Certified Robustness to Word Substitution Attack with Differential Privacy

Wenjie Wang, Pengfei Tang, Jian Lou* and **Li Xiong**
Emory University
Atlanta, GA, USA
{wang.wenjie, pengfei.tang, jian.lou, lxiong}@emory.edu

Abstract

The robustness and security of natural language processing (NLP) models are significantly important in real-world applications. In the context of text classification tasks, adversarial examples can be designed by substituting words with synonyms under certain semantic and syntactic constraints, such that a well-trained model will give a wrong prediction. Therefore, it is crucial to develop techniques to provide a rigorous and provable robustness guarantee against such attacks. In this paper, we propose *WordDP* to achieve certified robustness against word substitution attacks in text classification via differential privacy (DP). We establish the connection between DP and adversarial robustness for the first time in the text domain and propose a conceptual exponential mechanism-based algorithm to formally achieve the robustness. We further present a practical simulated exponential mechanism that has efficient inference with certified robustness. We not only provide a rigorous analytic derivation of the certified condition but also experimentally compare the utility of *WordDP* with existing defense algorithms. The results show that *WordDP* achieves higher accuracy and more than $30\times$ efficiency improvement over the state-of-the-art certified robustness mechanism in typical text classification tasks.

1 Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performance in many natural language processing (NLP) tasks, such as text classification (Zhang et al., 2015), sentiment analysis (Bakshi et al., 2016), and machine translation (Bahdanau et al., 2014), making the robustness and security of NLP models significantly important. Recent studies have shown that DNNs can be easily fooled by adversarial examples, which are carefully crafted by adding imperceptible perturbations to input examples during inference time (Szegedy et al., 2013). In the context of text classification tasks, adversarial examples can be designed by manipulating the word or characters under certain semantic and syntactic constraints (Ren et al., 2019; Jin et al., 2019; Zang et al., 2020; Gao et al., 2018). Among all the attack strategies, word substitution attacks, in which attackers attempt to alter the model output by replacing input words with their synonyms, can maximally maintain the naturalness and semantic similarity of the input. Therefore, in this paper, we consider such word substitution attacks and focus on defending against such attacks. Figure 1 shows an example of the word substitution attack where the clean input text is changed into adversarial text by substituting input words from a synonym list. Various mechanisms have been developed to

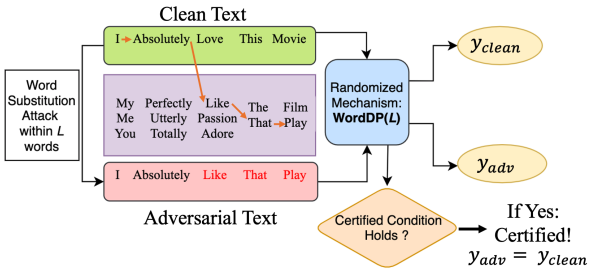


Figure 1: Word Substitution Attack and Certified Robustness via *WordDP*.

defend against adversarial examples in text classification models. Miyato *et al.* (2016) applied adversarial training to the text domain that involves adversarial examples in the training stage. Data augmentation in the training phase is another defense approach to improve model robustness. For example, Synonyms Encoding Method (SEM) proposed by Wang *et al.* (2019), Dirichlet Neighborhood Ensemble (DNE) proposed by Zhou *et al.* (2020), and Robust Encodings (RobEn) proposed by Jones *et al.* (2020) are different data augmentation methods on either embedding space or word space. How-

*J. Lou is the corresponding author.

ever, all the above-mentioned works are only evaluated empirically and have no theoretical analysis or guarantee on the robustness of the methods in that they may be broken by other adaptive attacks. Therefore, it is important to provide rigorous and provable certified defense.

There are several attempts to achieve certified robustness for word substitution attacks. Jia *et al*. (2019) and Huang *et al*. (2019) utilize Interval Bound Propagation (IBP) to compute an upper bound on the model's loss in the forward pass and minimize this bound via backpropagation. Although IBP gives a theoretical bound, it does not provide any certification condition. Another limitation is that it is not applicable to character-level DNNs, because IBP is limited to continuous space so that model input should be the word-level embedding. SAFER (Ye et al., 2020) achieves certified robustness with a new randomized smoothing technique. However, its computation of synonym set intersection greatly reduces the computation speed in the inference stage. Besides, SAFER only provides a theoretical certified accuracy and its empirical effectiveness on adversarial examples has not been evaluated.

In this paper, we propose a novel approach *WordDP* to certified robustness against word substitution attacks in text classification via differential privacy (DP) (Dwork, 2008). Figure 1 is a high-level illustration. In the inference phase, the input go through a randomized mechanism WordDP. If a clean input satisfies the certification condition of *WordDP*, its adversarial counterpart is guaranteed to predict the same output label. DP is a privacy framework that protects the information of individual record in the database by randomized computations, such that the change of the computation output is bounded when small perturbation is applied on the database. This stable output guarantee is in parallel with the definition of robustness: ensuring that small changes in the input will not result in dramatic shift of its output. The idea of providing robustness certification via DP was originally introduced in PixelDP (Lecuyer et al., 2019) which is specifically designed for norm-bounded adversarial examples in the continuous domain for applications like image classification. However, it is challenging to directly apply such an idea against word substitution attack, due to the discrete nature of the text input space. Therefore, in this work, we develop *WordDP* to achieve the DP and robustness connection in the discrete text space by exploring novel application of the exponential mechanism (McSherry and Talwar, 2007), conventionally utilized to realize DP for answering discrete queries. To achieve this, we present a conceptual certified robustness algorithm that randomly samples word-substituted sentences according to the probability distribution designated by the exponential mechanism and aggregates their inference result as the final classification for the input.

A fundamental barrier limiting the conceptual algorithm from being applied in practice is that the sampling distribution of the exponential mechanism requires an exhaustive enumeration-based sub-step, which needs to repeat the model inference for every neighboring sentences with word substitutions from the input sentence. To overcome this computational difficulty, we develop a practical *simulated exponential mechanism* via uniform sampling and re-weighted averaging, which not only lowers the computational overhead but also ensures uncompromising level of certified robustness.
Our contribution can be summarized as follows:
1) We propose *WordDP* to establish the connection between DP and certified robustness for the first time in text classification domain (Sec.4.1).
2) We leverage conceptual exponential mechanism to achieve *WordDP* and formally prove an L-word bounded certified condition for robustness against word substitution attacks (Sec.4.2).
3) We develop a simulated exponential mechanism via uniform sampling and weighted averaging to overcome the computation bottleneck of the conceptual exponential mechanism without compromising the certified robustness guarantee (Sec.4.3).
4) Extensive experiments validate that *WordDP* outperforms existing defense methods and achieves over 30$\times$ efficiency improvement in the inference stage than the state-of-the-art certified robustness mechanism (Sec.5).

2 Related Work

Word Substitution Attacks. Various attacks have been developed to fool DNNs in text classification, including substituting a word with its synonyms (Ren et al., 2019; Jin et al., 2019; Zang et al., 2020; Alzantot et al., 2018), manipulating the characters (Gao et al., 2018; Ebrahimi et al., 2018), and perturbation on the embedding space (Papernot et al., 2016; Liang et al., 2018; Sato et al., 2018; Cheng et al., 2019).

In word substitution attacks, attackers replace

words in a sentence with their synonyms according to a synonym table, including PWWS (Ren et al., 2019), TEXTFOOLER (Jin et al., 2019), among others (Zang et al., 2020). In particular, PWWS is the most widely used attack algorithm to evaluate defense mechanisms (Zhou et al., 2020; Jia et al., 2019; Ye et al., 2020). PWWS uses WordNet to build synonym set and only replaces named entities (NEs) with similar NEs in order to flip the prediction. It incorporates word saliency to determine the replacement order and selects the synonym that can cause the greatest prediction probability change.

Empirical Defenses to Word Substitution Attacks. Several existing empirical defenses are effective for adversarial word substitution. Miyato *et al*. (2016) applied adversarial training to the text domain. Wang *et al*. (2019) proposed Synonyms Encoding Method (SEM), which finds a mapping between the words and their synonyms before the input layer. Jones *et al*. (2020) proposed robust encodings (RobEn) that involves an encoding function to map sentences to a smaller, discrete space. Dirichlet Neighborhood Ensemble (DNE) (Zhou et al., 2020) creates virtual sentences by mixing the embedding of the original word with its synonyms' embedding via Dirichlet sampling, which is randomized smoothing based data augmentation.

Certified Robustness. Certified robustness has been first studied in image domain, which certifies that a model is robust to adversarial examples when its prediction result is stable when applying small perturbations to the input (Lecuyer et al., 2019; Cohen et al., 2019; Lee et al., 2019). In text domain, Jia *et al*. (2019) and Huang *et al*. (2019) both applied Interval Bound Propagation (IBP) for certification. The intuition is to compute an upper bound on the model's loss through the network in a standard forward pass and minimize this upper bound via backpropagation. One major limitation of IBP certification is that it is not applicable to character-level DNNs, because IBP is limited to continuous space (word-level embedding).

SAFER (Ye et al., 2020) is a certified robust method based on randomized smoothing. The certification is based on the intersection of synonym sets between perturbed examples and clean examples. However, its computation of synonym set intersection greatly reduces the inference efficiency. Besides, it lacks thorough evaluation of empirical effectiveness on adversarial examples.

3 Preliminaries

3.1 Adversarial Word Substitution and Certified Robustness

Adversarial Word Substitution. Consider a sentence of ω words $\mathbf{X} = (x_1, x_2, ..., x_i, ..., x_\omega)$, where each word x_i belongs to a synonym set of $\kappa(i)$ number of synonyms $\mathbf{S}(x_i) = \{x_i^1, x_i^2, ..., x_i^{\kappa(i)}\}$. Following common practice (Ye et al., 2020), we also assume the synonymous relation is symmetric, such that x_i is in the synonym set of all its synonyms $x_i^2, ..., x_i^{\kappa(i)}$ and $\mathbf{S}(x_i^j) = \mathbf{S}(x_i^k)$ for all $j, k \in [\kappa(i)]$. The synonym set $\mathbf{S}(x_i)$ can be built by following GLOVE (Pennington et al., 2014b).

Definition 3.1. **(L-Adversarial Word Substitution Attack)** For an input sentence $\mathbf{X}$, an L-adversarial word substitution attack perturbs the sentence by selecting at most L ($L \leq \omega$) words $x_{\tau_1}, ..., x_{\tau_L}$ and substitutes each selected word x_{τ_i} with one of its synonyms $x'_{\tau_i} \in \mathbf{S}(x_{\tau_i})$. We denote an attacked sentence by $\mathbf{X}'$ and the set of all possible attacked sentences by $\mathscr{S}(L)$.

Certified Robustness. In general, we say a model is robust to adversarial examples when its prediction result is stable when applying small perturbations to the input.

Definition 3.2. **(Certified Robustness to Word Substitution Attack)** Denote a multiclass classification model by $f(\mathbf{X}) : \mathcal{X} \mapsto c \in \mathcal{C}$, where c is a label in the possible label set $\mathcal{C} = \{1, ..., C\}$. In general, $f(\mathbf{X})$ outputs a vector of scores $f^{\mathbf{y}}(\mathbf{X}) = (f^{y_1}, ..., f^{y_C}) \in \mathcal{Y}$, where $\mathcal{Y} = \{\mathbf{y} : \sum_{i=1}^{C} f^{y_i} = 1, f^{y_i} \in [0, 1]\}$, and $c = \arg\max_{i \in \mathcal{C}} f^{y_i}$. A predictive model $f(\mathbf{X})$ is robust to L-adversarial word substitution attack on input $\mathbf{X}$, if for all $\mathbf{X}' \in \mathscr{S}(L)$, it has $f(\mathbf{X}) = f(\mathbf{X}')$, which is equivalent to

$$y_c(\mathbf{X}') > \max_{i \in \mathcal{C}: i \neq c} y_i(\mathbf{X}'). \quad (1)$$

In the following, we refer to the above robustness as L-certified robustness for short.

3.2 Differential Privacy and Exponential Mechanism

Differential Privacy. The concept of DP is to prevent the information leakage of an individual record in the database by introducing randomness into the computation. More specifically, DP guarantees the output of a function over two neighbouring databases are indistinguishable.

Definition 3.3. (Differential Privacy (Dwork et al., 2006)) A randomized mechanism $\mathcal{A}$ is ϵ-differentially private if, for all neighboring datasets $\mathbf{D} \sim \mathbf{D}'$ that differ in one record or are bounded by certain distance and for all events $\mathbf{O}$ in the output space $\mathcal{O}$ of $\mathcal{A}$, we have

$$\mathbb{P}[\mathcal{A}(\mathbf{D}) \in \mathbf{O}] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(\mathbf{D}') \in \mathbf{O}]. \quad (2)$$

Exponential Mechanism. The exponential mechanism is a commonly utilized DP mechanism in the discrete domain, which consists of the utility score function, sensitivity, and sampling probability distribution as its key ingredients.

Definition 3.4. (Exponential Mechanism (McSherry and Talwar, 2007)) Denote the score function $u(\mathbf{D}, \mathbf{r}) : \mathcal{D} \times \mathcal{R} \mapsto \mathbb{R}$, which maps each pair of input dataset $\mathbf{D} \sim \mathcal{D}$ and candidate result $\mathbf{r} \in \mathcal{R}$ to a real valued score. Denote the sensitivity by $\Delta_u := \max_{\mathbf{r}\in\mathcal{R}} \max_{\mathbf{D}\sim\mathbf{D}'} |u(\mathbf{D}, \mathbf{r}) - u(\mathbf{D}', \mathbf{r})|$. The exponential mechanism $\mathcal{M}_E(\mathbf{D}, u, \mathcal{R})$ selects and outputs an element $\mathbf{r} \in \mathcal{R}$ with probability proportional to $e^{\frac{\epsilon u(\mathbf{D},\mathbf{r})}{2\Delta_u}}$. The exponential mechanism is ϵ-differentially private.

4 Proposed Method

4.1 *WordDP* for Certified Robustness

***WordDP*.** We expand the intuition that DP can be applied to provide certified robustness against textual adversarial examples like word substitution attack by regarding the sentence as a database and each word as a record. If the randomized predictive model satisfies ϵ-DP during inference, then the output of a potentially adversarial input $\mathbf{X}' \in \mathscr{S}(L)$ and the output of the original input $\mathbf{X}$ should be indistinguishable. Thus, our proposed approach is to transform a multiclass classification model's prediction score into a randomized ϵ-*WordDP* score, which is formally defined below.

Definition 4.1. (Word Differential Privacy) Consider any input sentence $\mathbf{X}$ and its L-word substitution sentence set $\mathscr{S}(L)$. For a randomized function $f_{\mathcal{A}}(\mathbf{X})$, let its prediction score vector be $\mathbf{y} \in \mathcal{Y}$. $f_{\mathcal{A}}(\mathbf{X})$ satisfies ϵ-word differential privacy (*WordDP*), if it satisfies ϵ-differential privacy for any pair of neighboring sentences $\mathbf{X}_1, \mathbf{X}_2 \in \mathscr{S}(L)$ and the output space $\mathbf{y} \in \mathcal{Y}$.

Remark 1. We stress that *WordDP* does not seek DP protection for the training dataset as in the conventional privacy area. Instead, it leverages the DP randomness for certified robustness during inference with respect to a testing input.

In practice, for a base model f, a DP mechanism $\mathcal{A}$ will be introduced to randomize it to $f_{\mathcal{A}}$. For an ϵ-*WordDP* model $f_{\mathcal{A}}$, its expected prediction $\mathbb{E}[f_{\mathcal{A}}(\mathbf{X})]$ is certified robust. Denote the prediction score vector of $\mathbb{E}[f_{\mathcal{A}}(\mathbf{X})]$ by $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{A}}(\mathbf{X})] = (\mathbb{E}[f^{y_1}_{\mathcal{A}}(\mathbf{X})], ..., \mathbb{E}[f^{y_C}_{\mathcal{A}}(\mathbf{X})]) \in \mathcal{Y}$. Lemma 4.2 shows $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{A}}(\mathbf{X})]$ satisfies the certified robustness condition in eq.(1), based on Lemma 4.1 that shows each expected prediction score $\mathbb{E}[f^{y_i}_{\mathcal{A}}(\mathbf{X})]$ is stable.

Lemma 4.1. *For an ϵ-WordDP model $f_{\mathcal{A}}$, its prediction score satisfies the relation, $\forall i \in [C]$,*

$$\mathbb{E}[f^{y_i}_{\mathcal{A}}(\mathbf{X}_1)] \leq e^{\epsilon}\mathbb{E}[f^{y_i}_{\mathcal{A}}(\mathbf{X}_2)], \forall \mathbf{X}_1, \mathbf{X}_2 \in \mathscr{L}. \quad (3)$$

From the above property, we can derive the certified robustness condition to adversarial examples.

Lemma 4.2. *For an ϵ-WordDP model $f_{\mathcal{A}}$ and an input sentence $\mathbf{X}$, if there exists a label c such that:*

$$\mathbb{E}(f^{y_c}_{\mathcal{A}}(\mathbf{X})) > e^{2\epsilon} \max_{i\neq c} \mathbb{E}(f^{y_i}_{\mathcal{A}}(\mathbf{X})), \quad (4)$$

then the multiclass classification model $f_{\mathcal{A}}$ based on the expected label prediction score vector $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{A}}(\cdot)]$ is certified robust to L-adversary word substitution attack on $\mathbf{X}$.

The proofs of the above two lemmas can be adapted from the pixelDP to *WordDP* context based on *Lemma 1* and *Proposition 1* in Lecuyer *et al.* (2019). We relegate the proofs to Appendix A. Our focus is how to design the DP mechanism $\mathcal{A}$ to achieve *WordDP* (Subsection 4.2), and how to implement it for efficient inference that still ensures certified robustness (Subsection 4.3).

4.2 *WordDP* with Exponential Mechanism

In this subsection, we present the conceptual exponential mechanism-based algorithm to achieve *WordDP* and the certification procedure.

Exponential Mechanism for *WordDP*. To obtain the DP classifier $f_{\mathcal{A}}$ given the base model f, we introduce the exponential mechanism $\mathcal{M}_E$ as the randomization mechanism $\mathcal{A}$ and define $f_{\mathcal{A}} := f(\mathcal{M}_E)$. Given an input example, the mechanism selects and outputs L-substitution sentences with a probability based on exponential mechanism. It then aggregates the inferences of these samples by an average as the estimated prediction of the input. Figure 2 illustrates the algorithm.

Definition 4.2. (Exponential Mechanism for *WordDP* and L-Certified Robustness) Given the base model f, for any input sentence $\mathbf{X}$ and potential L-substitution sentence set $\mathscr{S}(L)$, we define the utility score function as:

$$u(\mathscr{S}(L), \mathbf{X}') = e^{-\|f^{\mathbf{y}}(\mathbf{X}') - f^{\mathbf{y}}(\mathbf{X})\|_1}, \quad (5)$$

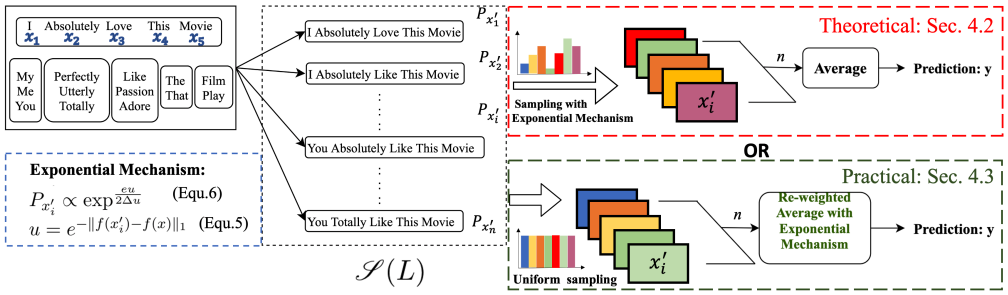


Figure 2: *WordDP* with Exponential Mechanism.

which associates a utility score to a candidate output $\mathbf{X}' \in \mathscr{S}(L)$. The sensitivity of the utility score is $\Delta_u = 1 - e^{-1}$. Then, the exponential mechanism selects and outputs $\mathbf{X}'$ with probability $\mathbb{P}_{\mathbf{X}'}$

$$\mathbb{P}_{\mathbf{X}'} = \frac{1}{\rho} \exp\left(\frac{\epsilon \cdot u(\mathscr{S}(L), \mathbf{X}')}{2\Delta u}\right), \tag{6}$$

where $\rho = \sum_{i=1}^{|\mathscr{S}(\mathbf{X},L)|} \exp\left(\frac{\epsilon \cdot u(\mathscr{S}(L), \mathbf{X}'_i)}{2\Delta u}\right)$ is the normalization factor.

Proposition 4.1. *The exponential mechanism $\mathcal{M}(E)$ satisfies ϵ-DP. The composition model function $f_{\mathcal{M}_E}(\mathbf{X}) := f(\mathcal{M}_E(X))$ is ϵ-DP and its prediction score vector $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$-based classification is certified robust to L-adversary word substitution attack on $\mathbf{X}$.*

Proof. To show $\mathcal{M}_E$ is ϵ-DP, we prove the sensitivity of the utility score (maximum difference between the utility scores given any two neighboring input) Δ_u is indeed $1 - e^{-1}$ and the remaining follows the definition of the exponential mechanism (c.f.Definition 3.4). Since $\|f^{\mathbf{y}}(\mathbf{X}'_i) - f^{\mathbf{y}}(\mathbf{X})\|_1$ is the prediction probability change which is in $[0, 1]$, we have $u(\mathscr{S}(L), \mathbf{X}'_i) \in [e^{-1}, 1]$, which leads to $\Delta_u = 1 - e^{-1}$. Next, since $\mathcal{M}_E(\mathbf{X})$ is ϵ-DP, by the post-processing property (i.e., any computation on the output of the DP mechanism remains DP, Proposition 2.1 in (Dwork et al., 2014).), $f_{\mathcal{M}_E}(\mathbf{X})$ is also ϵ-DP. Subsequently, by Lemma 4.2, $\mathbb{E}[f_{\mathcal{M}_E}(\mathbf{X})]$ is L-certified robust on $\mathbf{X}$. □

Remark 2. 1) The design of the utility function has the intuition that we wish to assign higher probability to sentences that have minimal impact on the prediction score function. 2) The privacy budget ϵ influences whether the sampling probability distribution is flat (lower ϵ) or peaky (greater ϵ). Too small of an ϵ value will clearly affect the prediction accuracy. For certification purpose, according to the certified condition Lemma 4.2, too large of an ϵ value will result in none certified, so ϵ can only be searched within a limited range.

Certification Condition. It is a common practice in certified robustness literature to estimate $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ via Monte Carlo estimation (Lecuyer et al., 2019; Cohen et al., 2019) in the form of $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$. That is, we repeat the exponential mechanism-based inference to draw n samples of $f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X}'_\tau)$, for $\tau \in [n]$ and let $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] = \frac{1}{n}\sum_{\tau=1}^{n} f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X}'_\tau)$. The estimation error between $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ and $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ can be bounded based on Hoeffding's inequality with probability η, which guarantees that $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] \in [\mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] - \sqrt{\frac{1}{2n} ln(\frac{2C}{1-\eta})}, \mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] + \sqrt{\frac{1}{2n} ln(\frac{2C}{1-\eta})}] := [\widehat{\mathbb{E}}^{lb}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})], \widehat{\mathbb{E}}^{ub}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]]$. The next proposition shows that the inference based on the estimated $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ (as versus $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$) can still ensure certified robustness.

Proposition 4.2. *Under the same condition with Proposition 4.1, if there exists a label c such that*

$$\widehat{\mathbb{E}}^{lb}[f^{\mathbf{y}_c}_{\mathcal{M}_E}(\mathbf{X})] > e^{2\epsilon} \max_{i \neq c} \widehat{\mathbb{E}}^{ub}[f^{\mathbf{y}_i}_{\mathcal{M}_E}(\mathbf{X})], \tag{7}$$

the prediction score vector $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$-based classification is certified robust with probability η to L-adversary word substitution attack on $\mathbf{X}$.

4.3 Simulated Exponential Mechanism

Simulated Exponential Mechanism. The conceptual exponential mechanism in Definition 4.2 is computationally impractical. The bottleneck is the need to enumerate the entire $\mathscr{S}(L)$ in order to calculate the probability distribution of $\mathbb{P}_{\mathbf{X}'}$ for each $\mathbf{X}' \in \mathscr{S}(L)$ and the normalization factor ρ, which essentially requires us to perform inference for $\mathscr{S}(L) \gg n$ times (n is the number of samples) for certifying a single input sentence $\mathbf{X}$.

In the following, we show that we can significantly reduce the computation cost by sampling via a simulated exponential mechanism, which suffices to sample n candidate $L-$substitution sentences

and calculate only n times, i.e., the same repetitions as the Monte Carlo estimation. The key insight is based on the different purpose of applying the exponential mechanism between the conventional scenario for achieving DP and our certified robustness scenario. For the former, in order to ensure DP of the final output $f_{\mathcal{M}_E}(\mathbf{X}'_\tau)$, the intermediate $\mathbf{X}'_\tau$ is forced to satisfy DP, i.e., drawn from the exact probability distribution designated by the exponential mechanism. For the latter, while the derivation of the certified robustness relied on the randomness of DP and the exponential mechanism, we do not actually require the DP of the intermediate $\mathbf{X}'_\tau$. As a result, it allows us to sample $\mathbf{X}'_\tau$ from other simpler distributions without calculating the probability distribution of the exponential mechanism, as long as the alternative approach can obtain the equivalent $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{A}}(\mathbf{X})]$ for robustness certification.

We develop a simulated exponential mechanism via *uniform sampling and re-weighted average prediction score calculation*. Figure 2 shows the simulated mechanism in contrast to the conceptual mechanism. In detail, we sample from $\mathscr{S}(L)$ with uniform probability, which can be efficiently implemented without generating $\mathscr{S}(L)$. Denoting a sample by $\mathbf{X}'_\tau$, we calculate its scaled exponential mechanism probability by

$$\bar{\mathbb{P}}_{\mathbf{X}'_\tau} = \exp(\frac{\epsilon \cdot u(\mathscr{S}(L), \mathbf{X}'_\tau)}{2\Delta u}), \tag{8}$$

which can be obtained via a single inference on $\mathbf{X}'_\tau$ and the inference on $\mathbf{X}$ due to the omission of the normalization factor ρ that requires the entire $\mathscr{S}(L)$. The inference on $\mathbf{X}$ only needs to be computed once and shared by all n Monte Carlo repetitions. Such uniform sampling and scaled probability calculation is repeated for n times, which requires only $n+1$ inferences. Finally, we use the following re-weighted average prediction score (weighted by the scaled exponential mechanism probability) for certified robust prediction,

$$\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] = \sum_{\tau=1}^{n} \bar{\mathbb{P}}_{\mathbf{X}'_\tau} \cdot f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X}'_\tau). \tag{9}$$

The following theorem shows that $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$-based prediction guarantees certified robustnessand the conceptual exponential mechanism-based inference in Proposition 4.2 is certified robust provided $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ is so.

Theorem 4.1. *For any input $\mathbf{X}$, let $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ be calculated by eq.(9). Denote $\bar{\mathbb{E}}^{lb}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ and $\bar{\mathbb{E}}^{ub}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ be η-confidence lower and upper bounds, respectively, i.e., $\bar{\mathbb{E}}^{lb}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] = \bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] - \sqrt{\frac{1}{2n} ln(\frac{2C}{1-\eta})}$ and $\bar{\mathbb{E}}^{ub}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] = \bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})] + \sqrt{\frac{1}{2n} ln(\frac{2C}{1-\eta})}$. If there exists a label c such that*

$$\bar{\mathbb{E}}^{lb}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X})] > e^{2\epsilon} \max_{i \neq c} \bar{\mathbb{E}}^{ub}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})], \tag{10}$$

the prediction score vector $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$-based classification is certified robust with probability η to L-adversary word substitution attack on $\mathbf{X}$.

The proof of Theorem 4.1 requires the following lemma, which is adapted from Lemma 4.1 from the accurate expectation of $\mathbb{E}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ to the simulated expectation $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$. We stress that during both proofs, we do not use the DP property of $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\cdot)]$, but only its equivalent relation to $\widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\cdot)]$.

Lemma 4.3. *For any label $i \in [C]$ and any $\mathbf{X}_1, \mathbf{X}_2 \in \mathscr{S}(L)$, let $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{M}_E}(\mathbf{X})]$ be computed by eq.(9). Then, we have*

$$\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}_1)] \leq e^{\epsilon} \bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}_2)]. \tag{11}$$

Proof. First, we notice that for any $X' \in \mathscr{S}(L)$, it has $\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')] = \frac{\rho}{|\mathscr{S}(L)|} \widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')]$ by $\bar{\mathbb{P}}[\mathbf{X}'] = \rho \mathbb{P}[\mathbf{X}']$ and the uniform sampling probability $\frac{1}{|\mathscr{S}(L)|}$. Second, since $\widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')]$ is ϵ-*WordDP* , we can show that it satisfies Lemma 4.1 by switching $\mathbb{E}[f^{y_i}_{\mathcal{M}_E}(\cdot)]$ there to $\widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\cdot)]$ here. It follows that:

$$\begin{aligned}
&\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}_1)] = \widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}_1)] \cdot (\frac{\rho}{|\mathscr{S}(L)|}) \\
&\leq e^{\epsilon} \widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}_2)] \cdot (\frac{\rho}{|\mathscr{S}(L)|}) = e^{\epsilon} \bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}_2)],
\end{aligned}$$

which proves the lemma. □

Proof. **(Proof of Theorem 4.1)** For any $\mathbf{X}' \in \mathscr{S}(L)$, by eq.(11), we have

$$\begin{aligned}
&e^{\epsilon} \bar{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X}')] \geq \bar{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X})] \\
&> \bar{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X})] - \sqrt{\frac{1}{2n} ln(\frac{2C}{1-\eta})} = \bar{\mathbb{E}}^{lb}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X})];
\end{aligned}$$

as well as

$$\begin{aligned}
&\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')] \leq e^{\epsilon} \bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})] \leq e^{\epsilon} \max_{i \neq c} \bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})] \\
&\leq e^{\epsilon} \max_{i \neq c} (\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})] + \sqrt{\frac{1}{2n} ln(\frac{2C}{1-\eta})}) \\
&= e^{\epsilon} \max_{i \neq c} \bar{\mathbb{E}}^{ub}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})].
\end{aligned}$$

Equipped with the above two relations, we can prove the claim in Theorem 4.1. We show that $\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})]$ is certified robust for any $\mathbf{X}' \in \mathscr{S}(L)$, as follows,

$$\begin{aligned}
&\bar{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X}')] > \bar{\mathbb{E}}^{lb}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X})]/e^{\epsilon} \\
&> e^{\epsilon} \max_{i \neq c} \bar{\mathbb{E}}^{ub}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})] > e^{\epsilon} \max_{i \neq c} \bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')].
\end{aligned} \tag{12}$$

which is $\bar{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X}')] > e^{2\epsilon}\max_{i\neq c}\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})]$. For completeness, we can also show that the certified robustness of $\bar{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{A}}(\mathbf{X})]$ implies the certified robustness of $\widehat{\mathbb{E}}[f^{\mathbf{y}}_{\mathcal{A}}(\mathbf{X})]$:

$$\begin{aligned}
&\widehat{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X}')] = (\frac{|\mathscr{S}(L)|}{\rho})\cdot\bar{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X}')]\\
&> (\frac{|\mathscr{S}(L)|}{\rho})\bar{\mathbb{E}}^{lb}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X})]/e^{\epsilon}\\
&> (\frac{|\mathscr{S}(L)|}{\rho})e^{\epsilon}\max_{i\neq c}\bar{\mathbb{E}}^{ub}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X})]\\
&> (\frac{|\mathscr{S}(L)|}{\rho})\max_{i\neq c}\bar{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')] = \max_{i\neq c}\widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')],
\end{aligned}$$

which proves $\widehat{\mathbb{E}}[f^{y_c}_{\mathcal{M}_E}(\mathbf{X}')] > \max_{i\neq c}\widehat{\mathbb{E}}[f^{y_i}_{\mathcal{M}_E}(\mathbf{X}')]$.□

Training procedure. To achieve a better certification result, we involve randomness in the training stage, which is also adopted by almost all certified robustness approaches. To do so, we use the data augmentation strategy that utilizes the perturbed sentences for training, i.e., $\mathbf{X}' \in \mathscr{S}(L) \setminus \mathbf{X}$ given the original training sample $\mathbf{X}$. In practice, we first train the model without data augmentation for several epochs to achieve a reasonable performance, followed by training with perturbed $\mathbf{X}'$. For each training data point, we randomly draw one neighbour sentence during training (as opposed to multiple draws during certified inference).

5 Experiments

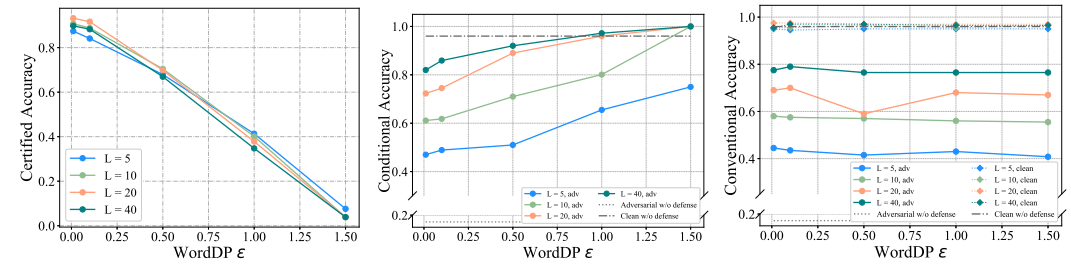
We evaluate *WordDP* on two classification datasets: Internet Movie Database (IMDB) (Maas et al., 2011) and AG News corpus (AGNews) (Zhang et al., 2015). IMDB is a binary sentiment classification dataset containing 50000 movie reviews. AGNews includes 30,000 news articles categorized into four classes. The target model architecture we select is a single-layer LSTM model with size of 128. We use Global Vectors for Word Representation (GloVe) (Pennington et al., 2014a) for word embedding. The LSTM model achieves 88.4% and 91.8% clean accuracy on IMDB and AGNews, respectively. We use PWWS (Ren et al., 2019) to generate adversarial examples on the test dataset. PWWS is a state-of-the-art attack method which uses WordNet to build synonym set and incorporates word saliency to replace selected named entities (NEs) with their synonyms in order to flip the prediction. The details about the datasets, model training and attack algorithm are in Appendix C.

5.1 Evaluation Metrics and Baselines

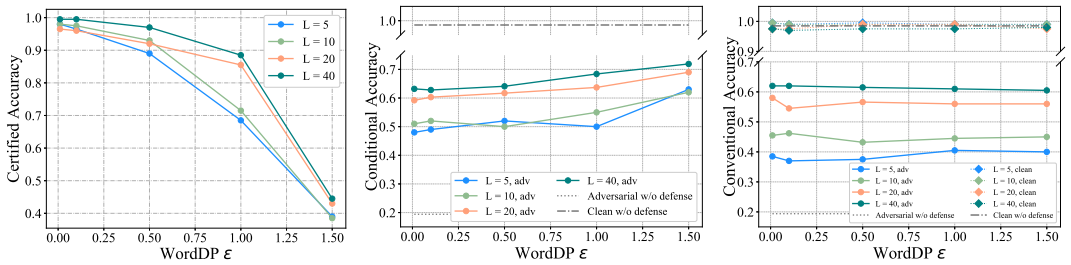
We use four metrics to evaluate the effectiveness of *WordDP*: certified ratio, certified accuracy, conditional accuracy, and conventional accuracy. **Certified Ratio** represents the fraction of testing set that the prediction satisfies the certification criteria: $\frac{\sum_{t=1}^{T} certifiedCheck(\mathbf{X}_t,L,\epsilon)}{T}$, where $certifiedCheck$ returns 1 if Theorem 4.1 is satisfied and T is the size of the test dataset. **Certified accuracy (CertAcc)** denotes the fraction of the clean testing set on which the predictions are both correct and satisfy the certification criteria. This is a standard metric to evaluate certified robust model (Lecuyer et al., 2019). Formally, it is defined as: $\frac{\sum_{t=1}^{T} certifiedCheck(\mathbf{X}_t,L,\epsilon)\&corrClass(\mathbf{X}_t,L,\epsilon)}{T}$, where $corrClass$ returns 1 if the classification output is correct. When the accuracy of a model is close to 100%, certified accuracy largely reflects certified ratio. **Conventional accuracy (ConvAcc)** is defined as the fraction of testing set that is correctly classified, $\frac{\sum_{t=1}^{T} corrClass(\mathbf{X}_t,L,\epsilon)}{T}$, which is a standard metric to evaluate any deep learning systems. Note that the input $\mathbf{X}_t$ can be both adversarial or clean inputs. We use this metric to evaluate how *WordDP* empirically works on adversarial examples.

Besides the above standard metrics, we introduce a new accuracy metric called **Conditional accuracy (CondAcc)** to evaluate the following: when a clean input $\mathbf{X}_t$ is certified within bound L, whether its corresponding L-word substitution adversarial example $\mathbf{X}^{adv}_t$ is indeed correctly classified. The CondAcc can be formulated as: $\frac{\sum_{t=1}^{T} certifiedCheck(\mathbf{X}_t,L,\epsilon)\&corrClass(\mathbf{X}^{adv}_t,L,\epsilon)}{\sum_{t=1}^{T} certifiedCheck(X_t,L,\epsilon)}$. While certified accuracy is typically evaluated on clean inputs in the literature to show the certified robustness property, conditional accuracy is evaluated on adversarial inputs and provides an informative measure of the classification result of adversarial examples when its counterpart clean input can be certified. This metric is aligned with the definition and purpose of certified robustness. Ideally, if a clean example is successfully certified, adversarial examples created from this clean example should have the same prediction. Therefore, the accuracy of adversarial examples is influenced by the ConvAcc of clean examples.

Comparison Methods. We compare *WordDP* with the state-of-the-art certified robust method SAFER for text classification. We note that SAFER only reports certified accuracy, without accuracy on adversarial examples. To conduct a fair comparison with *WordDP*, we rerun SAFER on the adversarial examples and report the comparison



(a) CertAcc on IMDB (b) CondAcc on IMDB (c) ConvAcc on IMDB



(d) CertAcc on AGNews (e) CondAcc on AGNews (f) ConvAcc on AGNews

Figure 3: Certified Accuracy, Conditional Accuracy and Conventional Accuracy on IMDB and AGNews

in CertAcc and CondAcc. Besides SAFER, we also compare the ConvAcc on adversarial examples with two state-of-the-art defense methods, i.e., IBP (Jia et al., 2019) and DNE (Zhou et al., 2020), which do not provide certified robustness guarantee. Thus, their defense may be broken by more powerful word substitution attacks in the future.

5.2 Certified Results

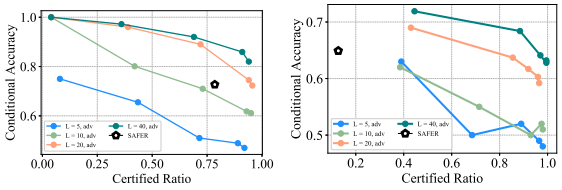
Certified Accuracy. Figure 3 presents the CertAcc, CondAcc and ConvAcc under different ϵ and L, respectively. Each line in the figures represents a certified bound L, which allows L number of words to be substituted. The first row is the results on IMDB, and the second row is on AGNews.

Figures 3(a) and 3(d) show the certified accuracy on the two datasets. Since the conventional accuracy on the clean examples of our mechanisms is close to 100% (as shown in Figures 3(c) and 3(f)), the certified accuracy mainly reflects the certified ratio (which we skip in the results). As shown, higher ϵ can result in lower CertAcc. This is intuitive as the condition in Theorem 4.1 is more difficult to satisfy when given higher epsilon, i.e. weaker requirement of indistinguishability of the output, hence results in lower certified ratio. As illustrated in 3(a), when ϵ is around 1.5, the mechanism will approach 0 certified ratio. This indicates that ϵ can only be searched within a limited range.

Comparing each line in 3(a) and 3(d), we note that greater L results in higher CertAcc in most cases for the AGNews dataset. This can be explained by the fact that a greater L means more word substitutions and randomness are introduced in *WordDP*, making it easier to ensure the indistinguishability of the output, and hence a higher certified ratio.

	ADV	IBP	DNE	SAFER	*WordDP*
IMDB	0.172	0.722	0.823	0.727	**0.972**
AGNews	0.194	0.823	**0.909**	0.647	0.719

Table 1: Empirical comparison on accuracy

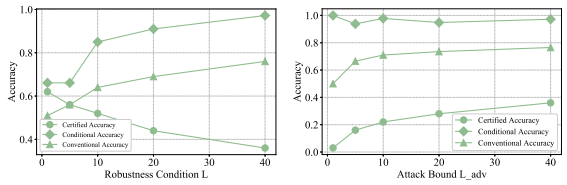


(a) IMDB (b) AGNews

Figure 4: Certified Ration vs. Conditional Accuracy

Accuracy on Adversarial Examples. Figures 3(b), 3(e), 3(c) and 3(f) present CondAcc and ConvAcc of the two datasets on adversarial examples, respectively. Note that we only test the adversarial examples that are within the L bound. We also show the CondAcc and ConvAcc for both clean and adversarial examples without any defense mechanisms as a reference. In addition, we show ConvAcc of *WordDP* with varying parameters on clean examples to show the impact of the mechanism on clean examples.

As shown in the figures, *WordDP* achieves significantly higher accuracy on adversarial examples compared to no defense while maintaining the close to 100% accuracy on clean examples. Conditional



(a) Fixed attack power 40 (b) Fixed defense power 40

Figure 5: The trend on accuracy under different defense and attack power

accuracy is higher than conventional accuracy as expected, since it is computed only on those adversarial examples with a certified counterpart clean example. Besides, we can observe that with higher ϵ, higher CondAcc on adversarial examples can be achieved. This is because less randomness is introduced in the inference.

In addition, by comparing different L bound under the same ϵ, larger L can yield more accuracy improvement on adversarial examples but less on clean examples. Intuitively, using the aggregated prediction of more distant neighbouring sentences (higher L) can benefit adversarial examples more than clean examples.

Trader-off between Certified Ratio and CondAcc. We can see that ϵ has an opposite impact on certified accuracy (certified ratio) and CondAcc, we present the trade-off between the certified ratio and CondAcc of *WordDP* in Figure 4 in comparison with the baseline method SAFER. Ideally, we want both high certified ratio and high condAcc to contribute to overall high accuracy. The black dot represents the baseline SAFER, since the neighbouring sentence generating method of SAFER does not depend on L or ϵ. As illustrated on these two datasets, with $L = 20$ and $L = 40$,*WordDP* can dominate SAFER and achieve a much better performance in both certified ratio and condAcc.

Relation between certified bound L and adversarial attack power L_{adv}. Figure 5 presents the three accuracy metrics under different attack power and defense power. In Figure 5(a), we fix the attack power L_{adv} to 40, which means allowing less than 40 word substitutions, and adjust the *WordDP* defense power by using different certified bound L. As discussed in Section 4, certified bound L determines the size of neighbouring set. Greater L leads to higher randomness and thus can benefit the CondAcc and ConvAcc on adversarial examples. On the other hand, greater L also makes the certified condition more difficult to be satisfied, which result in lower CertAcc.

In Figure 5(b), we fix the certified bound L to 40, which means using the same power of *WordDP* to defend against adversarial examples generated by varying attack power L_{adv}. As shown in the figure, the performance increases with higher attack power. This is because the adversarial examples with more word changes (higher L_{adv}) are more difficult to generate but easier to defend (due to the nature of PWWS attack algorithm).

Comparison with Empirical Defense. Besides certified robust method SAFER, we also compare CondAcc of *WordDP* with baseline empirical defense methods, IBP (Jia et al., 2019) and DNE (Zhou et al., 2020). Table 1 compares the highest CondAcc achieved by *WordDP* with the conventional accuracy reported by the baselines (ADV corresponds to no defense). *WordDP* achieves a much higher accuracy on IMDB dataset compared to IBP, DNE and SAFER. For AGNews, the accuracy of *WordDP* outperforms SAFER, but is lower than the two empirical defenses. We stress, however, the empirical defense methods do not provide any rigorous certified robustness guarantees and the performance can be significantly dependent on datasets and specific attacks.

Efficiency Comparison. We also compare the efficiency of *WordDP* with SAFER by computing the average time cost for certifying one input and producing the Monte Carlo sampling-based output. It takes *WordDP* 6.25s and 3.21s on IMDB and AGNews, respectively. As a comparison, it costs SAFER 230.35s and 96.68s. Thus, *WordDP* achieves more than $30\times$ efficiency improvement.

6 Conclusion

We proposed *WordDP*, a certified robustness method to adversarial word substitution attacks with the exponential mechanism-based algorithm. Compared with previous work, *WordDP* achieves notable accuracy improvement and $30\times$ efficiency improvement. In the future, it would be interesting to expand *WordDP* to other kinds of textual adversarial examples, such as character-level attacks. It is also worthwhile to study other certified approaches such as random smoothing.

Acknowledgement

We sincerely thank all anonymous reviewers for their constructive comments. This work is partially supported by the National Science Foundation (NSF) CNS-1952192, IIS-1838200, and National Institutes of Health (NIH) CTSA Award UL1TR002378.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. 2016. Opinion mining and sentiment analysis. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 452–455. IEEE.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. *ACL*.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. *ACL*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. *EMNLP-IJCNLP*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. *EMNLP-IJCNLP*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*, 2.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. *ACL*.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. 2019. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 4910–4921.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. *IJCAI*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *stat*, 1050:7.

Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM 2016-2016 IEEE Military Communications Conference*, pages 49–54. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014b. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Interpretable adversarial perturbation in input embedding space for text. *IJCAI*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Xiaosen Wang, Hao Jin, and Kun He. 2019. Natural language adversarial attacks and defenses in word level. *arXiv preprint arXiv:1909.06723*.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. *ACL*.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-wei Chang, and Xuanjing Huang. 2020. Defense against adversarial attacks in nlp via dirichlet neighborhood ensemble. *arXiv preprint arXiv:2006.11627*.

Appendix

This is the appendix for the submission: *Certified Robustness to Word Substitution Attack with Differential Privacy*. Section A contains additional proofs that are omitted in the paper; Section B presents additional experiment results; Section C provides additional details of the experiment.

A Proof

A.1 Proof for Lemma 4.1

Proof. Take y_1 as an example and $y_1 \in [0, 1]$.

$$\begin{aligned}\mathbb{E}[f_{\mathcal{A}}^{y_1}(\mathbf{X})] &= \int_0^1 \mathbb{P}(f_{\mathcal{A}}^{y_1}(\mathbf{X}) > t)dt \\ &\overset{(a)}{\leq} e^{\epsilon}(\int_0^1 f_{\mathcal{A}}^{y_1}(\mathbf{X}') > t)dt) \\ &= e^{\epsilon}\mathbb{E}[f_{\mathcal{A}}^{y_1}(\mathbf{X}')],\end{aligned} \tag{13}$$

where (a) is the by definition of DP. The same proof holds for any $y \in \mathcal{Y}$. □

A.2 Proof for Lemma 4.2

Proof. By eq.(3) in the paper, $\forall \mathbf{X}' \in \mathscr{S}(L)$ we have:

$$\mathbb{E}[f_{\mathcal{A}}^{y_n}(\mathbf{X})] \leq e^{\epsilon}\mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X}')] \tag{14}$$

$$\mathbb{E}[f_{\mathcal{A}}^{y_i}(\mathbf{X}')] \leq e^{\epsilon}\mathbb{E}[f_{\mathcal{A}}^{y_i}(\mathbf{X})], \quad i \neq n \tag{15}$$

Then we have:

$$\begin{aligned}\mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X}')] &\overset{Eq(16)}{\geq} \frac{\mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X})]}{e^{\epsilon}} \\ &\overset{eq(4)}{\geq} \frac{e^{2\epsilon}\max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X}))}{e^{\epsilon}} \\ &= e^{\epsilon}\max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X})) \\ &\overset{eq(17)}{\geq} \max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X}'))\end{aligned} \tag{16}$$

$$\forall \mathbf{X}' \in \mathscr{S}(\mathbf{X}, L), \mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X}')] \geq \max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X}')). \tag{17}$$

the definition of robustness at $\mathbf{X}$ holds. □

B Addition Experimental Results

B.1 Parameter Impact: Sampling Rate

As discussed in Section 4.3, we use Monte Carlo sampling to estimate the expected value of the randomized scoring function $\widehat{\mathbb{E}}[\mathcal{A}(\mathbf{X}')]$. The draw times n can influence the final estimation of the expected value. Here we present the result of tuning n on IMDB dataset. As shown in Table 2, we evaluate different n on two settings, $L = 3, \epsilon = 0.4$ and $L = 9, \epsilon = 1$. CP and CA represent certified percentage and certified accuracy respectively. With the increase of draw times n, while the certified percentage does not change significantly, the certified accuracy increases. The best certified accuracy is achieved when $n = 1000$.

C Experiment Details

C.1 Datasets

The detailed comparison between IMDB and AGNews are shown in Table 3.

	L = 9, eps = 1		l = 3, eps = 1.2	
N	CP	CA	CP	CA
10	0.551	0.836	0.511	0.761
50	0.55	0.841	0.45	0.777
100	0.547	0.845	0.484	0.781
500	0.553	0.846	0.468	0.782
1000	0.549	**0.857**	0.479	**0.814**

Table 2: Certified percentage and certified accuracy under different n

Dataset	IMDB	AGNews
Training size	20,000	120,000
Testing size	2000	2000
Task	binary	four-class
Vocab size	116,839	114,096
Average synonym size	3.52	3.79
Sentence length	269.97±200.88	44.97±12.55
embedding_dims	100	100

Table 3: Summary of datasets

C.2 Target Model

The architecture we use for both datasets are single-layer LSTM model with hidden size of 128. The batch size for IMDB is 32 and for AGNews is 63. The word embedding dimension is 100 for both of the datasets. The loss functions are binary crossentropy loss for IMDB and categorical crossentropy for AGNews. For both datasets. Both dataset are trained with 30 epoches. The optimizer is Adam with learning rate 1×10^{-2}. Dropout rate is 0.3.

C.3 Attack Algorithm

The attack algorithm we use to generate adversarial examples is PWWS(Ren et al., 2019), which calculates the word replacement order based on both the word saliency and the classification probability, and uses WordNet to build synonym set and replace named entities (NEs) with similar NEs to flip the prediction. In the experiments, we randomly generate 2000 adversarial examples.