Random Smooth-based Certified Defense against Text Adversarial Attack

Zeliang Zhang[1]*, **Wei Yao**[2]*, **Susan Liang**[1], **Chenliang Xu**[1]
[1] School of Computer Science, University of Rochester
[2] Gaoling School of Artificial Intelligence, Renmin University of China
{hust0426, busyweiyao}@gmail.com, {susan.liang,chenliang.xu}@rochester.edu,

Abstract

Certified defense methods have identified their effectiveness against textual adversarial examples, which train models on the worst-case text generated by substituting words in original texts with synonyms. However, due to the discrete word embedding representations, the large search space hinders the robust training efficiency, resulting in significant time consumption. To overcome this challenge, motivated by the observation that synonym embedding has a small distance, we propose to treat the word substitution as a continuous perturbation on the word embedding representation. The proposed method Text-RS applies random smooth techniques to approximate the word substitution operation, offering a computationally efficient solution that outperforms conventional discrete methods and improves the robustness in training. The evaluation results demonstrate its effectiveness in defending against multiple textual adversarial attacks.

1 Introduction

Language models are powerful tools for natural language processing; however, they have been found to be vulnerable to textual adversarial examples (Jia and Liang, 2017), which are carefully crafted through human-imperceptible changes. These textual adversarial examples pose a significant threat to real-world applications, such as text classification (Song et al., 2021; Kwon and Lee, 2022), text translation (Zhang et al., 2021; Sadrizadeh et al., 2023), question answering (Wallace et al., 2019; Sheng et al., 2021), text-driven image generation (Liu et al., 2023; Millière, 2022), *etc*. Textual adversarial attacks can be categorized into three types, namely character-level perturbation (Ebrahimi et al., 2018; Eger and Benz, 2020), word-level substitution (Ren et al., 2019; Zang et al., 2020; Wang et al., 2021b), and sentence-level rephrasing (Pei and Yue, 2022). Among these, word-level substitution attracts most of the research interest due to its preservation of sentence structure and transferability across various models (Ren et al., 2019). Therefore, our work focuses on defending against word-level substitution adversarial attacks.

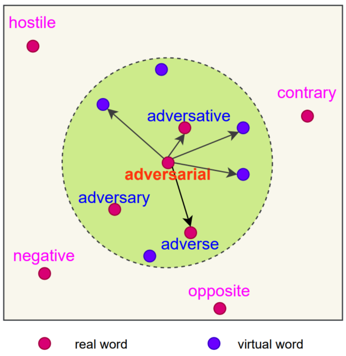


Fig. 1: By adding continuous permutations on word embeddings, our method maps one word to both real words and virtual words, which potentially broadens the optimized region and improves the training efficiency.

Various defense approaches have been proposed to mitigate the impact of word-level text perturbations, such as input transformations (Wang et al., 2021a), adversarial training (Morris et al., 2020), and certified defense (Jia et al., 2019). For example, Wang et al. (2021a) insert a synonym encoder before the input layer to eliminate adversarial substitutions by mapping various synonyms into the same tokens. Adversarial training methods train models on adversarial examples to improve robustness (Wang et al., 2021b; Ke et al., 2022; Zheng et al., 2022). Certified defense methods provide a provable defense radius that theoretically blocks all adversarial examples within that radius (Wang et al., 2021a; Atmakuri et al., 2022). Among these defense methods, certified defense methods achieve a strong defense performance with a theoretical robustness guarantee. However, it is time-consuming because of the construction of a word substitution-based candidate set for the worst-case optimization

[*] Equal contribution. Listing order is random.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1251–1265
March 17-22, 2024 ©2024 Association for Computational Linguistics

for training.

The aforementioned defense methods, especially certified defense methods, mainly perform word substitution in the **discrete** token space, which has an enormous search space and usually results in low efficiency during optimization due to the enumeration and substitution operations for each word. However, for modern language models, input tokens are commonly projected into continuous word embeddings before being fed into subsequent neural networks. The L_2 distance between synonyms in the embedding space approximately follows a compact exponential distribution (Sec. 2.2). This observation naturally motivates us to continuously treat text manipulation and design efficient adversarial defense techniques.

In this work, we propose manipulating texts in the **continuous** embedding space to approximate the word substitution operation for certified defense. Fig. 1 shows an intuitive example of our approach. For the word "adversarial", conventional methods that operate on the word level would map the "adversarial" to the real word "adverse" as an adversarial example, while our method can map the "adversarial" to both real and virtual words by adding permutations on embedding representations. Besides, such a continuous assumption allows us to perturb multiple words in parallel, which significantly broadens the optimized region for compact text representation and improves the training efficiency for certified defense.

On top of continuous perturbation, we further propose a random smooth-based certified adversarial defense framework Text-RS. We integrate the continuous perturbation for word substitution into the certified defense, thus achieving smooth text representation for better model robustness against the text adversarial attack. Extensive results of experiments on popular datasets using different models demonstrate the effectiveness of our method against advanced adversarial text attacks.

2 Method

2.1 Notations

For the text classification task, we define $\mathcal{X}$ as the input text space, $\mathcal{X}_e$ as the embedding space, and $\mathcal{Y}$ as the output category space. Given a text $x = (w_1, w_2, \dots, w_n) \in \mathcal{X}$, an embedding network f_e projects the discrete x to the continuous $x_e \in \mathcal{X}_e$. Subsequently, a text encoder f_p predicts x's category $y \in \mathcal{Y}$ based on x_e. The embedding network f_e and the text encoder f_p are combined as a text classifier $f = f_p \circ f_e$.

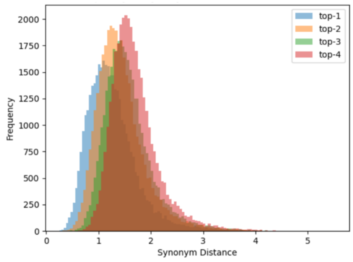


Fig. 2: Statistics of the L_2 distance of GloVe embedding between each word and its i-th synonym, $i = 1, 2, 3, 4$. Results are from the IMDB dataset.

In this work, our main focus is on synonym substitution-based attacks and their defense. We denote the synonyms of a word w as $\mathcal{S}(w)$, which typically consists of the top-k nearest words to w within the Euclidean distance δ in the third-party GloVe embedding space (Pennington et al., 2014) and are post-processed by counter-fitting. Synonym substitution-based attacks commonly replace words $w_i \in x$ with their synonyms $\mathcal{S}(w_i)$ to create an adversarial example x^{adv} such that $f(x^{adv}) = y^{adv} \neq y$, s.t. $d(x, x^{adv}) \leq \epsilon$, where ϵ is a small constant constraining the maximum magnitude of perturbation added to x, and d measures the distance between two texts by counting their differing words. The adversarial defense is to ensure robust estimation against such adversarial samples x^{adv}.

2.2 Motivation

We calculate the L_2 distance between each word and its corresponding i-th synonym, $i = 1, 2, 3, 4$. As depicted in Fig. 2, the distance between one word and its i-th synonym approximately follows an exponential family distribution, with the majority of distance values concentrating around the *mean* value. Additionally, *mean* values of different synonyms are close to each other. Based on these two observations, we make an assumption that discrete word substitutions can be approximated through continuous perturbations in word embedding representations. Consequently, we propose Text-RS, which incorporates continuous perturbation into the model training for certified defense, leading to a broader optimized region and improved training efficiency.

2.3 Practical Algorithm

Specifically, we propose Text-RS to enhance the robustness of a text classifier f when faced with continuous perturbation. Given a text $x \in \mathcal{X}$ and its corresponding word embeddings $x_e \in \mathcal{X}_e$, we simulate the perturbation by injecting random noise ξ into the embeddings, resulting in $f_p(f_e(x)+\xi)$. Our objective is to train f to accurately predict the category of x despite this perturbation. To achieve this, we present two training objectives and introduce an adaptive variable to control the magnitude of the injected noise.

Perturbation loss: We first present a perturbation loss function to smooth the classification surface:

$$\mathcal{L}_s = \|f_p(f_e(x)) - f_p(f_e(x)+\xi)\|_2. \quad (1)$$

$\mathcal{L}_s$ supervises a text classifier to make consistent estimations on noisy and noise-free texts, boosting the classifier's robustness (Peng et al., 2022).

Triplet loss: To achieve more compact text representations for continuous word embeddings, we employ the word-level triplet loss introduced in Yang et al. (2022) to reduce the discrepancy between embedding values of synonyms and simultaneously increase the differentiation among other words, which can be expressed as follows,

$$\begin{aligned}\mathcal{L}_{tr} = &\frac{1}{k}\sum_{w' \in \mathrm{Syn}(w,k)} \|f_e(w) - f_e(w')\|_2 - \\ &\frac{1}{m}\sum_{\hat{w} \notin \mathrm{Syn}(w,k)} \|f_e(w) - f_e(\hat{w})\|_2,\end{aligned} \quad (2)$$

where we utilize top-k synonyms $w' \in \mathrm{Syn}(w,k)$ as positive words and randomly sample m non-synonyms $\hat{w} \notin \mathrm{Syn}(w,k)$ as negative words.

Adaptive variable: In this work, we instantiate ξ as Gaussian noise $\mathcal{N}(0,\sigma^2)$, where σ represents the maximum Euclidean distance between the top-k synonyms. We leave the exploration of other noise types for future work. By assigning the maximum synonym distance as the standard deviation of ξ, we increase the certified robustness radius and enhance the robustness of the text classifier. However, the considerable perturbation on feature representation caused by large k makes it difficult to optimize the parameters and usually leads to substantial performance degradation in the text classification task as identified in Cohen et al. (2019).

Motivated by He et al. (2019) and Xiao et al. (2022), we introduce an adaptive variable α to regulate the magnitude of noise injected into word embeddings $\xi \sim \mathcal{N}\left(0, \mathrm{diag}\left(\{\alpha_i \sigma_i^2 I\}_{i=1}^{n}\right)\right)$, where $\alpha_i \in [0,1]$ and σ_i is the maximum distance between top-k synonyms. We initialize all α_i to 1 and jointly optimize α_i with all model parameters. The introduction of adaptive variables facilitates the optimization of a strongly robust classifier even when k is large.

Overall training objective: In our training process, we integrate perturbation loss (Eq. 1) and triplet loss (Eq. 2) alongside the generally used classification loss $\mathcal{L}_{cls}$ as follows,

$$\mathcal{L}(x,y) = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_{tr}, \quad (3)$$

where λ_1 and λ_2 are two hyper-parameters used to adjust the weight of each loss.

Certified Prediction: Once the text classifier f is trained, we perform certified prediction. Given an input x, we utilize the well-trained f to predict the categories on multiple noisy copies, each crafted with perturbations. We then select the two most common categories as the observation list and employ Bernoulli hypothesis testing to determine their distribution. Based on the significance level, we decide whether to output the most common category as the certified final prediction or reject the prediction to ensure the certified robustness. An overview of the proposed certified prediction is depicted in Fig. A1 of Appendix A.

2.4 Robustness Guarantee

Let a word $w_i \in \mathbb{R}^d$, a sentence containing n words: $x = (w_1, w_2, ..., w_n)$ and function $f : \mathbb{R}^{dn} \to \mathcal{Y}$. Let $\xi \sim \mathcal{N}(0,\Sigma)$, where $\Sigma = \mathrm{diag}(\{\sigma_i^2 I_{d\times d}\}_{i\in[n]}) \in \mathbb{R}^{nd\times nd}$. Let $g(x) = \mathrm{argmax}_c \mathbb{P}(f(x+\xi) = c)$. Suppose that for a specific $x \in \mathbb{R}^{nd}$, there exist $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ such that: $\mathbb{P}(f(x+\xi) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\xi) = c)$. The following Theorem 1 investigates the noise added to the word embedding to guarantee a successful defense for one-word substitution.

Theorem 1 (One-word substitution) *An attacker replaces w_i with $w_i' \in syn(w,k)$, leading to a perturbation $\delta = [0, \cdots, \delta_i, \cdots, 0]$, where $\delta_i = f_e(w_i) - f_e(w_i')$. Then $g(x+\delta) = c_A$ for all $\|\delta_i\| < r$, where*

$$r = \frac{\sigma_i}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})). \quad (4)$$

The proof can be found in Appendix C. We can enumerate the perturbations caused by word-level attacks on each synonym and flexibly select an appropriate Σ to meet our need. For example, for the word w_i to be substituted, we consider

Table 1: Classification accuracy (%) on **IMDB** with various adversarial attack and defense methods.

Defense	CNN						Bi-LSTM					
	Clean	GA	PWWS	PSO	HLA	FGPM	Clean	GA	PWWS	PSO	HLA	FGPM
Standard	**88.8**	7.3	5.3	6.8	14.5	4.4	**89.2**	4.9	3.6	4.3	12.3	4.3
ATFL	86.5	70.7	69.7	72.5	74.0	79.0	86.8	71.1	75.0	73.8	75.6	72.5
ASCC	84.7	79.0	77.2	77.9	78.3	80.9	86.5	73.5	77.8	78.2	80.2	71.7
SEM	86.9	69.2	70.4	70.3	72.2	77.3	87.1	77.4	79.0	79.2	79.9	75.9
ASCL	87.1	79.7	77.5	78.8	79.9	81.5	87.0	79.0	78.5	82.0	82.5	77.3
IBP	83.2	77.5	77.4	77.4	78.7	81.4	82.3	77.0	78.3	79.5	80.2	76.7
RanMASK	85.6	75.0	75.4	70.6	75.1	77.6	82.7	76.1	77.3	78.7	80.1	73.1
Text-RS	86.7	**82.3**	**81.8**	**80.6**	**80.8**	**85.1**	87.9	**83.2**	**81.3**	**82.3**	**83.9**	**78.9**

Table 2: Classification accuracy (%) on **IMDB** with various adversarial attack and defense methods.

Defense	Bert				RoBERTa			
	Clean	BAE	BERT-Attack	CLARE	Clean	BAE	BERT-Attack	CLARE
Standard	**91.4**	13.1	10.5	7.3	**93.7**	12.9	12.6	10.1
ATFL	88.2	33.2	32.6	29.3	91.5	34.7	35.2	30.3
ASCC	87.5	33.9	34.5	35.2	91.1	38.6	39.2	35.5
SEM	90.2	34.8	36.2	37.0	92.4	41.5	41.3	36.7
ASCL	89.5	37.2	37.1	36.5	90.6	40.3	40.5	35.9
RanMASK	90.4	36.8	35.2	33.2	93.1	39.4	39.6	35.3
Text-RS	91.2	**40.5**	**38.3**	**37.8**	92.9	**44.2**	**43.9**	**39.1**

top-k synonyms of it and record the most serious perturbation $\|\delta_i^{max}\| = \max_{j\in[k]} \|f_e(w) - f_e(\text{Syn}(w,j))\|_2$. To successfully defend such an attack with top-k synonyms of w_i, we may apply a large σ_i to make sure $r \geq \|\delta_i^{max}\|$, i.e.,

$$\sigma_i \geq \frac{2\|\delta_i^{max}\|}{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}. \tag{5}$$

Take the example of an attacker replacing only one word in a sentence at a time. For a word w under consideration, the sorted list of top-k synonym substitution perturbation is

$$L = \{\|\delta_i\|_2 | i \in [k]\},$$

where

$$\begin{aligned}\|\delta_i\|_2 &= \|f_e(x) - f_e(x^{adv})\|_2 \\ &= \|f_e(w) - f_e(\text{Syn}(w,i))\|_2.\end{aligned}$$

If we require a successful defense with probability t for that word, we can specify $\|\delta_{\lceil kt \rceil}\|_2$ as the radius r. In other words, to meet our need, we should select a σ_{min} to let $r \geq \|\delta_{\lceil kt \rceil}\|_2$, which means that

$$\sigma_{min} \geq \frac{2\|\delta_{\lceil kt \rceil}\|_2}{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}.$$

In summary, Theorem 1 indicates that the word with a large $\|\delta_i\|_2$ is easier to be attacked and should be protected by adding a Gaussian noise with large σ_i^2. In practice, our adaptive algorithm tends to select larger Gaussian noise for more vulnerable words, which is suggested in Figure A3 in Appendix B.3. Next, we extend the above to the case of multi-word substitution.

Theorem 2 (Multi-word substitution) *Consider an attacker that replaces multiple words at a time. The list $L = [L_1, \cdots, L_n] \in [0,1]^n$ records the positions of all the replaced words. If w_i is replaced, then $L_i = 1$. An attacker replaces w_i with its top-k synonyms $w_i' \in syn(w,k)$. There are $d(x,x') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(w_i, w_i')$ words been replaced. Denote the perturbation of each word $\delta_i = f_e(w_i) - f_e(w_i')$ and the overall perturbation of this sentence $\delta = [L_i\delta_i]_{i\in[n]} \in \mathbb{R}^{nd}$. For each word w_i to be substituted, we record the most serious possible perturbation $\|\delta_i^{max}\| = \max_{j\in[k]} \|f_e(w) - f_e(Syn(w,j))\|_2$. If $\forall i \in [n]$, we have*

$$\sigma_i \geq \frac{2\sqrt{d(x,x')}\|\delta_i^{max}\|}{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}. \tag{6}$$

Then the attack is successfully defended, i.e., $g(x+\delta) = c_A$.

The full proof is in Appendix C. One-word substitution attack means $d(x, x') = 1$. In this case, the result of (6) recovers (5). Intuitively, if an attacker can cause dramatic perturbation to the embedding by replacing some words, then we should add stronger noises to the embedding of such vulnerable words. To protect the model from being attacked, one may add Gaussian noise with different variance to the embedding of the words depending on $\|\delta_i^{max}\|$. A word with large $\|\delta_i^{max}\|$ requires gaussian noise with a large σ_i^2, which is consistent with (6).

3 Experiment

3.1 Experiment Setup

We evaluate our method Text-RS on the IMDB dataset (Maas et al., 2011), which is a classification dataset consisting of $25,000$ movie reviews for training and $25,000$ for testing.

In our evaluation, we first use different defense methods to train two classic architectures, namely the Convolutional Neural Network (CNN) (LeCun et al., 2015) and Bidirectional Long Short-Term Memory (Bi-LSTM) network (Hochreiter and Schmidhuber, 1997) on the IMDB dataset to defend against various attacks. For defense methods, we select ATFL (Wang et al., 2021b), ASCC (Dong et al., 2021), SEM (Wang et al., 2021a), ASCL (Shi et al., 2022), IBP (Jia et al., 2019), and RanMASK (Zeng et al., 2021). For attack methods, we select GA (Alzantot et al., 2018), PWWS (Ren et al., 2019), PSO (Zang et al., 2020), HLA (Maheshwary et al., 2021), and FGPM (Wang et al., 2021b).

Then, we compare the effectiveness of different defense methods on improving the robustness of the advanced Bert architecture (BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019)) against the Bert-related attacks (BAE (Garg and Ramakrishnan, 2020), BERT-Attack (B.A.) (Li et al., 2020b), and CLARE (Li et al., 2020a)). For defense methods, we don't consider the IBP, which lacks the scalability of Bert.

For all experiments, we adopt the classification accuracy as the performance metric. We measure the model's performance on both the benign and adversarial samples to assess whether defense methods can achieve a balance between robustness against adversarial attacks and stability on original, non-adversarial data. A detailed experiment setup can be found in Appendix B.1.

3.2 Numerical Results

Results on CNN and BiLSTM. We present the classification results of CNN and BiLSTM on the IMDB dataset in Table A1, where each row represents a defense method while each column corresponds to an attack method. Among various defense methods, Text-RS demonstrates superior defense performance against all attack methods. Specifically, Text-RS outperforms the runner-up defense method, achieving up to 3.2% and 3.4% improvement for CNN and BiLSTM models, respectively. When compared with certified defense methods such as IBP and RanMask, Text-RS (1) enhances robustness against adversarial attacks with a notable margin and (2) maintains the performance on clean (unmodified) data, indicating Text-RS is a generic framework for handling diverse data.

Results on Bert and RoBERTa. We present the classification results of Bert and RoBERTa on the IMDB dataset in Table 2. Our proposed Text-RS method achieves consistent robustness improvement under different advanced Bert-related attacks. Compared with the runner-up certified defense approach RanMASK, Text-RS boosts a 3% accuracy improvement on average.

In the supplementary material, we also provide results on Ag-News and SST-2 datasets (see Appendix B.2) along with ablation studies of different components in (3) (see Appendix B.3).

4 Conclusion

In our work, motivated by the compact exponential distribution of word embedding space, we propose approximating the discrete word substitution operation as a continuous perturbation on the word embedding representation, thus achieving efficient certified defense training. Numeric results demonstrate the effectiveness of our proposed method.

Limitations

In our work, we use continuous perturbation on word embedding representations for certified robustness training. Although this method enables efficient multi-word substitution in parallel, it incurs inevitable computational costs during noise generation, making it impractical for processing long sentences. Hence, it is worthwhile to explore the possibility of identifying keywords for perturbation. In contrast to perturbing all words in a text, keyword perturbation can enhance both robustness and efficiency.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant 1909912 and 2202124 and by the Center of Excellence in Data Science, an Empire State Development-designated Center of Excellence. This paper does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Shriya Atmakuri, Tejas Chheda, Dinesh Kandula, Nishant Yadav, Taesung Lee, and Hessel Tuinhof. 2022. Robustness of Explanation Methods for NLP Models. *arXiv preprint arXiv:2206.12284*.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *Proceedings of the International Conference on Machine Learning*, pages 1310–1320.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021. Towards robustness against natural language word substitutions. In *International Conference on Learning Representations*.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On Adversarial Examples for Character-Level Neural Machine Translation. In *Proceedings of the International Conference on Computational Linguistics*, pages 653–663.

Steffen Eger and Yannik Benz. 2020. From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 786–803.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*.

Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. 2019. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 588–597.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9-th International Joint Conference on Natural Language Processing*.

Jianpeng Ke, Lina Wang, Aoshuang Ye, and Jie Fu. 2022. Combating Multi-level Adversarial Text with Pruning Based Adversarial Training. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Hyun Kwon and Sanghyun Lee. 2022. Ensemble Transfer Attack Targeting Text Classification Systems. *Comput. Secur.*, 117:102695.

Yann LeCun et al. 2015. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020a. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. 2023. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L. 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*.

Rishabh Maheshwary, Saket Maheshwary, and Vikram Pudi. 2021. Generating Natural Language Attacks in a Hard Label Black Box Setting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13525–13533.

Raphaël Millière. 2022. Adversarial Attacks on Image Generation With Made-Up Words. *arXiv preprint arXiv:2208.04135*.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 119–126.

Weiping Pei and Chuan Yue. 2022. Generating Content-Preserving and Semantics-Flipping Adversarial Text. In *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, pages 975–989.

Yijie Peng, Li Xiao, Bernd Heidergott, L. Jeff Hong, and Henry Lam. 2022. A New Likelihood Ratio Method for Training Artificial Neural Networks. *INFORMS J. Comput.*, 34(1):638–655.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. 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 Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097.

Sahar Sadrizadeh, AmirHossein Dabiri Aghdam, Ljiljana Dolamic, and Pascal Frossard. 2023. Targeted Adversarial Attacks against Neural Machine Translation. *arXiv preprint arXiv:2303.01068*.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-Adversarial Visual Question Answering. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 20346–20359.

Jiahui Shi, Linjing Li, and Daniel Zeng. 2022. ASCL: Adversarial Supervised Contrastive Learning for Defense against Word Substitution Attacks. *Neurocomputing*, 510:59–68.

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal Adversarial Attacks with Natural Triggers for Text Classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3724–3733.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Xiaosen Wang, Jin Hao, Yichen Yang, and Kun He. 2021a. Natural Language Adversarial Defense through Synonym Encoding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 823–833.

Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. 2021b. Adversarial Training with Fast Gradient Projection Method against Synonym Substitution Based Text Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13997–14005.

Li Xiao, Zeliang Zhang, Jinyang Jiang, and Yijie Peng. 2022. Noise Optimization in Artificial Neural Networks. In *18th IEEE International Conference on Automation Science and Engineering, CASE 2022, Mexico City, Mexico, August 20-24, 2022*, pages 1595–1600.

Yichen Yang, Xiaosen Wang, and Kun He. 2022. Robust Textual Embedding against Word-level Adversarial Attacks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 2214–2224.

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.

Jiehang Zeng, Xiaoqing Zheng, Jianhan Xu, Linyang Li, Liping Yuan, and Xuanjing Huang. 2021. Certified Robustness to Text Adversarial Attacks by Randomized[MASK]. *arXiv preprint arXiv:2105.03743*.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting Adversarial Examples for Neural Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1967–1977.

Rui Zheng, Rong Bao, Qin Liu, Tao Gui, Qi Zhang, Xuanjing Huang, Rui Xie, and Wei Wu. 2022. PlugAT: A Plug and Play Module to Defend against Textual Adversarial Attack. In *Proceedings of the International Conference on Computational Linguistics*, pages 2873–2882.

Appendix A Overview of Text-RS

We use Fig. A1 to present more details of our method. Given a sentence x,

1. First, we transform x to the embedding representation $f(x)$;

2. Second, we generate N random noise from the optimized distribution (see the adaptive variable in Section 2.3) to perturb $f(x)$ and generate N noisy embeddings $f(x) + \xi_1, f(x) + \xi_2, ..., f(x) + \xi_N$, which corresponds to the word replacement in sentence level.

3. Next, we forward the N noisy inputs to model and get N predictions $\{y_n\}_{n=1}^{N}$.

4. Last, use the Bernoulli hypothesis testing to decide whether to predict the label with confidence (see the certified prediction in Section 2.3).

Appendix B Experiment Details

B.1 Experiment Setup

Datasets: We evaluate Text-RS on three benchmark datasets, namely IMDB, Ag-News, and SST-2 datasets. IMDB dataset is a binary classification dataset that consists of $25,000$ movie reviews for training and $25,000$ for testing. Ag-News dataset is a topic classification dataset consisting of four classes: World, Sports, Business, and Sci/Tech. There are $30,000$ in news articles for training and $19,000$ for testing in each class. SST-2 dataset is a binary classification dataset on sentiment analysis, which contains $67,000$ movie reviews for training and $1,800$ for testing.

Models: We use two generally used architectures to conduct experiments, including the convolution neural network (CNN) and bidirectional long short-term memory (Bi-LSTM) network. Specifically, we implement the CNN, which contains 3 layers with the filter size 3, 4, and 5, respectively, followed by a max pooling layer and a fully connected layer for classification. We use a one-layer Bi-LSTM, consisting of 128 LSTM units for forward and reverse. We use the pre-trained Glove embedding, which maps the words into a $\mathbb{R}^{300}$ vector.

Baselines: We adopt five advanced adversarial defense techniques for our baselines, including ATFL, ASCC, SEM, ASCL, IBP, and RanMASk. Besides, we use five adversarial attacks to evaluate the performance of the defense methods, including GA, PWWS, PSO, HLA, and FGPM.

Hyper-parameter setting: We train 20 epochs for CNN and BiLSTM on all three datasets to ensure convergence. We follow the same hyper-parameter setting in studied attack and defense methods. For Text-RS, we set $k = 5$, $\lambda_1 = \lambda_2 = 1$, and $n = 20$. Besides, due to the low efficiency of synonym substitution-based attacks, we only evaluate the defensive performance against attacks on 500 samples for each dataset. We use Pytorch to run our experiments. We conduct our experiments on a server which has two Intel(R) Xeon(R) Gold 5118 CPUs. Each of CPUs has 12 cores @2.30GHz supporting 24 hardware threads. There is a Titan RTX GPU which consists of 24 GB device memory. There are 256 GB DDR4 memories on the server. The mean training time of all models is 3.35 hours.

B.2 Evaluations on Ag-News and SST-2

Among various defense methods, Text-RS demonstrates superior defense performance against different attack methods. On the IMDB dataset, Text-RS outperforms the runner-up defense method, achieving up to 3.2% and 3.4% improvement for CNN and BiLSTM models, respectively. On Ag-News, Text-RS shows 0.2% and 1.7% improvement over the runner-up, and on SST-2, Text-RS demonstrates 4.1% and 5.0% improvement. While certified defense methods such as IBP and RanMask fail to deliver good results on BiLSTM with the three datasets, Text-RS still performs well. It is worth noting that Text-RS not only improves adversarial robustness but also maintains the original task performance (Clean), unlike certified defense methods.

B.3 Ablation Study

On the optimized noise: To evaluate the efficacy of the proposed noise injection method in enhancing adversarial robustness, we established two baseline models: the standard training model with noise prediction (Standard$_\mathrm{r}$) and the random smoothing training model with unoptimized noise (RS$_\mathrm{u}$). The results, presented in Table A5, reveal that while random smoothing during inference (Standard$_r$) provides a significant improvement in adversarial robustness, it also impairs the performance on benign samples. In contrast, the noise injection-based training approach enhances both adversarial robustness and task performance. These results affirm the

Defense	CNN						BiLSTM					
	Clean	GA	PWWS	PSO	HLA	FGPM	Clean	GA	PWWS	PSO	HLA	FGPM
Standard	**88.8**	7.3	5.3	6.8	14.5	4.4	**89.2**	4.9	3.6	4.3	12.3	4.3
ATFL	86.5	70.7	69.7	72.5	74.0	79.0	86.8	71.1	75.0	73.8	75.6	72.5
ASCC	84.7	79.0	77.2	77.9	78.3	80.9	86.5	73.5	77.8	78.2	80.2	71.7
SEM	86.9	69.2	70.4	70.3	72.2	77.3	87.1	77.4	79.0	79.2	79.9	75.9
ASCL	87.1	79.7	77.5	78.8	79.9	81.5	87.0	79.0	78.5	82.0	82.5	77.3
IBP	83.2	77.5	77.4	77.4	78.7	81.4	82.3	77.0	78.3	79.5	80.2	76.7
RanMASK	85.6	75.0	75.4	70.6	75.1	77.6	82.7	76.1	77.3	78.7	80.1	73.1
Text-RS	86.7	**82.3**	**81.8**	**80.6**	**80.8**	**85.1**	87.9	**83.2**	**81.3**	**82.3**	**83.9**	**78.9**

Table A1: Classification accuracy (%) on **IMDB**.

Defense	CNN						BiLSTM					
	Clean	GA	PWWS	PSO	HLA	FGPM	Clean	GA	PWWS	PSO	HLA	FGPM
Standard	**93.3**	33.2	32.9	32.9	43.5	32.3	**92.4**	32.8	32.8	32.7	43.1	32.1
ATFL	92.7	87.9	88.0	86.8	**90.3**	**89.5**	91.6	88.2	87.1	87.4	90.1	88.2
ASCC	89.4	83.3	83.0	83.0	81.7	86.2	89.5	74.4	73.6	74.1	75.8	74.9
SEM	91.8	80.1	79.2	83.8	86.7	79.6	88.6	87.6	87.5	87.9	90.9	88.3
ASCL	90.9	85.0	85.1	84.8	83.9	85.4	88.7	68.6	86.9	86.2	88.6	87.1
IBP	89.4	84.2	87.6	86.2	87.0	87.2	87.9	76.3	74.0	73.5	77.1	74.6
RanMASK	88.9	83.7	84.5	86.2	86.4	87.6	88.2	72.6	69.4	69.3	75.4	74.4
Text-RS	90.4	**88.5**	**89.8**	**87.6**	88.5	89.1	92.3	**90.5**	**89.1**	**90.5**	**91.5**	**89.5**

Table A2: Classification accuracy (%) on **Ag-News**.

Defense	CNN						BiLSTM					
	Clean	GA	PWWS	PSO	HLA	FGPM	Clean	GA	PWWS	PSO	HLA	FGPM
Standard	**91.8**	3.1	2.4	2.4	13.3	2.6	**92.5**	2.8	2.7	2.9	12.9	2.0
ATFL	91.2	64.2	62.7	62.1	72.1	65.8	92.3	63.1	62.8	63.6	74.2	64.6
ASCC	**91.8**	68.6	68.3	68.4	69.5	63.9	91.9	67.8	68.5	68.2	74.1	71.7
SEM	91.1	67.5	67.1	66.8	68.5	64.5	91.4	67.0	66.1	66.8	70.5	66.1
ASCL	91.1	69.5	69.9	70.5	70.5	65.2	92.0	69.8	69.0	69.2	75.4	73.1
IBP	90.4	69.8	69.6	69.7	72.0	64.3	91.0	69.0	67.9	69.3	71.4	66.7
RanMASK	91.5	67.9	68.7	67.1	69.7	61.9	90.7	67.3	66.5	67.6	68.3	64.8
Text-RS	**91.8**	**73.5**	**72.1**	**72.8**	**75.7**	**72.3**	91.9	**74.8**	**74.2**	**75.3**	**78.6**	**74.8**

Table A3: Classification accuracy (%) on **SST-2**.

Table A4: Classification accuracy (%) against various adversarial attacks on three datasets for CNN and BiLSTM.

Table A5: Classification accuracy (%) against various adversarial attacks on IMDB dataset for CNN. NI: Noise Injection, SO: Scale Optimization, PLoss: Perturbation Loss, SLoss: Synonym Loss.

Method	NI	SO	PLoss	SLoss	Clean	GA	PWWS	PSO	HLA	FGPM
Standard	✗	✗	✗	✗	88.8	7.3	5.3	6.8	14.5	4.4
Standard$_{\text{r}}$	✓	✗	✗	✗	78.4	65.3	66.8	68.5	75.0	62.4
RS$_{\text{u}}$	✓	✗	✓	✗	85.1	67.2	67.2	68.6	74.8	65.5
RS$_{\text{-s}}$	✓	✓	✓	✗	85.6	78.1	76.9	77.3	74.7	80.2
Text-RS	✓	✓	✓	✓	86.7	**82.3**	**81.8**	**80.6**	**80.8**	**85.1**

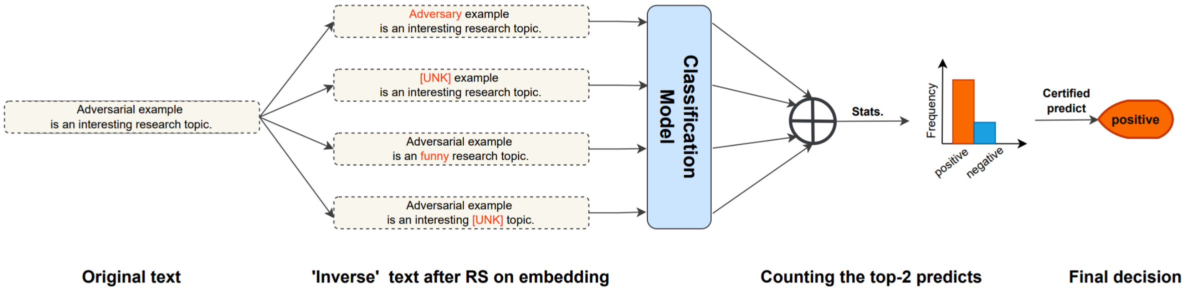
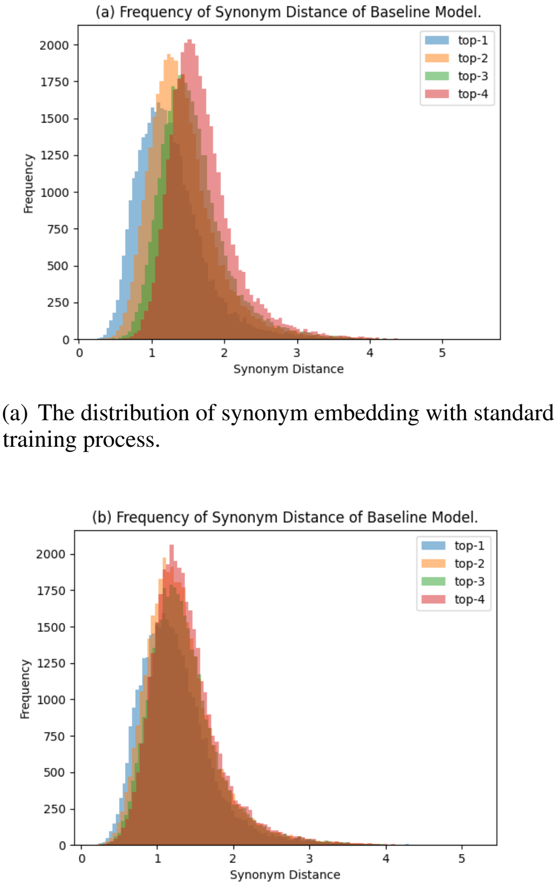

Adversarial example
is an interesting research topic.
Adversary example
is an interesting research topic.
[UNK] example
is an interesting research topic.
Adversarial example
is an funny research topic.
Adversarial example
is an interesting [UNK] topic.
Classification Model
Stats.
Frequency
positive
negative
Certified predict
positive
Original text
'Inverse' text after RS on embedding
Counting the top-2 predicts
Final decision

Fig. A1: Overview of our proposed certified prediction method based on the assumption of continuous perturbation.

effectiveness of our proposed method.

On the synonym embedding: Text-RS narrows the synonym and moves away from other words to achieve the certified defense by introducing the loss function (2) of SEM. Here, to study the influence of the synonym loss, we use Text-RS without synonym loss (RS$_{\text{-s}}$ in Tab. A5) to train a model and evaluate the performance to validate the performance. From the result, the effectiveness of the introduction of synonym loss can be verified. On the other hand, randomized smoothing training compact with synonym loss contributes to improving the adversarial transferability. Besides, as discussed in Section, we visualize the mean distance of the top-k synonym. 2.2 again. Comparing Fig. 2(a) and Fig. 2(b), it can be clearly identified that the L_2 distance of synonym has been reduced compared with the baseline.

On the learning of σ: To guarantee the robustness under the multi-word substitution, the learned σ_i for word w_i should be proportional to the minimum distance between the synonyms, as analyzed in (6). To further verify the robustness guarantee theory, we collect the minimum distance d between synonyms and corresponding σ for every world as (d, σ) and present the distribution relationship in Fig. A3. From the scatter plot, it can be noticed that with an increasing magnitude of the minimum distance between synonyms, the learned σ corresponding increases in statistics. We also use a linear model to fit the distribution, which is presented in red. The slope ratio for the linear model is 0.17, which shows the positive correlation between d and σ, thus providing more evidence for (13).



(a) The distribution of synonym embedding with standard training process.

(b) The distribution of synonym embedding with Text-RS.

Fig. A2: Ablation study on Text-RS.

Appendix C Proof

C.1 Theorems

Theorem 3 (Anisotropic Gaussians) *Let* $f : \mathbb{R}^d \to \mathcal{Y}$ *be any deterministic or random function. Let* $\varepsilon \sim \mathcal{N}(0, \Sigma)$, *where* $\Sigma = diag\{\sigma_i^2\}(i \in [d])$, *and* $\min_{i \in [d]} \sigma_i = \sigma_{min}$. *Let* $g(x) = \operatorname{argmax}_c \mathbb{P}(f(x+\varepsilon) = c)$. *Suppose that for a specific* $x \in \mathbb{R}^d$, *there exist* $c_A \in \mathcal{Y}$ *and* $\underline{p_A}, \overline{p_B} \in [0, 1]$ *such that:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \quad (7)$$

Then $g(x+\delta) = c_A$ *for all* $\|\delta\| < r$, *where*

$$r = \frac{\sigma_{min}}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \quad (8)$$

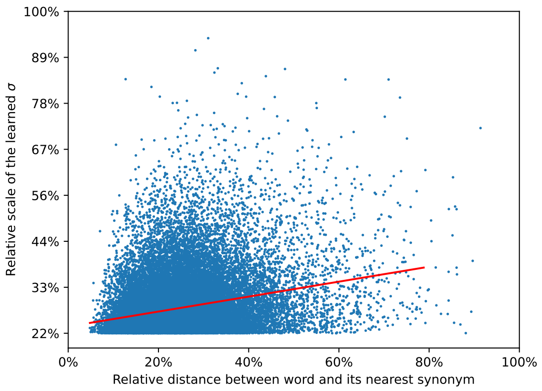


Fig. A3: Visualization of the relationship between the synonym distance and optimized σ for given words.

Analysis The above theorem is appropriate for images. We extend (Cohen et al., 2019) from isotropic gaussian to anisotropic gaussian. The only difference is σ and σ_{min}. And $\sigma = \sigma_{min}$ will recover the result in (Cohen et al., 2019).

However, considering the nature of word-level substitution, only some specific part of $x = (w_1, w_2, ..., w_n)$ will be affected. The following theorem extends the result to one-word substitution.

Theorem 4 (One-word substitution) *Let a word $w_i \in \mathbb{R}^d$, a sentence containing n words: $x = (w_1, w_2, ..., w_n)$ and function $f : \mathbb{R}^{dn} \to \mathcal{Y}$. Let $\xi \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = diag(\{\sigma_i^2 I_{d\times d}\}_{i\in[n]}) \in \mathbb{R}^{nd\times nd}$. Let $g(x) = \text{argmax}_c \mathbb{P}(f(x + \xi) = c)$. Suppose that for a specific $x \in \mathbb{R}^{nd}$, there exist $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ such that:*

$$\mathbb{P}(f(x+\xi) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c\neq c_A} \mathbb{P}(f(x+\xi) = c). \tag{9}$$

An attacker replaces w_i with $w_i' \in syn(w, k)$, leading to a perturbation $\delta = [0, \cdots, \delta_i, \cdots, 0]$, where

$$\delta_i = f_e(w_i) - f_e(w_i').$$

Then $g(x + \delta) = c_A$ for all $\|\delta_i\| < r$, where

$$r = \frac{\sigma_i}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \tag{10}$$

Analysis With Theorem 3 and the experiments, we can enumerate the perturbations caused by word-level attacks on each synonym and flexibly select an appropriate Σ to meet our need. For example, for the word w_i to be substituted, we consider top-k synonyms of it and record the most serious perturbation

$$\|\delta_i^{max}\| = \max_{j\in[k]} \| f_e(w) - f_e(\text{Syn}(w, j)) \|_2.$$

To successfully defend such an attack with top-k synonyms of w_i, we may apply a large σ_i to make sure $r \geq \|\delta_i^{max}\|$, i.e.,

$$\sigma_i \geq \frac{2\|\delta_i^{max}\|}{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}. \tag{11}$$

Next, we extend the above to the case of multi-word substitution.

Theorem 5 (Multi-word substitution) *Let a word $w_i \in \mathbb{R}^d$, a text containing n words: $x = (w_1, w_2, ..., w_n)$ and function $f : \mathbb{R}^{dn} \to \mathcal{Y}$. Let $\xi \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = diag(\{\sigma_i^2 I_{d\times d}\}_{i\in[n]}) \in \mathbb{R}^{nd\times nd}$. Let $g(x) = \text{argmax}_c \mathbb{P}(f(x+\xi) = c)$. Suppose that for a specific $x \in \mathbb{R}^{nd}$, there exist $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ such that:*

$$\mathbb{P}(f(x+\xi) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c\neq c_A} \mathbb{P}(f(x+\xi) = c). \tag{12}$$

Consider an attacker that replaces multiple words at a time. The list $L = [L_1, \cdots, L_n] \in [0, 1]^n$ records the positions of all the replaced words. If w_i is replaced, then $L_i = 1$. An attacker replaces w_i with its top-k synonyms $w_i' \in syn(w, k)$. There are $d(x, x') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(w_i, w_i')$ words been replaced. Denote the perturbation of each word $\delta_i = f_e(w_i) - f_e(w_i')$ and the overall perturbation of this sentence $\delta = [L_i\delta_i]_{i\in[n]} \in \mathbb{R}^{nd}$.

For each word w_i to be substituted, we record the most serious possible perturbation

$$\|\delta_i^{max}\| = \max_{j\in[k]} \| f_e(w) - f_e(Syn(w, j)) \|_2.$$

If $\forall i \in [n]$, we have

$$\sigma_i \geq \frac{2\sqrt{d(x, x')}\|\delta_i^{max}\|}{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}. \tag{13}$$

Then the attack is successfully defended, i.e., $g(x + \delta) = c_A$.

Analysis One-word substitution attack means $d(x, x') = 1$. In this case, the result of (13) recovers (11). To protect the model from being attacked, one may add Gaussian noise with different variance to the embedding of the words depending on $\|\delta_i^{max}\|$. A word with large $\|\delta_i^{max}\|$ requires a large σ_i^2. Intuitively, if an attacker can cause dramatic perturbation to the embedding by replacing some words, then we should add a stronger noise to the embedding of such vulnerable words.

C.2 Lemmas

Lemma 1 (Neyman-Pearson) *Let X and Y be random variables in $\mathbb{R}^d$ with densities μ_X and μ_Y. Let $h : \mathbb{R}^d \to \{0, 1\}$ be a random or deterministic function. Then:*

1. *If $S = \left\{z \in \mathbb{R}^d : \frac{\mu_Y(z)}{\mu_X(z)} \le t\right\}$ for some $t > 0$ and $\mathbb{P}(h(X) = 1) \ge \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \ge \mathbb{P}(Y \in S)$.*

2. *If $S = \left\{z \in \mathbb{R}^d : \frac{\mu_Y(z)}{\mu_X(z)} \ge t\right\}$ for some $t > 0$ and $\mathbb{P}(h(X) = 1) \le \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \le \mathbb{P}(Y \in S)$.*

The following is Neyman-Pearson lemma for Anisotropic Gaussians with different means.

Lemma 2 (Neyman-Pearson (Anisotropic)) *Let $X \sim \mathcal{N}(x, \Sigma)$ and $Y \sim \mathcal{N}(x + \delta, \Sigma)$, where $\Sigma = diag\{\sigma_i^2\}(i = 1, \cdots, d)$. Let $h : \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function. Then:*

1. If $S = \left\{z \in \mathbb{R}^d : (\Sigma^{-\frac{1}{2}}\delta)^T z \le \beta\right\}$ for some β and $\mathbb{P}(h(X) = 1) \ge \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \ge \mathbb{P}(Y \in S)$

2. If $S = \left\{z \in \mathbb{R}^d : (\Sigma^{-\frac{1}{2}}\delta)^T z \ge \beta\right\}$ for some β and $\mathbb{P}(h(X) = 1) \le \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \le \mathbb{P}(Y \in S)$

This lemma is the special case of Lemma 1 when X and Y are anisotropic Gaussians with means x and $x + \delta$.

By Lemma 1 it suffices to simply show that for any β, there is some $t > 0$ for which:

$$\{z : (\Sigma^{-\frac{1}{2}}\delta)^T z \le \beta\} = \left\{z : \frac{\mu_Y(z)}{\mu_X(z)} \le t\right\} \quad \text{and}$$
$$\{z : (\Sigma^{-\frac{1}{2}}\delta)^T z \ge \beta\} = \left\{z : \frac{\mu_Y(z)}{\mu_X(z)} \ge t\right\} \tag{14}$$

The likelihood ratio for this choice of X and Y turns out to be:

$$\begin{aligned}\frac{\mu_Y(z)}{\mu_X(z)} &= \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{(z_i - (x_i + \delta_i))^2}{\sigma_i^2})\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{(z_i - x_i)^2}{\sigma_i^2}\right)} \\ &= \exp\left(\frac{1}{2}\sum_{i=1}^d \frac{2z_i\delta_i - \delta_i^2 - 2x_i\delta_i}{\sigma_i^2}\right) \\ &= \exp((\Sigma^{-1}\delta)^T z + b)\end{aligned}$$

where $b = -(\Sigma^{-1}\delta)^T x - \frac{1}{2}\|\Sigma^{-1}\delta\|_2^2$ is a constant w.r.t z. Therefore, given any $\beta = \sum_{i=1}^d \beta_i$, where $\beta_i \le \frac{\delta_i}{\sigma_i} z_i$. we may take $t = \exp(\sum_{i=1}^d \frac{\beta_i}{\sigma_i} + b)$, noticing that

$$(\Sigma^{-\frac{1}{2}}\delta)^T z \le \beta \iff \exp((\Sigma^{-1}\delta)^T z + b) \le t$$
$$(\Sigma^{-\frac{1}{2}}\delta)^T z \ge \beta \iff \exp((\Sigma^{-1}\delta)^T z + b) \ge t$$

So the proof is complete.

C.3 Proof of Theorem 3

To show that $g(x + \delta) = c_A$, it follows from the definition of g that we need to show that

$$\mathbb{P}(f(x + \delta + \varepsilon) = c_A) > \max_{c_B \ne c_A} \mathbb{P}(f(x + \delta + \varepsilon) = c_B) \tag{15}$$

We will prove that $\mathbb{P}(f(x+\delta+\varepsilon) = c_A) > \mathbb{P}(f(x+\delta+\varepsilon) = c_B)$ for every class $c_B \ne c_A$. Fix one such class c_B without loss of generality.

For brevity, define the random variables

$$\begin{aligned}X &:= x + \varepsilon = \mathcal{N}(x, \Sigma) \\ Y &:= x + \delta + \varepsilon = \mathcal{N}(x + \delta, \Sigma)\end{aligned}$$

In this notation, we know that

$$\mathbb{P}(f(X) = c_A) \ge \underline{p_A} \quad \text{and} \quad \mathbb{P}(f(X) = c_B) \le \overline{p_B} \tag{16}$$

and our goal is to show that

$$\mathbb{P}(f(Y) = c_A) > \mathbb{P}(f(Y) = c_B) \tag{17}$$

Define the half-spaces:

$$\begin{aligned}A &:= \{z : (\Sigma^{-\frac{1}{2}}\delta)^T (z - x) \le \|\delta\|\Phi^{-1}(\underline{p_A})\} \\ B &:= \{z : (\Sigma^{-\frac{1}{2}}\delta)^T (z - x) \ge \|\delta\|\Phi^{-1}(1 - \overline{p_B})\}\end{aligned}$$

Algebra (deferred to C.6) shows that $\mathbb{P}(X \in A) = \underline{p_A}$. Therefore, by (16) we know that $\mathbb{P}(f(X) =$

$c_A) \geq \mathbb{P}(X \in A)$. Hence we may apply Lemma 2 with $h(z) := \mathbf{1}[f(z) = c_A]$ to conclude:

$$\mathbb{P}(f(Y) = c_A) \geq \mathbb{P}(Y \in A) \tag{18}$$

Similarly, algebra shows that $\mathbb{P}(X \in B) = \overline{p_B}$. Therefore, by (16) we know that $\mathbb{P}(f(X) = c_B) \leq \mathbb{P}(X \in B)$. Hence we may apply Lemma 2 with $h(z) := \mathbf{1}[f(z) = c_B]$ to conclude:

$$\mathbb{P}(f(Y) = c_B) \leq \mathbb{P}(Y \in B) \tag{19}$$

To guarantee (17), we see from (18, 19) that it suffices to show that $\mathbb{P}(Y \in A) > \mathbb{P}(Y \in B)$, as this step completes the chain of inequalities

$$\begin{aligned}\mathbb{P}(f(Y) = c_A) \geq \mathbb{P}(Y \in A) > \\ \mathbb{P}(Y \in B) \geq \mathbb{P}(f(Y) = c_B)\end{aligned} \tag{20}$$

Let $R(A, x) = \frac{x^T A x}{x^T x}$ be the Rayleigh quotient for symmetric matrix A and vector x. In our setting, Σ is a symmetric and positive-definite matrix, so its eigenvalues are all greater than zero. Based on the deferred derivation in C.6, we know that $R(\Sigma^{-\frac{1}{2}}, \delta) > 0$.

We can compute the following:

$$\mathbb{P}(Y \in A) = \Phi\left(\Phi^{-1}(\underline{p_A}) - \|\delta\| R(\Sigma^{-\frac{1}{2}}, \delta)\right) \tag{21}$$

$$\mathbb{P}(Y \in B) = \Phi\left(\Phi^{-1}(\overline{p_B}) + \|\delta\| R(\Sigma^{-\frac{1}{2}}, \delta)\right) \tag{22}$$

Finally, $\mathbb{P}(Y \in A) > \mathbb{P}(Y \in B)$ holds if and only if:

$$\|\delta\| < \frac{1}{2R(\Sigma^{-\frac{1}{2}}, \delta)}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \tag{23}$$

Furthermore, we just need to let the Rayleigh quotient takes the maximum. We know that

$$\max R(\Sigma^{-\frac{1}{2}}, \delta) = \lambda_{max}(\Sigma^{-\frac{1}{2}}) = \frac{1}{\sigma_{min}}$$

Therefore, we have $R(\Sigma^{-\frac{1}{2}}, \delta) \geq \sigma_{min}$, which means that

$$\begin{aligned}\|\delta\| &< \frac{\sigma_{min}}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \\ &\leq \frac{1}{2R(\Sigma^{-\frac{1}{2}}, \delta)}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))\end{aligned} \tag{24}$$

The proof is complete.

C.4 Proof of Theorem 4

Before (23), the proof for Theorem 3 and 4 are the same. Recall that $\delta = [0, \cdots, \delta_i, \cdots, 0]$, so we have

$$R(\Sigma^{-\frac{1}{2}}, \delta) = \frac{\delta^T \Sigma^{-\frac{1}{2}} \delta}{\delta^T \delta} = \frac{\delta_i^T (\frac{1}{\sigma_i} I) \delta_i}{\delta_i^T \delta_i} = \frac{1}{\sigma_i}.$$

Finally, Combining it with (23) and we obtain:

$$\|\delta\| < \frac{\sigma_i}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \tag{25}$$

The proof is complete.

C.5 Proof of Theorem 5

Before (23), the proof for Theorem 3 and 5 are the same. Recall that the list $L = [L_1, \cdots, L_n] \in [0,1]^n$ records the positions of all the replaced words. An attacker replaces w_i with w_i', where $L_i = 1$. The perturbation of each word $\delta_i = f_e(w_i) - f_e(w_i')$. The overall perturbation of this sentence satisfies: $\|\delta\| = \sqrt{\sum_{i \in [n], L_i = 1} \|\delta_i\|^2}$.

Therefore, for multi-word substitution, we have

$$\begin{aligned}R(\Sigma^{-\frac{1}{2}}, \delta) &= \frac{\delta^T \Sigma^{-\frac{1}{2}} \delta}{\delta^T \delta} \\ &= \frac{\sum_{i \in [n], L_i = 1} \frac{1}{\sigma_i} \|\delta_i\|^2}{\sum_{i \in [n], L_i = 1} \|\delta_i\|^2} \\ &= \sum_{i \in [n], L_i = 1} \frac{1}{\sigma_i} \frac{\|\delta_i\|^2}{\|\delta\|^2}.\end{aligned} \tag{26}$$

If $\forall i \in [n]$, we have

$$\sigma_i \geq \frac{2\sqrt{d(x, x')}\|\delta_i^{max}\|}{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}.$$

Then

$$\begin{aligned}R(\Sigma^{-\frac{1}{2}}, \delta) &= \sum_{i \in [n], L_i = 1} \frac{1}{\sigma_i} \frac{\|\delta_i\|^2}{\|\delta\|^2} \\ &\leq \sum_{i \in [n], L_i = 1} \frac{\|\delta_i\|^2}{\|\delta\|^2} \cdot \frac{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}{2\sqrt{d(x, x')}\|\delta_i^{max}\|} \\ &\leq \sum_{i \in [n], L_i = 1} \frac{\|\delta_i\|}{2\|\delta\|^2} \cdot \frac{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}{\sqrt{d(x, x')}} \\ &= \frac{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}{2\|\delta\|} \\ &\quad \cdot \left(\frac{1}{\sqrt{d(x, x')}} \sum_{i \in [n], L_i = 1} \frac{\|\delta_i\|}{\|\delta\|}\right)\end{aligned} \tag{27}$$

Notice that $\sum_{i\in[n],L_i=1} 1 = d(x,x')$. According to AM-QM Inequality mentioned in C.6, we have

$$\frac{1}{\sqrt{d(x,x')}} \sum_{i\in[n],L_i=1} \frac{\|\delta_i\|}{\|\delta\|} \le \sum_{i\in[n],L_i=1} \frac{\|\delta_i\|^2}{\|\delta\|^2} = 1.$$

In other words,

$$R(\Sigma^{-\frac{1}{2}}, \delta) \le \frac{\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})}{2\|\delta\|},$$

which is consistent with (23), i.e., $\mathbb{P}(Y \in A) > \mathbb{P}(Y \in B)$. So the attack is defended successfully. The proof is complete.

C.6 Deferred Algebra

C.6.1 The properties of Rayleigh quotient

$$R(A,x) = \frac{x^T A x}{x^T x} \in [\lambda_{min}, \lambda_{max}],$$

where $R(A,x)$ is the Rayleigh quotient for symmetric matrix A and vector x. And $\lambda_{max}, \lambda_{min}$ are the maximum and minimum eigenvalues of A.

We introduce Lagrange multiplier $\lambda \ge 0$. Without loss of generality, we set $\|x\|_2^2 = 1$ to obtain the extreme value of $R(A,x)$. So

$$L(x,\lambda) = x^T A x - \lambda(\|x\|_2^2 - 1).$$

Taking the derivative w.r.t. x and set it to zero:

$$\frac{\partial L(x,\lambda)}{\partial x} = Ax - \lambda x = 0.$$

So λ is one of the eigenvalues of A when $L(x,\lambda)$ takes an extreme value. Based on such result, when $R(A,x)$ takes an extreme value, there holds:

$$R(A,x) = \frac{x^T \lambda x}{x^T x} = \lambda \in [\lambda_{min}, \lambda_{max}].$$

Further, in our setting, Σ is a symmetric and positive-definite matrix, so its eigenvalues are all greater than zero, which means that $R(\Sigma^{-1}, x) > 0$.

C.6.2 Others

A frequently used derivation.

$$(\Sigma^{-\frac{1}{2}}\delta)^T \mathcal{N}(0,\Sigma) = \|\delta\| Z,$$

where $Z \sim \mathcal{N}(0,1)$.

Let $T = (t_1, t_2, \cdots, t_d)^T \sim \mathcal{N}(0,\Sigma)$. So we have $t_i \sim \mathcal{N}(0,\sigma_i^2)$, where $i = 1, \cdots, d$.

$$\begin{aligned}
&(\Sigma^{-\frac{1}{2}}\delta)^T \mathcal{N}(0,\Sigma) \\
=&(\Sigma^{-\frac{1}{2}}\delta)^T (t_1, t_2, \cdots, t_d)^T \\
=&\sum_{i=1}^{d} \frac{\delta_i}{\sigma_i} t_i \\
=&\mathcal{N}(0, \sum_{i=1}^{d} \delta_i^2) \qquad (t_i \sim \mathcal{N}(0,\sigma_i^2)) \\
=&\mathcal{N}(0, \|\delta\|^2) \\
=&\|\delta\| \mathcal{N}(0,1) \\
=&\|\delta\| Z \qquad (Z \sim \mathcal{N}(0,1))
\end{aligned}$$

Claim. $\mathbb{P}(X \in A) = \underline{p_A}$

Recall that $X \sim \mathcal{N}(x,\Sigma)$ and $A = \{z : (\Sigma^{-\frac{1}{2}}\delta)^T(z - x) \le \|\delta\|\Phi^{-1}(\underline{p_A})\}$.

$$\begin{aligned}
\mathbb{P}(X \in A) &= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T (X - x) \\
&\le \|\delta\|\Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T \mathcal{N}(0,\Sigma) \\
&\le \|\delta\|\Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(\|\delta\| Z \le \|\delta\|\Phi^{-1}(\underline{p_A})) \\
&\qquad (Z \sim \mathcal{N}(0,1)) \\
&= \Phi(\Phi^{-1}(\underline{p_A})) \\
&= \underline{p_A}
\end{aligned}$$

Claim. $\mathbb{P}(X \in B) = \overline{p_B}$

Recall that $X \sim \mathcal{N}(x,\Sigma)$ and $B = \{z : (\Sigma^{-\frac{1}{2}}\delta)^T(z - x) \le \|\delta\|\Phi^{-1}(1 - \overline{p_B})\}$.

$$\begin{aligned}
&\mathbb{P}(X \in A) \\
&= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T (X - x) \ge \|\delta\|\Phi^{-1}(1 - \overline{p_B})) \\
&= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T \mathcal{N}(0,\Sigma) \ge \|\delta\|\Phi^{-1}(1 - \overline{p_B})) \\
&= \mathbb{P}(\|\delta\| Z \ge \|\delta\|\Phi^{-1}(1 - \overline{p_B})) \\
&\qquad (Z \sim \mathcal{N}(0,1)) \\
&= \mathbb{P}(Z \ge \Phi^{-1}(1 - \overline{p_B})) \\
&= 1 - \Phi(\Phi^{-1}(1 - \overline{p_B})) \\
&= \overline{p_B}
\end{aligned}$$

Claim.

$$\mathbb{P}(Y \in A) = \Phi\left(\Phi^{-1}(\underline{p_A}) - \|\delta\| R(\Sigma^{-\frac{1}{2}}, \delta)\right)$$

Recall that $Y \sim \mathcal{N}(x + \delta, \Sigma)$ and $A = \{z :$

$(\Sigma^{-\frac{1}{2}}\delta)^T(z-x) \leq \|\delta\|\Phi^{-1}(\underline{p_A})\}$.

$$\begin{aligned}&\mathbb{P}(Y \in A)\\ &= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T(Y-x)\\ &\leq \|\delta\|\Phi^{-1}(\underline{p_A}))\\ &= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T\mathcal{N}(0,\Sigma) + \delta^T\Sigma^{-\frac{1}{2}}\delta\\ &\leq \|\delta\|\Phi^{-1}(\underline{p_A}))\\ &= \mathbb{P}(\|\delta\|Z \leq \|\delta\|\Phi^{-1}(\underline{p_A}) - \delta^T\Sigma^{-\frac{1}{2}}\delta)\\ &\qquad\qquad\qquad\qquad (Z \sim \mathcal{N}(0,1))\\ &= \mathbb{P}\left(Z \leq \Phi^{-1}(\underline{p_A}) - \frac{\delta^T\Sigma^{-\frac{1}{2}}\delta}{\|\delta\|}\right)\\ &= \Phi\left(\Phi^{-1}(\underline{p_A}) - \|\delta\|R(\Sigma^{-\frac{1}{2}},\delta)\right).\end{aligned}$$

Claim.

$$\mathbb{P}(Y \in B) = \Phi\left(\Phi^{-1}(\overline{p_B}) + \|\delta\|R(\Sigma^{-\frac{1}{2}},\delta)\right)$$

Recall that $Y \sim \mathcal{N}(x+\delta,\Sigma)$ and $B = \{z : (\Sigma^{-\frac{1}{2}}\delta)^T(z-x) \geq \|\delta\|\Phi^{-1}(1-\overline{p_B})\}$.

$$\begin{aligned}&\mathbb{P}(Y \in B)\\ &= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T(Y-x)\\ &\geq \|\delta\|\Phi^{-1}(1-\overline{p_B}))\\ &= \mathbb{P}((\Sigma^{-\frac{1}{2}}\delta)^T\mathcal{N}(0,\Sigma) + \delta^T\Sigma^{-\frac{1}{2}}\delta\\ &\geq \|\delta\|\Phi^{-1}(1-\overline{p_B}))\\ &= \mathbb{P}(\|\delta\|Z + \delta^T\Sigma^{-\frac{1}{2}}\delta\\ &\geq \|\delta\|\Phi^{-1}(1-\overline{p_B})) \qquad (Z \sim \mathcal{N}(0,1))\\ &= \mathbb{P}\left(Z \geq \Phi^{-1}(1-\overline{p_B}) - \frac{\delta^T\Sigma^{-\frac{1}{2}}\delta}{\|\delta\|}\right)\\ &= \mathbb{P}\left(Z \leq \Phi^{-1}(\overline{p_B}) + \frac{\delta^T\Sigma^{-\frac{1}{2}}\delta}{\|\delta\|}\right)\\ &= \Phi\left(\Phi^{-1}(\overline{p_B}) + \|\delta\|R(\Sigma^{-\frac{1}{2}},\delta)\right)\end{aligned}$$

C.6.3 AM-QM Inequality

For $x_1, \cdots, x_n \in \mathbb{R}_+$, we have

$$\frac{\sum_{i=1}^n x_i}{n} \leq \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}. \tag{28}$$

According to the Jensen's inequality,

$$f\left(\frac{\sum_{i=1}^n x_i}{n}\right) \leq \frac{\sum_{i=1}^n f(x_i)}{n}.$$

For a convex function $f(x) = x^2$, (28) holds. So the proof is complete.

Furthermore, it is obvious that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i \leq \sqrt{\sum_{i=1}^n x_i^2},$$

which will be used in our proof.