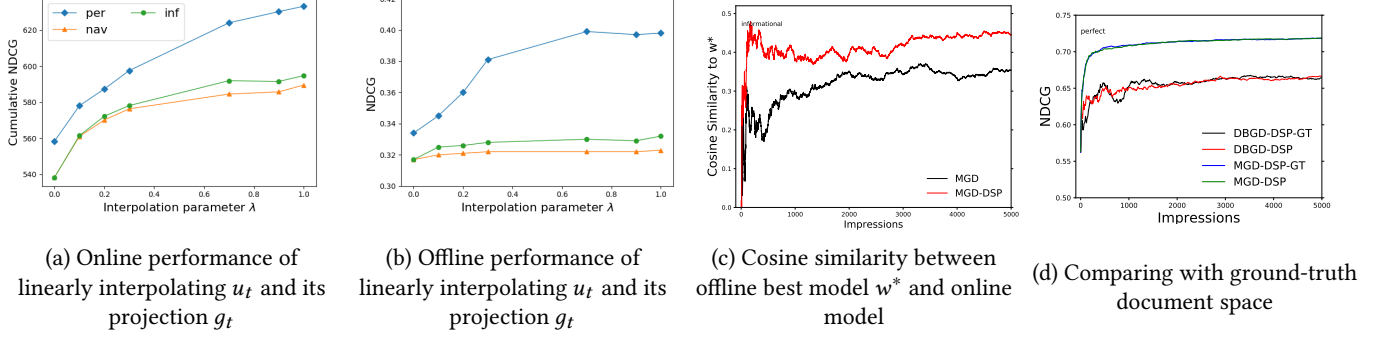


Figure 3: Analyzing Document Space Projection.

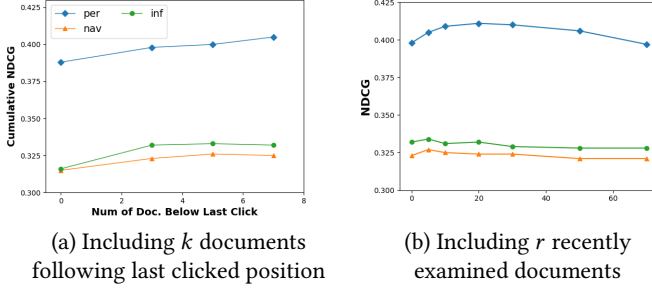


Figure 4: Hyper-parameter tuning for Document Space Projection.

the examined documents with a reasonable choice of k . More sophisticated click models can also be introduced. We will analyze the effect of k in Section 4.3. In addition, we also observe that under informational click model the performance of NSGD-DSP is slightly decreased compared with original NSGD over three datasets. Note that since NSGD does not guarantee its gradient exploration is unbiased, further projecting its gradient may also lead to a biased gradient update and thus a sub-optimal model.

4.3 Analysis of Document Space Projection

To answer the second evaluation question, we design two experiments to show the effectiveness of document space projected gradient. In the first experiment, we study the utility of document space projected gradient. We compare the ranking performance of linearly interpolating the unrectified direction u_t and its document space projected version g_t , i.e., $\lambda g_t + (1 - \lambda)u_t$, based on the MGD algorithm on MSLR-WEB10K dataset. Similar observations were obtained on other datasets, but due to space limit we have to omit those detailed results. We report the online and offline performance by varying λ from 0 (which is equivalent to the original MGD algorithm) and 1 (which is MGD-DSP) in Figure 3 (a) and (b). We can clearly observe a trend of increasing online performance over all three click models when we increase λ , i.e., trust more on the projected direction g_t for model update. This confirms the effectiveness of the projected direction g_t within document space comparing with the unrectified direction u_t from the entire parameter space. The offline performance is generally robust to the setting of λ for navigational and information click models. This is expected since

both MGD and MGD-DSP are unbiased and will eventually converge to similar offline performance after sufficiently large number of iterations (we had 10,000 iterations in our experiments).

In the second experiment, we trained an offline LambdaRank model [1] using the complete annotated relevance labels in the large-scale MSLR-WEB10K dataset. Then given this w^* , we compared cosine similarity between the online estimated model parameters with and without DSP in each iteration using MGD as the baseline. We show the result of first 5,000 iterations. In Figure 3 (c) we can observe that MGD-DSP converges faster and better to w^* than MGD. This suggests the rectified gradient is more effective than the original one. We also compared with an oracle algorithm that knows the ground-truth examined documents, denoted as DSP-GT, to validate the effectiveness of our approximated document space. We show the result on DBGD and MGD under the perfect click model in Figure 3(d). We notice that oracle algorithms performed similarly as our proposed algorithm with an approximated document space, which confirms the effectiveness of the approximation heuristics.

To answer the third evaluation question, we compare different hyper-parameters used for constructing the document space on MSLR-WEB10K dataset. We vary k from 0 to 7 and report the result in Figure 4 (a). We notice that for navigational and informational click models, a relatively small k achieved the best performance, i.e., $k = 3$. This corresponds to the observation that users do not continue to examine many documents after their last click under these two click models. However, under the perfect click model, the models' performance increases with a larger k . This aligns with the conclusions from our discussion in Section 4.2 that under the perfect click model, we need to set a much larger k to accurately construct the document space and guarantee an unbiased gradient estimate.

In Figure 4(b), we vary r . As we discussed in Section 3.5, we are motivated to add recently examined documents to compensate for potentially overlooked examined documents in the current query. The effect of different choices of r is more noticeable under the perfect click model. This echoes our analysis above that under the perfect click model some examined documents may be overlooked when k is not large enough. Thus correctly setting up r could reduce the bias in document space construction and compensate the final performance. From the result figure, we notice that setting $r = 20$ provides the best result. Under navigational and informational click models, the algorithm is generally robust to the choice of r . This

is because the approximations of examined documents are already accurate with a reasonable setting of k .

5 CONCLUSION

In this paper, we propose and develop the Document Space Projection (DSP) method for reducing variance in gradient estimation and improving online learning to rank performance. The key insight of DSP is to recognize that the interleaved test only reveals the projection of true gradient on the spanned space of *examined* documents. Including anything beyond this space for model update only introduces noise. Thus our method projects the selected model update direction back to the document space to reduce its variance. We proved that DSP maintains an unbiased gradient estimate, and it can substantially improve the regret bound for DBGD-style algorithms via the reduced variance. Through our extensive experiments, we found that DSP is able to provide statistically significant improvements to several best performing DBGD-type OL2R models, both in terms of variance reduction and overall performance, especially when the number of ranking features is large.

Currently, we are using a heuristic method to construct the document space. However, we did observe that the performance of DSP varies under different click models for simulated user click feedback, i.e., different underlying examination behaviors. As for our future work, we plan to incorporate different click modeling solutions for more accurate document space construction. It would also be meaningful to study how to perform document space based exploratory direction generation, before the interleaved test. Exploratory direction pre-selection is expected to further accelerate the gradient exploration and improve user satisfaction during online learning, but we also need to ensure it is unbiased. We leave the exploration of our projection-based solution to other types of OL2R algorithms as another future work.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This work was supported in part by National Science Foundation Grant IIS-1553568 and IIS-1618948 and Bloomberg Data Science Ph.D. Fellowship.

REFERENCES

- [1] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.
- [2] Olivier Chapelle and Yi Chang. 2011. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge*. 1–24.
- [3] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*. ACM, 1–10.
- [4] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.
- [5] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. 2005. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 385–394.
- [6] Artem Grotov and Maarten de Rijke. 2016. Online learning to rank for information retrieval: SIGIR 2016 Tutorial. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 1215–1218.
- [7] Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke. 2013. Reusing historical interaction data for faster online learning to rank for IR. In *Proceedings of the sixth ACM international conference on WSDM*. ACM, 183–192.
- [8] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. 2011. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 249–258.
- [9] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm, 4–11.
- [10] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. 2016. DCM bandits: Learning to rank with multiple clicks. In *International Conference on Machine Learning*. 1215–1224.
- [11] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. 2015. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 767–776.
- [12] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.
- [13] Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li. 2007. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval*, Vol. 310.
- [14] Harrie Oosterhuis and Maarten de Rijke. 2017. Balancing Speed and Quality in Online Learning to Rank for Information Retrieval. In *Proceedings of the 2017 ACM CIKM*. ACM, 277–286.
- [15] Harrie Oosterhuis and Maarten de Rijke. 2018. Differentiable Unbiased Online Learning to Rank. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18* (2018). <https://doi.org/10.1145/3269206.3271686>
- [16] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *arXiv:cs.LR/1306.2597*
- [17] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does click-through data reflect retrieval quality?. In *Proceedings of the 17th ACM CIKM*. ACM, 43–52.
- [18] Anne Schuth, Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. 2013. Lerot: An online learning to rank framework. In *Proceedings of the 2013 workshop on Living labs for information retrieval evaluation*. ACM, 23–26.
- [19] Anne Schuth, Harrie Oosterhuis, Shimon Whiteson, and Maarten de Rijke. 2016. Multileave gradient descent for fast online learning to rank. In *Proceedings of the Ninth ACM International Conference on WSDM*. ACM, 457–466.
- [20] Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. 2014. Multileaved comparisons for fast online evaluation. In *Proceedings of the 23rd ACM CIKM*. ACM, 71–80.
- [21] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 1. MIT press Cambridge.
- [22] Huazheng Wang, Ramsey Langley, Sonwoo Kim, Eric McCord-Snoek, and Hongning Wang. 2018. Efficient exploration of gradient space for online learning to rank. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 145–154.
- [23] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 115–124.
- [24] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. 2012. The k-armed dueling bandits problem. *J. Comput. System Sci.* 78, 5 (2012), 1538–1556.
- [25] Yisong Yue, Yue Gao, Oliver Chapelle, Ya Zhang, and Thorsten Joachims. 2010. Learning more powerful test statistics for click-based retrieval evaluation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 507–514.
- [26] Yisong Yue and Thorsten Joachims. 2009. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 1201–1208.
- [27] Tong Zhao and Irwin King. 2016. Constructing reliable gradient exploration for online learning to rank. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. ACM, 1643–1652.
- [28] Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. 2017. Online Learning to Rank in Stochastic Click Models. In *International Conference on Machine Learning*. 4199–4208.