Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models

Wu, C.; Zhang, R.; Guo, J.; Chen, W.; Fan, Y.; de Rijke, M.; Cheng, X.

DOI
10.1145/3511808.3557256

Publication date
2022

Document Version
Final published version

Published in
CIKM '22

License
Unspecified

Citation for published version (APA):
Wu, C., Zhang, R., Guo, J., Chen, W., Fan, Y., de Rijke, M., & Cheng, X. (2022). Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models. In CIKM '22: proceedings of the 31st ACM International Conference on Information & Knowledge Management : October 17-21, 2022, Atlanta, GA, USA (pp. 2128-2137). The Association for Computing Machinery. https://doi.org/10.1145/3511808.3557256
Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models

Chen Wu
wuuchen17z@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS University of Chinese Academy of Sciences Beijing, China

Ruqing Zhang
zhangruqing@ict.ac.cn
guojiafeng@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS University of Chinese Academy of Sciences Beijing, China

Wei Chen
chenwei2022@ict.ac.cn
fanyixing@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS University of Chinese Academy of Sciences Beijing, China

Maarten de Rijke
m.derijke@uva.nl
University of Amsterdam Amsterdam, The Netherlands

Xueqi Cheng
cxq@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS University of Chinese Academy of Sciences Beijing, China

ABSTRACT

Neural ranking models (NRMs) have achieved promising results in information retrieval. NRM have also been shown to be vulnerable to adversarial examples. A typical Word Substitution Ranking Attack (WSRA) against NRM was proposed recently, in which an attacker promotes a target document in rankings by adding human-imperceptible perturbations to its text. This raises concerns when deploying NRM in real-world applications. Therefore, it is important to develop techniques that defend against such attacks for NRM. In empirical defenses adversarial examples are found during training and used to augment the training set. However, such methods offer no theoretical guarantee on the models’ robustness, and may eventually be broken by other sophisticated WSRA. To escape this arms race, rigorous and provable certified defense methods for NRM are needed.

To this end, we first define the Certified Top-\(K\) Robustness for ranking models since users mainly care about the top ranked results in real-world scenarios. A ranking model is said to be Certified Top-\(K\) Robust on a ranked list when it is guaranteed to keep documents that are out of the top \(K\) away from the top \(K\) under any attack. Then, we introduce a Certified Defense method, named CertDR, to achieve certified top-\(K\) robustness against WSRA, based on the idea of randomized smoothing. Specifically, we first construct a smoothed ranker by applying random word substitutions on the documents, and then leverage the ranking property jointly with the statistical property of the ensemble to provably certify top-\(K\) robustness. Extensive experiments on two representative web search datasets demonstrate that CertDR can significantly outperform state-of-the-art empirical defense methods for ranking models.

CCS CONCEPTS

- Information systems → Retrieval models and ranking; Adversarial retrieval.

KEYWORDS

Certified Top-\(K\) Robustness, Certified Defense, Word Substitution Ranking Attack, Ranking Models

ACM Reference Format:

Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. 2022. Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models. In Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM ’22), October 17–21, 2022, Atlanta, GA, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557256

1 INTRODUCTION

Neural ranking models (NRMs) [6, 30, 36], especially pre-trained ranking models [22, 27, 34], have led to substantial performance improvements in a wide range of search tasks [14, 20, 27]. We have also seen NRM being used in various practical usages in the enterprise [25]. Despite their success, recent observations [48, 49] show that NRM are vulnerable to adversarial examples. A typical word substitution ranking attack (WSRA) [48] was proposed and proved successful for attacking NRM. In this setting, an attacker promotes a target document in rankings by replacing important words in its text with their synonyms in a semantic-preserving way.
To answer the second question, we propose a novel **Certified Defense method** for Ranking models, CertDR for short, to enhance a model’s certified robustness against WSRA. To avoid exponential computational costs, our method is based on the idea of randomized smoothing [5, 52], which replaces the ranking model with a smoothed ranker for which it is easier to verify the certified robustness. Specifically, we first construct a smoothed ranker by averaging the output ranking scores of randomly perturbed documents. Then, we obtain a certification criterion to judge models’ certified top-K robustness by leveraging the ranking property and statistical property of randomized ensembles. Finally, we design a practical certified defense algorithm, including a noise data augmentation strategy based on the perturbed documents for training and a rigorous statistical procedure to certify the top-K robustness.

We conduct experiments on two web search benchmark datasets, i.e., the MS MARCO document ranking dataset and the MS MARCO passage ranking dataset. We first compare the certified robustness among different ranking models (i.e., traditional probabilistic ranking models, and advanced neural ranking models) under CertDR. Based on the evaluation results, there clearly remains room for future certified robustness improvements. Besides, we compare with several state-of-the-art empirical defense methods and our experimental results show that CertDR can achieve the best defense performance against WSRA.

### 2 RELATED WORK

#### 2.1 Text Ranking Models

Ranking models lie at the heart of IR. Many different ranking models have been proposed over the past decades, including probabilistic models [39, 43] (e.g., BM25 [43]) and learning-to-rank (LTR) models [26]. With the advance of deep learning, we have witnessed a substantial growth of interest in NRMs [6, 30, 36], achieving promising results in a variety of search tasks [14, 20, 27]. Recently, pre-trained ranking models such as BERT-based models [22, 34] have shown substantial performance improvements both in academic research and industry [25]. However, recent observations [48, 49] have shown that NRMs are vulnerable to adversarial examples. In this paper, we study how to defend against adversarial attacks for NRMs.

#### 2.2 Adversarial Attacks

Adversarial attacks aim to generate human-imperceptible adversarial examples by perturbing inputs to maximally increase a model’s risk of making wrong predictions. Adversarial examples were first discovered in the image domain [45], where early research has developed powerful white-box attack methods, e.g., Fast Gradient Sign Method [FGSM, 11] and Projected Gradient Descent [PGD, 28], for attacking continuous image data.

The existence and pervasiveness of adversarial examples have also been observed in the text domain [54]. Despite the fact that generating adversarial examples for texts has proven to be more challenging than for images due to their discrete nature, prior work has explored adversarial attacks for many language tasks, including text classification [10, 24], dialogue systems [4], and natural language inference [1]. Among these attacks, word substitution attacks [1, 41, 53], which replace words in a sentence with their...
We first introduce the WSRA attack we consider in this paper. Then, we introduce the definition of our proposed notion of Certified Top-$K$ Attacks in Web Search. The web search ecosystem is, perhaps, the largest-scale adversarial setting in which search methods operate [13]. For many queries in the web retrieval setting there exists an on-going ranking competition, i.e., many web document authors manipulate their documents deliberately to promote them in rankings [12]. This practice is often referred to as search engine optimization (SEO) [15]. The consequences of SEO are that the quality of search results may rapidly decrease since many irrelevant documents are ranked higher than they deserve.

Very recently, a typical black-box word substitution ranking attack (WSRA) [48] was proposed to simulate such real-world ranking competitions. Specifically, WSRA could successfully attack NRMs by generating human-imperceptible adversarial documents for rank promotion. The synonymous word substitution it employs could maximally maintain the naturalness and semantic similarity of the original document, making it easy for the generated adversarial documents to evade spam detection. Due to the popularity of NRMs and the challenges of defending against human-imperceptible perturbations, we focus on WSRA attacks and design a corresponding defense in this paper.

Notation. In ad-hoc retrieval, given a query $q$ and a set of document candidates $D = \{d_1, d_2, \ldots, d_N\}$ selected from a collection $C$, a ranking model $\mathcal{f}$ aims to predict the relevance score $(f(q, d_n) : n = 1, 2, \ldots, N)$ between every pair of query $q$ and candidate document for ranking the whole candidate set. For example, $f$ outputs the ranked list $L = \{d_{N_1}, d_{N_2}, \ldots, d_1\}$ if it determines $f(q, d_{N_1}) > f(q, d_{N_2}) > \cdots > f(q, d_1)$. In this paper, we assume the ranking score $f(q, d_n)$ is the probability of relevance from 0 to 1 [6], which can be easily achieved by adding a sigmoid operation on the output given by the ranking model.

In WSRA, an attacker replaces the important words in the document with their synonyms by maximizing the adversarial ranking loss to promote the target document in rankings. The number of important words is a hyper-parameter. Specifically, for any word $w$, we consider a pre-defined synonym set $S_w$ containing the synonyms of $w$ (including $w$ itself). Following Ye et al. [52], we assume the synonymous relation is symmetric, that is, $w$ is in the synonym set of any of its synonyms. The synonym set $S_w$ can be built based on GLOVE [37].

Definition 3.1. ($\delta$-Word Substitution Ranking Attack). For an input document $d = \{w_1, w_2, \ldots, w_M\} \in D$, a $\delta$-word substitution ranking attack constructs an adversarial document $d' = (w'_1, w'_2, \ldots, w'_M)$ by perturbing at most $\delta \cdot M$ ($\delta \leq 1$) words in $d$ to any of their synonyms $w'_m \in S_{w_m}$. We denote the candidate set of adversarial documents as $S_d$, i.e.,

$$S_d := \{d' : \|d' - d\|_0/\|d\| \leq \delta\},$$

where $\|d' - d\|_0 := \sum_{m=1}^{M} I\{w'_m \neq w_m\}$ is the Hamming distance, with $I\{\cdot\}$ the indicator function. $\|d\|$ denotes the number of words in the document $d$ and $w'_m \in S_{w_m}$. Ideally, all $d' \in S_d$ have the same semantic meaning as $d$ for human judges, but their ranks may be promoted by the ranking model $f$. The goal of the attacker is to find $d' \in S_d$ such that $f(q, d') > f(q, d)$. Note that we do not attack the documents ranked from 1 to $K$, since there is no need to attack user’s top search results.

3.2 Definition of Certified Top-$K$ Robustness

Certified Top-$K$ Robustness. In general, a model is said to be certified robust if an attack is guaranteed to fail, no matter how the attacker manipulates the input [52]. In a real web search scenario, it is known that users usually care much more about the top ranking results than others [33]. For example, the traffic and click-through
rate (CTR) both fall off as users work their way down the search results in major search engines: while the first and second search results may achieve 36.4% and 12.5% CTR, the 10th search result may achieve a CTR of only 2.2%. Moreover, many widely-used ranking metrics [2, 40] focus on the top-K ranking results, e.g., MRR@K and nDCG@K.

Therefore, protecting the results ranked at the top positions is of great importance [33, 50], not only for real-world applications, but also for the robustness guarantee of widely-used ranking metrics. Inspired by this, we define the Certified Top-K Robustness of ranking models in IR, where a ranking model \( f \) is said to be certified robust at the ranked list \( L \) if it is guaranteed that the documents ranked after top \( K \) will not be attacked to be ranked into top \( K \) in \( L \). Since we focus on the WSRAs in this work, based on this basic definition, we further define Certified Top-K Robustness to WSRA.

**Definition 3.2. (Certified Top-K Robustness to WSRA).** Formally, a ranking model \( f \) is said to be Certified Top-K Robust against WSRAs on the ranked list \( L_q \) with respect to a query \( q \) if it can keep all the documents \( \delta \in L_q[K+1:] \) away from the top-K for all the possible \( \delta \)-word substitution ranking attacks (as defined in Definition 3.1), i.e.,

\[
\text{Rank}_{L_q}(f(q, d')) > K, \text{ for all } d \in L_q[K+1:] \text{ and any } d' \in S_d. \quad (1)
\]

where \( \text{Rank}_{L_q}(f(q, d')) \) denotes the rank position of the adversarial document \( d' \) in \( L_q \) given by the ranking model \( f \). It is highly challenging to judge if \( f \) is certified robust since all the candidate adversarial documents in \( S_d \) should be checked and the size of possible perturbations grows exponentially with \( \delta \). Following existing work [18, 52], we mainly consider the worst case when \( \delta = 1 \), which is the most challenging case.

4 OUR CERTIFIED DEFENSE METHOD

Based on the definition of certified top-K robustness to WSRA, we introduce a novel Certified Defense method for Ranking models (CertDR) to enhance the certified robustness. We first introduce a randomized smoothing function for ranking and how to use it to certify the robustness theoretically. Then, we propose a practical certified defense algorithm for ranking models. Proofs are at the end of this section.

4.1 Randomized Smoothing Function for Ranking

To circumvent the computationally expensive combinatorial optimization (e.g., enumerating all the candidate adversarial documents in \( S_d \)), we borrow the idea from the randomized smoothing technique [5, 52], which could provably defend against the adversarial attacks by leveraging the voting of randomly perturbed samples derived from the original input. We target to replace the ranking model \( f \) with a smoothed ranking model \( \tilde{f} \) for which it is easier to verify the certified robustness.

Specifically, we construct the smoothed ranker \( \tilde{f} \) by averaging the output ranking scores of a set of randomly perturbed documents based on random perturbations, i.e.,

\[
\tilde{f}(q, d) = \mathbb{E}_{R \sim \Pi_d} f(q, R),
\]

where \( R \) is a randomly perturbed document and \( \Pi_d \) is the corresponding probability distribution that prescribes a random perturbation around \( d \). In our work, we define \( \Pi_d \) to be the uniform distribution on a set of random word substitutions following [52].

In previous classification tasks [5, 52], the output of the smoothed classifier is the class with the largest probability “voting” by all randomly perturbed inputs. Different from these works, we compute the output of the smoothed ranker by averaging the ranking scores of all randomly perturbed documents originated from the \( d \), which is more suitable for the ranking problem. In this way, we could obtain the ranked list \( \tilde{L}_q^K \) based on the averaged scores produced by the smoothed ranker \( \tilde{f} \). Here, we leave the query \( q \) free from attack. In the future work, we would like to explore the defense against query attacks by focusing on \( q \) in this formulation.

To obtain random perturbations in defense methods effectively, we propose to build a perturbation set \( T_w \) for each word \( w \). Specifically, we construct \( T_w \) from the synonym set \( S_w \) used in the attack method, i.e., WSRA in this work, by choosing the top \( J \) nearest neighbors via the cosine similarity of GLOVE vectors. Then, for a document \( d = (w_1, w_2, ..., w_M) \), we define the perturbation distribution \( \Pi_d \) by perturbing each word \( w_i \) in \( d \) to a word in its perturbation set \( T_{w_i} \) randomly and independently, i.e.,

\[
\Pi_d(R) = \prod_{i=1}^{M} \frac{|T_{w_i} \cap R|}{|T_{w_i}|},
\]

where \( R = (r_1, ..., r_M) \) is the perturbed document and \( |T_{w_i}| \) denotes the size of \( T_{w_i} \).

4.2 Certifying Smoothed Ranking Models

Given the smoothed ranking model \( \tilde{f} \), in this section, we introduce how to certify its top-K robustness. For all the documents in \( L_q^K[K+1:] \), if their adversarial documents could achieve lower scores than the document \( d_K \) ranked at the position \( K \), we think these documents \( L_q^K[K+1:] \) cannot be attacked into top \( K \). Formally, the condition that \( \tilde{f} \) is certified top-K robust on \( L_q^K \) can be defined as,

\[
\max_{d \in L_q^K[K+1:] \text{ and any } d' \in S_d} \tilde{f}(q, d') < \tilde{f}(q, d_K). \quad (2)
\]

In general, there are two difficulties to complete the above certification case by case, i.e., the inner maximum and the outer maximum. **Inner Maximum.** The first difficulty is to examine all candidate adversarial documents in \( S_d \) for the inner maximum, where the computation cost grows exponentially with the attacked word number \( \delta M \). We address this problem in the following Theorem 4.1.

**Theorem 4.1 (Certified Upper Bound).** Assume that the perturbation set \( T_w \) is constructed such that \( |T_w| = |T_{w'}| \) for every word \( w \) and its synonym \( w' \) in \( S_w \). Define

\[
\sigma_w = \min_{w' \in S_w} \frac{|T_w \cap T_{w'}|}{|T_w|},
\]

where \( \sigma_w \) indicates the overlap between the two different perturbation sets. For a given document \( d = (w_1, ..., w_M) \), we sort the words according to \( \sigma_w \), such that \( \sigma_{w_1} \leq \sigma_{w_2} \leq \cdots \leq \sigma_{w_M} \). Then

\[
\max_{d' \in S_d} \tilde{f}(q, d') \leq \min_{d' \in S_d} \tilde{f}(q, d) + (1 - \sigma_w_1) \leq \min_{d' \in S_d} \tilde{f}(q, d) + \sigma_w 1,
\]

where \( \sigma_d := 1 - \prod_{j=1}^{E} \sigma_{w_j} \),
The idea is that for any adversarial document $d' \in S_d$, the upper bound of $\tilde{f}(q, d')$ can be bounded by its original document ranking score $f(q, d)$ with randomized smoothing. As a result, this theorem avoids the difficult adversarial optimization of $\tilde{f}(q, d')$ on $d' \in S_d$, and only needs to evaluate $\tilde{f}(q, d)$ at the original document $d$. Note that the difference between our Theorem 4.1 with Theorem 1 in [52] is that we extend the classification prediction in Theorem 1 to ranking scores between queries and documents, and prove the theorem in an upper bound situation, which has not been proved by Ye et al. [52]. The proof is provided in Section 4.4.1. Besides, the upper bound in Theorem 4.1 is sufficiently tight, which is shown in Section 4.4.2.

**Outer Maximum.** The second difficulty is to examine all documents in $L_q^K[1 : K]$ for the outer maximum, where the computational costs grow linearly with the list length $N$. We address the outer maximum in the following. To simplify our notation, we define $A_L = \tilde{f}(q, d_K) = \max_{d \in L_q^K[1 : 1]} \max_{d' \in S_d} \tilde{f}(q, d')$, where Eq. (2) is equivalent to $A_L > 0$. By applying Theorem 4.1 to Eq. (2), we have

$$A_L \geq \tilde{f}(q, d_K) \geq \max_{d \in L_q^K[K : 1]} (\tilde{f}(q, d) + o_d)$$

$$= \tilde{f}(q, d_K) - \tilde{f}(q, d_{K+1}) - \max_{d \in L_q^K[K : 1]} o_d,$$

where $d_{K+1}$ denotes the document which is ranked at the position $K + 1$. The proof is achieved by utilizing the ranking property that $\tilde{f}(q, d_1) > \tilde{f}(q, d_2) > \cdots$. The idea is that we can compute $A_L$ (in other words, certifying $\tilde{f}$) by comparing the ranking scores of documents ranked at $K$ and $K + 1$. Note that the computational cost of $\max_{d \in L_q^K[1 : K + 1]} o_d$ is negligible.

Based on the above solutions of the inner and outer maximum in Eq. (2), it is possible to introduce a certification criterion for checking the certified top-$K$ robustness for ranking models.

**Proposition 4.1.** For a ranked list $L_q^K$ with respect to a query $q$, under the condition of Theorem 4.1, we can certify that $\Delta L_q$ is computed, where if $\Delta L_q > 0$, we can certify that $\tilde{f}$ is top-$K$ robust at the ranked list $L_q^K$.

**4.3 Practical Certified Defense Algorithm**

Based on the above theoretical analysis, we now present a practical certified defense algorithm for ranking models. Formally, we write $S_C$ for the synonym dictionary of the collection. Specifically, $S_C$ contains the synonym set $S_w$ for all words in the collection $C$ and is often presented as a synonym network [52]. Similar to the process of obtaining $T_w$ from $S_w$, we achieve the collection perturbation dictionary $T_C$ by keeping the top $f$ nearest neighbors in $S_C$ for each word. The overall architecture is shown in Figure 2.

The certified defense algorithm contains two key steps, i.e., noised data augmentation and Top-$K$ Robustness Certification. We describe the two steps in the following.

**Noise Data Augmentation Strategy.** The robustness certification holds regardless of how the original ranker $f$ is trained. However, to rank the document $d$ with respect to the query correctly and robustly by $f$, it is expected that $f$ properly ranks the perturbed document $R$ (recall that $R \sim \Pi_d$) such that it is close to the rank position of the original document $d$. That is, the noise of $R$ should have little effect on the ranking, making the ensembled ranking score $f(q, d)$ close to the original ranking score $f(q, d)$. However, if $f$ is trained via standard supervised learning without any noised documents, it will not necessarily learn how to rank $R$ properly.

Inspired by previous works [23, 52, 55], we introduce a noise data augmentation strategy for ranking. Specifically, we first generate a perturbed document $d_{\text{noised}}$ for each $d$ in the collection. The perturbation is achieved by randomly sampling every word from $d$ using the perturbation distribution $\Pi_d$. Then, we train the original ranker $f$ using the training triples equipped with the noised documents via the following objective:

$$L_{\text{dat}} = \max(0, 1 - f(q, d_{\text{noised}}^+ + f(q, d_{\text{noised}}^-)),$$

where $d_{\text{noised}}^+$/$d_{\text{noised}}^-$ is the perturbed document from $d^+$/d$. And $d^+$/d$^-$ denotes the positive/negative document in original training triples. Then, we can obtain a better smoothed rank $\tilde{f}$ by Monte Carlo estimation in the following.

**Top-$K$ Robustness Certification.** In theory, since the perturbation space can be extremely large, it is impossible to exactly obtain...
the prediction of \( \hat{f} \) at each \((q, d)\). Therefore, based on the ranker \( f \) obtained from the noisy data augmentation strategy, we estimate \( \hat{f}(q, d_k) \) and \( \hat{f}(q, d_{k+1}) \) by Monte Carlo estimation. Take \( \hat{f}(q, d_K) \) as an example, we can estimate it like

\[
\hat{f}(q, d_K) = \mathbb{E}_{R_k \sim d_k} f(q, d_K) = \frac{1}{n} \sum_{i=1}^{n} f(q, R_k^{(i)}),
\]

where \( R_k^{(i)} \) are i.i.d. samples from \( \Pi_{d_k} \) and thus \( \Delta L_q \) can be approximated accordingly. We can construct rigorous statistical procedures to reject the null hypothesis that \( \hat{f} \) is not certified robust at \( L_q \) (e.g., \( \Delta L_q < 0 \)) with a given significance level (e.g., 5\%) following \([52]\).

Finally, if \( \Delta L_q > 0 \), then \( \hat{f} \) is not certified top-\( K \) robust at the ranked list \( L_q^\lambda \). Otherwise, we will judge it is not certified top-\( K \) robust at the \( L_q^\lambda \). We can see that our practical certified defense algorithm could be achieved by assembling the ranking outputs and that it does not require any further information about the ranking models. Thus, it can be applied to any ranking model.

### 4.4 Proofs

Here we provide all the necessary proofs of Theorem 4.1 and the tightness of the bound in Theorem 4.1 with its proof. The upper bound of Theorem 4.1 is achieved by introducing an auxiliary function cluster based on the relevance between the query and document, and solving the constraint optimization problem by Lagrange and properties of randomly perturbed sets. Tightness is proved by constructing the randomized smoothing ranker that satisfies the desired property we want.

#### 4.4.1 Proof of Theorem 4.1

Our goal is to calculate the upper bound \( \max_{d \in S_d} \hat{f}(q, d') \). The key idea is to frame the computation of the upper bound into a variational optimization.

**Lemma 4.1.** Define \( \mathcal{G}_{[0,1]} \) to be the set of all bounded functions mapping from \( Q \times D \) to \([0, 1]\). For any \( g \in \mathcal{G}_{[0,1]} \), define

\[
\Pi_d[g] = \mathbb{E}_{R \sim \Pi_d} [g(q, R)].
\]

Then we have for any \( R \),

\[
\max_{d \in S_d} \hat{f}(q, d') = \max_{d \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \{ \Pi_d[g] \ such \ that \ \Pi_d[g] = \hat{f}(q, d) \}. \tag{2133}
\]

**Proof of Lemma 4.1.** Define \( g_0(q, d) = f(q, d) \). Then we have

\[
\hat{f}(q, d) = \mathbb{E}_{R \sim \Pi_d} [f(q, R)] = \Pi_d[g_0].
\]

Therefore, \( g_0 \) satisfies the constraints in the optimization, which makes it obvious that

\[
\hat{f}(q, d') = \mathbb{E}_{R \sim \Pi_d} [f(q, R)] = \max_{g \in \mathcal{G}_{[0,1]}} \{ \Pi_d[g] \ such \ that \ \Pi_d[g] = \hat{f}(q, d) \}.
\]

Taking \( \max_{d \in S_d} \) on both sides yields the upper bound and thus the problem reduces to deriving bounds for the optimization problems.

**Theorem 4.2.** Under the assumption of Theorem 4.1, for the optimization problem in Lemma 4.1, we have

\[
\hat{f}_{up}(q, d') \leq \min(\hat{f}(q, d) + o_d, 1),
\]

where \( o_d \) is the quantity defined in Theorem 4.1.

**Proof of Theorem 4.2.** For notation, we denote \( p = \hat{f}(q, d) \). Applying the Lagrange multiplier to the constraint optimization problem and exchanging the min and max, we have

\[
\hat{f}_{up}(q, d') = \max_{d \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \{ \Pi_d[g] \ such \ that \ \Pi_d[g] = \hat{f}(q, d) \}
\]

where \( \Delta L_q < 0 \) with a given significance level (e.g., 5\%) following \([52]\).

Finally, if \( \Delta L_q > 0 \), then \( \hat{f} \) is not certified top-\( K \) robust at the ranked list \( L_q^\lambda \). Otherwise, we will judge it is not certified top-\( K \) robust at the \( L_q^\lambda \). We can see that our practical certified defense algorithm could be achieved by assembling the ranking outputs and that it does not require any further information about the ranking models. Thus, it can be applied to any ranking model.

#### 4.4.1 Proof of Theorem 4.1

Our goal is to calculate the upper bound \( \max_{d \in S_d} \hat{f}(q, d') \). The key idea is to frame the computation of the upper bound into a variational optimization.

**Lemma 4.1.** Define \( \mathcal{G}_{[0,1]} \) to be the set of all bounded functions mapping from \( Q \times D \) to \([0, 1]\). For any \( g \in \mathcal{G}_{[0,1]} \), define

\[
\Pi_d[g] = \mathbb{E}_{R \sim \Pi_d} [g(q, R)].
\]

Then we have for any \( R \),

\[
\max_{d \in S_d} \hat{f}(q, d') = \max_{d \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \{ \Pi_d[g] \ such \ that \ \Pi_d[g] = \hat{f}(q, d) \}. \tag{2133}
\]

**Proof of Lemma 4.1.** Define \( g_0(q, d) = f(q, d) \). Then we have

\[
\hat{f}(q, d) = \mathbb{E}_{R \sim \Pi_d} [f(q, R)] = \Pi_d[g_0].
\]

Therefore, \( g_0 \) satisfies the constraints in the optimization, which makes it obvious that

\[
\hat{f}(q, d') = \mathbb{E}_{R \sim \Pi_d} [f(q, R)] = \max_{g \in \mathcal{G}_{[0,1]}} \{ \Pi_d[g] \ such \ that \ \Pi_d[g] = \hat{f}(q, d) \}.
\]

Taking \( \max_{d \in S_d} \) on both sides yields the upper bound and thus the problem reduces to deriving bounds for the optimization problems.

**Theorem 4.2.** Under the assumption of Theorem 4.1, for the optimization problem in Lemma 4.1, we have

\[
\hat{f}_{up}(q, d') \leq \min(\hat{f}(q, d) + o_d, 1),
\]

where \( o_d \) is the quantity defined in Theorem 4.1.

**Proof of Theorem 4.2.** For notation, we denote \( p = \hat{f}(q, d) \). Applying the Lagrange multiplier to the constraint optimization problem and exchanging the min and max, we have

\[
\hat{f}_{up}(q, d') = \max_{d \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \{ \Pi_d[g] \ such \ that \ \Pi_d[g] = \hat{f}(q, d) \}
\]

where \( \Delta L_q < 0 \) with a given significance level (e.g., 5\%) following \([52]\).

Finally, if \( \Delta L_q > 0 \), then \( \hat{f} \) is not certified top-\( K \) robust at the ranked list \( L_q^\lambda \). Otherwise, we will judge it is not certified top-\( K \) robust at the \( L_q^\lambda \). We can see that our practical certified defense algorithm could be achieved by assembling the ranking outputs and that it does not require any further information about the ranking models. Thus, it can be applied to any ranking model.
Lemma 4.3. For any word $w$, define $\hat{w}^* = \arg \min_{w \in S_d} n_{w,w} / n_w$. For a given document $d = (w_1, \ldots, w_M)$, we define an ordering of the words $w_{P_1}, \ldots, w_{P_N}$ such that $n_{w_{P_i}, w_{P_{i+1}}} / n_{w_{P_i}} \leq n_{w_{P_i}, w_{P_{i+1}}} / n_{w_{P_{i+1}}}$ for any $i \leq j$. For a given $d$ and $E = \mathcal{E}_M$, we define an adversarial perturbed document $\tilde{d}^* = (w_1^*, \ldots, w_M^*)$, where

$$w_i^* = \begin{cases} 
\hat{w}_i & \text{if } i \in [w_1, \ldots, w_E] \\
w_i & \text{if } i \notin [w_1, \ldots, w_E]
\end{cases}$$

Then for any $\lambda > 0$, we have that $d^*$ is the optimal solution of the problem:

$$\max_{d \in S_d} \int (d \Pi_d(R) - \lambda d \Pi_d(R))_+$$

Now by Lemma 4.3, the upper bound becomes

$$\bar{f}_\mu(q, d^*) = \min_{\lambda \geq 0} \frac{n_{w_i, w_j}}{n_{w_j}} = 1 - \sum_{j=1}^E \min_{j \in [E], w_j \neq \hat{w}_j} \frac{n_{w_i, w_j}}{n_{w_j}}$$

$$= \min_{\lambda \geq 0} \frac{n_{w_i, w_j}}{n_{w_j}} = 1 - \sum_{j=1}^E \min_{j \in [E], w_j \neq \hat{w}_j} \frac{n_{w_i, w_j}}{n_{w_j}}$$

$$= \min_{\lambda \geq 0} \frac{n_{w_i, w_j}}{n_{w_j}} = 1 - \sum_{j=1}^E \min_{j \in [E], w_j \neq \hat{w}_j} \frac{n_{w_i, w_j}}{n_{w_j}}$$

$$= \min_{\lambda \geq 0} \frac{n_{w_i, w_j}}{n_{w_j}} = 1 - \sum_{j=1}^E \min_{j \in [E], w_j \neq \hat{w}_j} \frac{n_{w_i, w_j}}{n_{w_j}}$$

$$= \min_{\lambda \geq 0} \frac{n_{w_i, w_j}}{n_{w_j}} = 1 - \sum_{j=1}^E \min_{j \in [E], w_j \neq \hat{w}_j} \frac{n_{w_i, w_j}}{n_{w_j}}$$

Here, Eq. (5) is calculated using the assumption of Theorem 4.1. The optimization of $\min_{\lambda \geq 0}$ in (5) is an elementary step: if $p + a_d > 1$, we have $\lambda^* = 0$ with solution 1; if $p + a_d < 1$, we have $\lambda^* = 1$ with solution $p + a_d$. For the proof of Lemma 4.2 and 4.3, we refer readers to [52, Lemma 2 and 3].

4.4.2 Tightness. Whether the bound in Theorem 4.4 is sufficiently tight is of great importance. In the following, we provide a theorem to state its tightness.

Theorem 4.3 (Tightness). Assume the conditions of Theorem 4.1 hold. For a ranking model $f$ that maps $O \times D$ to $[0, 1]$, there exists a model $\hat{f}$ such that its related smoothed ranker $\bar{f}_\mu$ satisfies

$$\bar{f}_\mu(q, d) = f(q, d),$$

and

$$\max_{d \in S_d} \bar{f}_\mu(q, d^*) = \min_{\lambda \geq 0} (\bar{f}_\mu(q, d) + a_d, 1),$$

where $a_d$ is defined in Theorem 4.1.

As shown in Theorem 4.3, the upper bound in Theorem 4.1 is tight and can not be further improved if we do not know any other structural information about $f$. In the following, we provide the proof of Theorem 4.3.

Proof of Tightness. We denote $\bar{f}_\mu(q, d) = p_r$ in this proof for simplicity. The $d^*$ below is the optimal adversarial document defined in the proof of Lemma 4.3. Note that $a_d = |T_d - T_{d^*}| / |T_d|$ and $1 - a_d = |T_d \cap T_{d^*}| / |T_d|$ as defined in Theorem 4.1. Our proof is based on constructing a randomized smoothing ranker that satisfies the desired property we want to prove.

Case 1 $p_r \leq 1 - a_d$. Note that in this case $|T_d \cap T_{d^*}| = 1 - a_d \geq p_r$. Therefore, we can choose set $U$ such that $U \subseteq T_d \cap T_{d^*}$ and $|U| / |T_d| = p_r$. We define the ranker:

$$\bar{f}_\mu(q, R) = \begin{cases} 
1, & \text{if } R \in U \cup T_{d^*} \\
0, & \text{otherwise}
\end{cases}$$

Case 2 $p_r > 1 - a_d$. We choose set $U$ such that $U \subseteq T_d \cap T_{d^*}$ and $|U| / |T_d| = p_r$. We define the ranker:

$$\bar{f}_\mu(q, R) = \begin{cases} 
1, & \text{if } R \in U \cup (T_{d^*} - T_d) \\
0, & \text{otherwise}
\end{cases}$$

It can easily be verified that for each case, the defined ranker satisfies all the conditions in Theorem 4.3. This indicates the bound can be achieved by learning a gold ranker that can judge some specific documents as relevant and others as irrelevant for the query.

5 EXPERIMENTAL SETUP

5.1 Datasets

To evaluate the effectiveness of our proposed methods, we conduct experiments on two representative web search benchmark datasets.

- MS MARCO Document Ranking dataset [32] (MS-MARCO-Doc) is a large-scale benchmark dataset for web document retrieval, with about 3.21M web documents and 0.37M training queries. The average length of the document is about 1129.

- MS MARCO Passage Ranking dataset [32] (MS-MARCO-Pas) is a large-scale benchmark dataset for passage retrieval, with about 8.84M passages from web pages and 0.5M training queries. The average length of the passage is about 58.

5.2 Baselines

Our CertDr can certify the top-K robustness of different ranking models. We first compare the certified top-K robustness among different ranking models (i.e., BM25 [43], Duet [30] and BERT [7]) under CertDr. Then, since the defense methods for NRMs have not been well explored yet, we adopt the representative empirical defense, i.e., Data Augmentation (DA), in text classification task [17, 42], for NRMs as a baseline. Specifically, for each training document $d$, we augment the collection with 2 new documents $\tilde{d}$ by sampling $\tilde{d}$ uniformly from $S_d$, then train on the normal hinge loss following [18]. We do not use adversarial training [11] here because it would require running an adversarial search procedure at each training step, which would be prohibitively slow.

5.3 Evaluation Metrics

We evaluate the robustness to all WSRAs of models. We propose a metric to directly measure the Certified Robust Query (CRQ) percentage, the percentage of test queries for which the model is certified robust at the query $q$ if all the documents out of top-K are not attacked into the top-K. Evaluating this exactly involves enumerating exponentially many perturbations, which is intractable (Section 4.2). Instead, we evaluate the CRQ under randomized smoothing, i.e.,

$$CRQ = \frac{\sum_{q \in Q} \mathbb{1} (\Delta_{q} > 0)}{|Q|},$$

where $\Delta_{q}$ is the criterion mentioned in Section 4.2. The ranking model is more certified robust with a higher CRQ (%)
To compare the defense ability of CertDR with empirical defense methods, we also leverage two metrics, i.e., success rate and conditional success rate. **Success Rate (SR)** [48] evaluates the percentage of the after-attack documents that are ranked higher than original documents. The robustness of a ranking model is better with a lower SR (%).

Inspired by CondAcc [46], which enables the comparison certified and empirical defense, we introduce **Conditional Success Rate (CondSR)**. CondSR evaluates whether the rankings of the adversarial documents in an attacked ranked list indeed cannot be improved when its counterpart clean ranked list is certified robust:

$$
\text{CondSR} = \frac{\sum_{q \in Q} \mathbb{I}\{\Delta L_q > 0\} \frac{1}{N_q} \sum_{i=1}^{N_q} \mathbb{I}\{\text{Rank}_L(d_i + p) < \text{Rank}_L(d_i)\}}{\sum_{q \in Q} \mathbb{I}\{\Delta L_q > 0\}}.
$$

While CRQ is evaluated on clean ranked lists to show certified robustness, CondSR is tested on attacked ranked lists to show the empirical robustness of models on these certified ranked lists. The robustness of a ranking model is better with a lower CondSR (%).

### 5.4 Implementation Details

We implement ranking models following previous work [6, 49]. For the MS-MARCO-Doc collection, we use the official top 100 (i.e., $N = 100$) ranked documents retrieved by the QL model. For the MS-MARCO-Pas, initial retrieval is performed using the Anserini toolkit [51] with the BM25 model to obtain the top 100 ranked passages. We evaluate all ranking models on 200 queries (i.e., $|Q| = 200$) randomly sampled from the dev set of two datasets following [48].

For the Monte Carlo estimation of $\Delta L_q$, we use 1,000 random perturbed documents to accept $\Delta L_q > 0$ with probability of at least 0.95. The corresponding estimation error is 0.086 and is considered during the estimation following Ye et al. [52]. Further implementation details and the code can be found online.²

### 6 EXPERIMENTAL RESULTS

We evaluate our defense method to address the following research questions: *(RQ1)* What is the certified robustness among different ranking models via CertDR? *(RQ2)* How does the randomized smoothed ranker perform compared with the original ranker? *(RQ3)* How does $K$ affect certified top-$K$ robustness? *(RQ4)* How does CertDR perform compared with empirical defense baselines?

#### 6.1 Certified Top-$K$ Robustness of Different Ranking Models

To answer *(RQ1)*, we analyze certified top-$K$ robustness of different ranking models using CertDR on MS-MARCO-Doc and MS-MARCO-Pas. See Figure 3. We find that (i) Overall, the certified robustness of the ranking model is lower than that of the text classification models [18, 52], indicating that ranking models are vulnerable to adversarial attacks. There are two potential reasons: First, the text ranking task itself needs to model cross-document interactions to capture query-document relevance, which is more complex than classifying a single sentence independently as in text classification. Second, certified top-$K$ robustness imposes requirements on the

---

²https://github.com/ict-bigdatalab/CertDR/

---

![Figure 3: Certified top-$K$ robustness of different ranking models in terms of CRQ (%) on MS-MARCO-Doc (a) and MS-MARCO-Pas (b).](image)

ranked list, which is demanding than the point-wise classification scenario. (ii) Pre-trained model BERT generally outperforms other models, indicating that BERT is more certified robust than other ranking models. The reason might be that pre-training on a large text corpus can improve the out-of-distribution generalizability to adversarial examples attacked by synonyms substitution. Therefore, it is worthwhile to leverage pre-training techniques to enhance the robustness of NRM in the future. (iii) BM25 is less certified robust than Duet and BERT on MS-MARCO-Doc when $K$ is small, while it is more certified robust than Duet on MS-MARCO-Pas for all $K$s. BM25 depends on exact matching signals between the query and document; therefore, a possible explanation is that there are fewer options of attacked words in the short passage than the long document, contributing to the robustness on short texts. Besides, we leave the analysis of different $K$s to Section 6.3.

### Table 1: Comparing the ranking performance between the original and randomized smoothed ranker in terms of MRR@10 and MRR@100 on MS-MARCO-Doc. * denotes significant degradation w.r.t. the randomized smoothed ranker w/o noise data augmentation (p-value<0.05).

| Method               | MRR@10  | MRR@100 |
|----------------------|---------|---------|
| Original $f$         | 0.4428* | 0.4470* |
| Smoothed $\hat{f}$ w/o noise data aug | 0.2259 | 0.2416 |
| Smoothed $\hat{f}$   | 0.3635* | 0.3722* |

#### 6.2 Smoothed Ranker vs. Original Ranker

To answer *(RQ2)*, we compare the ranking performance of the randomized smoothed ranker with the original ranker. We select BERT as the original ranker and conduct experiments on MS-MARCO-Doc. We also show the ranking performance of the randomized smoothed ranker without the noised data augmentation.

From Table 1, we observe that: (i) The ranking performance of smoothed ranker without noised data augmentation drops dramatically (e.g., 0.2259 vs. 0.4428 in terms of MRR@100). The reason might be that the smoothed ranker ranks documents based on the ensemble ranking scores of perturbed documents, which are far away from the original documents. (ii) By applying a noised data augmentation strategy, the ranking performance of the smoothed ranker improves significantly and becomes closer to the original ranker. The reason might be that the augmented training documents are generated from the same perturbation distribution with
the perturbed documents, which helps the smoothed ranker learn to rank the perturbed documents properly. (iii) The smoothed ranker has a moderate drop in terms of MRR@100 compared with the original ranker with normal training (0.3722 vs. 0.4470). Similar drops on clean acc (accuracy on clean examples) are also seen for robust models in previous work [18, 31]. Future work could explore how to achieve the trade-off between clean and robust performance.

Table 2: The CRQ (%) of different ranking models with different K on MS-MARCO-Doc and MS-MARCO-Pas.

| K  | MS-MARCO-Doc | MS-MARCO-Pas |
|----|--------------|--------------|
|    | BM25 | Duet | BERT | BM25 | Duet | BERT |
| 1  | 3.0  | 9.5  | 15.5 | 12.5 | 7.0  | 14.0 |
| 2  | 2.0  | 1.5  | 12.5 | 6.0  | 2.0  | 15.0 |
| 3  | 1.5  | 1.5  | 11.5 | 8.0  | 0.5  | 14.5 |
| 4  | 3.5  | 1.5  | 10.0 | 5.5  | 0    | 7.5  |
| 5  | 3.5  | 0.5  | 13.0 | 4.5  | 0.5  | 15.0 |
| 10 | 1.5  | 0    | 3.0  | 2.5  | 0    | 9.5  |
| 20 | 1.5  | 0    | 0    | 0.5  | 0    | 3.0  |
| 30  | 0.5  | 0    | 0    | 0    | 1.5  |
| 40  | 0.5  | 0    | 0    | 1.5  | 0    | 0    |
| 50  | 0.5  | 0    | 0    | 0.5  | 0    | 0    |
| 60  | 0.5  | 0    | 0    | 0    | 0    | 0    |
| 70  | 0    | 1    | 0    | 1    | 0    | 0    |
| 80  | 0    | 0    | 0    | 0    | 0    | 0    |
| 90  | 0    | 0    | 0    | 0.5  | 0    | 0    |

6.3 Analysis of the Effect of K

To answer RQ3, we analyze the effect of K for CertDR when we certify the top-K robustness. Specifically, we analyze the ranking performance of different ranking models in terms of CRQ, and set K to 14 different values. As shown in Table 2, we can find that: (i) Overall, the model becomes less certified top-K robust with the increase of K on both datasets. Intuitively, it is more difficult to attack a document to a higher rank position than a lower rank position. (ii) However, an interesting finding is that the certified top-K robustness with a larger K is even greater than a smaller K in a certain range. For example, the CRQ of BERT is 15.0 with K = 5 while 7.5 with K = 4 on the MS-MARCO-Pas dataset. By conducting further analysis, we find that although documents ranked out of top 5 are not easily attacked into the top 5, the 4-th document could be attacked into the top 4 easily. (iii) While the CRQ of NMRs reduces to 0 after a point (e.g., the CRQ of Duet reduces to 0 after K=10), it is interesting to find that the CRQ of BM25 remains at a low positive value when K is very large (e.g., CRQ of BM25 remains at 0.5 when K=30 to 90 on MS-MARCO-Doc). The reason may be that BM25 relies on the statistical features, which is more robust than word embeddings of NMRs against adversarial attacks. This is consistent with the findings in [49].

6.4 Comparison with Empirical Defenses

Based on the above analysis of certified robustness of different models, we further compare CertDR with baseline empirical defense methods (i.e., DA) following [46]. The WSRA is conducted by PRADA [48], and we set K = 10 and 5 for CertDR on MS-MARCO-Doc and MS-MARCO-Pas, respectively.

Table 3: Comparisons between our proposed CertDR and the baseline on the BERT. Adversarial attacks are conducted by PRADA [48]. ADV corresponds to no defense. ADV and DA are evaluated under SR (%) and CertDR is evaluated under CondSR (%).

| Dataset          | ADV | DA | CertDR |
|------------------|-----|----|--------|
| MS-MARCO-Doc     | 96.7 | 57.0 | 40.0 |
| MS-MARCO-Pas     | 91.4 | 64.6 | 57.4 |

7 CONCLUSION

In this paper, we defined the notion of Certified Top-K Robustness for ranking models focusing on the characteristics of IR. We proposed a certifiably robust defense method called CertDR, based on randomized smoothing. The key idea is to smooth the ranking model with random word substitutions, and construct provable certification bounds based on the ranking property. Extensive experiments validate that CertDR outperforms existing defense methods and improves the certifiable robustness of ranking models.

In future work, it is worth to strengthen the notion of Certified Top-K Robustness to guarantee that the order of top-K ranking results remains unchanged. We hope that our study helps to put concerns about the robustness of NMRs on the research agenda and to motivate new defense ideas, including empirical and certified defenses of ranking models.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62006218 and 61902381, the Youth Innovation Promotion Association CAS under Grants No. 20144310 and 2021100, the Innovation Project of ICT, CAS under Grants No. E261090, the Young Elite Scientist Sponsorship Program by CAST under Grants No. YESS20200121, and the Lenovo-CAS Joint Lab Youth Scientist Project. This work was also (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Dutch Research Council, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.
