Density-Aware Differentially Private Textual Perturbations Using Truncated Gumbel Noise

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, Nathanael Teissier
Amazon
nanx@usc.edu, {sey,aggabhin,zeku,natteis}@amazon.com

Abstract

Deep Neural Networks, despite their success in diverse domains, are provably sensitive to small perturbations which cause the models to return erroneous predictions to minor transformations. Recently, it was proposed that this effect can be addressed in the text domain by optimizing for the worst case loss function over all possible word substitutions within the training examples. However, this approach is prone to weighing semantically unlikely word replacements higher, resulting in accuracy loss. In this paper, we study robustness to adversarial perturbations by using differentially private randomized substitutions while training the model. This approach has two immediate advantages: (1) by ensuring that the word replacement likelihood is weighted by its proximity to the original word in a metric space, we circumvent optimizing for worst case guarantees thereby achieve performance gains; and (2) the calibrated randomness results in training a privacy preserving model, while also guaranteeing robustness against adversarial attacks on the model outputs. Our approach uses a novel density-based differentially private mechanism based on truncated Gumbel noise. This ensures training on substitutions of words in dense and sparse regions of a metric space while maintaining semantic similarity for model robustness. Our experiments on two datasets suggest an improvement of up to 10% on the accuracy metrics.

Deep Neural Networks (DNNs) have found applications within multiple domains: from computer vision (Krizhevsky, Sutskever, and Hinton, 2012), and Natural Language Processing (Mikolov et al., 2013), to robotics (Kober, Bagnell, and Peters, 2013) and self-driving cars (Bojarski et al., 2016). However, DNNs have been shown to be vulnerable to adversarial examples. These are small perturbations of examples that are correctly classified by well-trained models but incorrectly classified in the target (Goodfellow, Shlens, and Szegedy, 2014).

A few approaches have been proposed to defend against such adversarial attacks. One of the most widely used methods is adding the adversarial examples to the original training set and retraining the model. On most kinds of perturbations, such augmented training approach has achieved improved robustness without harming accuracy on the original testing sets (Jia and Liang, 2017; Iyyer et al., 2018; Ribeiro, Singh, and Guestrin, 2018; Belinkov and Bisk, 2017; Ebrahimi et al., 2017). However, this often leads to the augmented neural network over-fitting to the additional data (Matyasko and Chau, 2017), but failing to perform robustly against other types of adversarial examples (Jia and Liang, 2017; Belinkov and Bisk, 2017). Recently, certified defences have been adopted in the computer vision domain (Lecuyer et al., 2019; Dvijotham et al., 2018; Gowal et al., 2018). To defend against perturbations on text data, the Interval Bounded Propagation (IBP) approach was proposed by Jia et al. (2019) to minimize the upper bound on the worst-case loss that word substitutions can induce during the training procedure. However, this approach is prone to weighing semantically unlikely word replacements higher, resulting in accuracy loss. This is due to the fact that the loss optimization is done to cater to the worst case guarantees.

In this paper, we propose a new approach to generate adversarial examples via word substitutions in textual analysis. Our approach is based on randomized mechanisms satisfying Metric Differential Privacy (also known as d_χ-privacy (Andrés et al., 2013)), which is a variant of traditional Differential privacy (DP). DP was proposed by Dwork et al. (2006) and has been established as a *de facto* standard for privacy-preserving data analysis. It mathematically guarantees, given a privacy parameter ϵ, that an adversary observing separate outputs of computations over adjacent databases (described by a Hamming distance) will make essentially the same inference. As opposed to standard DP, with d_χ-privacy, the guarantees are scaled by a (different) distance metric between adjacent databases, and privacy preserving noise is sampled from a distribution such as the multivariate Laplacian. The distances are over a metric space, usually Euclidean, over the space defined by word embeddings such as GloVe Pennington, Socher, and Manning (2014) or fastText Bojanowski et al. (2017), while the data points are vector representations of the words. The mechanism assigns higher substitution probability, based on the noise added, to words closer to the original one than those further away. The private text mechanisms proposed by Fernandes, Dras, and McIver (2019); Feyisetan et al. (2020); and, Xu et al. (2020) work this way.

However, for words with embedding vectors in dense areas, these existing mechanisms fail to distinguish nearer (*i.e.*, more relevant) words from other close but less relevant words. As a result, for a given value of the privacy param-

eter ϵ, an irrelevant word could have a similar substitution probability as a relevant word. We propose a new metric-DP mechanism called the truncated Gumbel perturbation mechanism to allow a smaller range of nearby words to be considered than the multivariate Laplace mechanism. The new mechanism samples a k value from a truncated Poisson distribution as the number of substitution candidates before perturbation, and hence, words nearby with irrelevant meanings are disregarded.

We carry out experiments to investigate the performance of text classification models trained to be robust to adversarial substitutions. Unlike existing work such as Jia and Liang (2017), the input text is perturbed by a metric DP mechanism with varying noise levels corresponding to different degrees of semantic preservation. This helps us attain both adversarial robustness and differentially private guarantees. Furthermore, our results suggest that this approach better preserves word semantics and improves utility of models trained on perturbed datasets in downstream tasks. We summarize our main contributions as follows:

- We propose a novel metric DP mechanism called the truncated Gumbel mechanism, which better preserves semantic meanings than the existing multivariate Laplace mechanisms. We formally prove its privacy guarantees and analyze relevant privacy statistics.
- To the best of our knowledge, we are the first to leverage metric DP mechanisms to generate adversarial examples and study the performance of different adversarial training approaches at different noise levels.
- We empirically demonstrate the benefit of the truncated Gumbel mechanism in preserving semantics and show that augmented training performs better than certifiably robust training, both in clean and adversarial accuracy.

Technical Preliminaries. We begin with providing some background on metric DP and the multivariate Laplace mechanism, which has previously been used for privacy-preserving textual analysis. We also provide details on the truncated Gumbel distribution and some other mathematical preliminaries that will be used throughout this paper.

Differential Privacy (DP). First proposed by (Dwork et al., 2006), DP provides a strong mathematical framework for guaranteeing that the output of a randomized mechanism will remain essentially unchanged on any two neighboring input databases. Formally, a randomized mechanism $M : \mathcal{X} \to \mathcal{Y}$ satisfies (ϵ, δ)-DP if for any $x, x' \in \mathcal{X}$ that differ in only one entry, then it holds for all $Y \subseteq \mathcal{Y}$ that:

$$\Pr[M(x) \in Y] \leq e^{\epsilon} \Pr[M(x') \in Y] + \delta, \tag{1}$$

where $\epsilon > 0$ and $\delta \in [0, 1]$ are parameters that quantify the strength of the privacy guarantee. If $\delta = 0$, we say that the mechanism M is ϵ-DP.

This definition can be generalized to other metrics for capturing dataset proximity depending on the application, e.g., the Manhattan distance metric used to provide indistinguishability if the individual's registration date differs at most 5 days in two databases, and the Euclidean distance on the 2-dimensional space used to preserve the user's longitude and latitude information (Chatzikokolakis, Palamidessi, and Stronati, 2015). In particular, for text data, we adopt metric Differential Privacy (a.k.a. d_χ-privacy), following (Chatzikokolakis et al., 2013; Fernandes, Dras, and McIver, 2019; Feyisetan et al., 2020; Xu et al., 2020). In this framework, we ensure that for all $y \in \mathcal{Y}$, it holds that:

$$\Pr[M(x) = y] \leq e^{\epsilon d(x,x')} \Pr[M(x') = y], \tag{2}$$

where the metric $d(x, x') = \|\phi(x) - \phi(x')\|$ describes the Euclidean distance of the word representations for x, x' in some semantic embedding space like GLOVE (Pennington, Socher, and Manning, 2014). Under this definition, the likelihood of a similar output from the mechanism is weighted in proportion to distance of the word being substituted.

Multivariate Laplace Mechanism. A popular approach for achieving metric-DP is to use a multivariate Laplace Mechanism for high-dimensional data (Wu et al., 2017; Feyisetan et al., 2020). Given the embedding vector $\phi(x) \in \mathcal{R}^n$ for each word in the vocabulary, an n-dimensional noise κ is sampled following the distribution $p(\kappa) \propto \exp(-\epsilon \|\kappa\|)$. This variate is obtained by first sampling a uniform vector in the n-dimensional unit ball and scaling it using a Gamma variate sampled from $\Gamma(n, 1/\epsilon)$. The perturbed word x' is the nearest word to $\phi(x) + \kappa$ in the embedding space.

Truncated Poisson Sampling. The mechanism we define in this paper uses random variates sampled from a Poisson distribution, but truncated in value if it gets too large. Let $\lambda > 0$ be a real and a, b be two integers with $1 \leq a < b$. We say that a random variable X follows a $\mathsf{TruncatedPoisson}\,(\lambda; a, b)$ distribution if the following holds:

$$\Pr(X = k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{if } a \leq k < b \\ 1 - \sum_{k=a}^{b-1} \frac{e^{-\lambda}\lambda^k}{k!} & \text{if } k = b \\ 0 & \text{otherwise.} \end{cases}$$

Truncated Gumbel Distribution. Our mechanism uses random variables sampled from the truncated Gumbel distribution for location parameter $\mu \in \mathbb{R}$ and scale parameter $\beta > 0$, with the density function proportional to:

$$\mathsf{TruncatedGumbel}(x; \mu, \beta, C) \propto \exp\left(-\frac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}\right),$$

for all $x \in [-C, C]$, where $C > 0$ is a constant. The distribution has no support in the interval $[-C, C]$. We write $X \sim \mathsf{TruncatedGumbel}(0, b, c)$ to denote a truncated Gumbel distributed random variable with $\mu = 0$, $\beta = b$, and $C = c$. Further, if $C = \infty$, the truncated Gumbel distribution reduces to the Gumbel distribution, which we write as $\mathsf{Gumbel}(0, b)$. Samples from $\mathsf{TruncatedGumbel}(x; \mu, \beta, C)$ can be obtained using standard rejection sampling.

The Truncated Gumbel Mechanism

Motivated by the approach proposed by (Durfee and Rogers, 2019), our density-aware word substitution mechanism uses a truncated Gumbel random variable for selecting amongst a list of candidate perturbations (see Algorithm 1. We provide an overview of the main steps involved in our algorithm. Due to space constraints, we refer the reader to the Appendix for missing details in this section.

Algorithm 1: Truncated Gumbel Perturbation Mechanism

Input : String $x = w_1 w_2 \dots w_\ell \in \mathcal{W}^\ell$, privacy parameter ϵ, word set $\mathcal{W}$.

1 Let $\Delta = \max_{w,w' \in \mathcal{W}} \|\phi(w) - \phi(w')\|_2$ be the maximum inter-word distance, and $\Delta_0 = \min_{\substack{w,w' \in \mathcal{W} \\ w \neq w'}} \|\phi(w) - \phi(w')\|_2$ be the minimum inter-word distance in the embedding space.

2 Set $b = \frac{2\Delta}{\min\{W(2\alpha\Delta),\, \log_e(\alpha\Delta_0)\}}$, where $\alpha = \frac{1}{3}\left(\epsilon - \frac{2(1+\log|\mathcal{W}|)}{\Delta_0}\right)$ and W denotes the principal branch of the Lambert-W function.

3 Initialize an empty string $\tilde{x}$.

4 **for** $w_i \in x$ **do**

5 Sample $k \sim \mathsf{TruncatedPoisson}\,(\log|\mathcal{W}|; 1, |\mathcal{W}|)$ and find the top k closest words to w_i as $\mathbf{u} = [u_1, \dots, u_k]$, where $u_1 = w_i$.

6 Compute the distances $\mathbf{d} = [d_1, d_2, \dots d_j, \dots, d_k]$, where $d_j = ||w_i - u_j||_2$.

7 Set $\widetilde{w}_i = u_j$, where $j = \arg\min\{d_1 + g_1, d_2 + g_2, \dots, d_k + g_k\}$ and $g_1, \dots, g_k \sim_{i.i.d.} \mathsf{TruncatedGumbel}(0, b, \Delta)$.

8 Add $\widetilde{w}_i$ to $\tilde{x}$.

9 **end**

10 **Return** $\tilde{x}$.

For each sentence in the database, we independently perturb each word using the following two steps. First, we randomly select k nearest neighbors of the original word using a truncated Poisson variable, with support over the entire vocabulary (see Step 5). This is done to ensure plausible deniability in our algorithm, for which the support of the substitution mechanism must include all the words in the vocabulary. Ideally, limiting the set of candidate substitutions to only the semantically similar words is necessary to maintain utility. Our approach addresses this trade-off efficiently, by setting the mean number of candidates to the natural logarithm of the vocabulary size. This ensures that the number of candidate substitutions is neither too small, nor too large.

Next, we select the closest $k-1$ words to the original word (using a nearest neighbor search in the embedding space) and compute their distances to the original word (see Step 6). A random choice over this set as follows: the distances are first noised with i.i.d. truncated Gumbel distributed random variables, and then, the smallest noised distance determines the new word (see Step 7). The noise is scaled using the privacy parameter ϵ, the diameter Δ and the minimum inter-word distance Δ_0 of the embedding space, and then clipped using a truncation parameter $C > 0$. We set $C = \Delta$ in our algorithm to ensure that the noised distances are not larger than the inter-word distance, which helps bound the sensitivity of our substitution mechanism. The process is repeated independently for each word in the input string. We formally show that the mechanism described in Algorithm 1 satisfies metric-DP.

Theorem 1. *The Truncated Gumbel mechanism, defined in Algorithm 1, is ϵd_χ-private with respect to the Euclidean metric, for any given privacy parameter $\epsilon > 0$.*

Empirical Evaluation

We now give an overview of approaches discussed in this paper. Given text input $x \in \mathcal{X}$, we consider classification tasks where a model $f(x;\theta)$, parametrized by θ, should predict a label $y \in \mathcal{Y}$. For sentiment classification tasks, the input x is composed of a string of l words $x_1, x_2, \cdots, x_l$ and labelled by one of the two classes $y \in \{1, -1\}$, where the positive sentiment is denoted by 1 while the negative by -1. For textual entailment tasks, two texts are given, one is the premise x and the other is the hypothesis x', and a label is provided based on the relationship between the two: $y \in \{0, 1, 2\}$ denoting the entailment, contradiction or neutral relationship, respectively. Performance of the classification model is evaluated by the percentile of correct predictions inferred on the testing set: $\sum_{x_i \in \mathcal{D}_{\text{test}}} \mathbb{1}(f(x_i;\theta) = y_i)/|\mathcal{D}_{\text{test}}|$, where $\mathbb{1}$ is an indicator function equal to 1 if the predicted label $f(x_i;\theta)$ is identical to the ground-truth y_i, 0 otherwise; $|\mathcal{D}_{\text{test}}|$ represents the size of the test set.

Adversarial Attacks by Word Substitutions. We evaluate the performance of existing certifiably robust trained models when perturbed texts are provided as inputs. Formally, a word-level perturbation is obtained by substituting a given word x_i by another word $\widetilde{x}_i$ in a way that the semantic similarity between the two is determined by the leveraged metric DP mechanism. To achieve this, the additive noise is parametrized by the privacy parameter ϵ: a larger value of ϵ corresponds to less noise, and vice versa.

For the multivariate Laplace Mechanism of (Feyisetan et al., 2020), since the noise is scaled purely as a function of the distance from the original word, when ϵ is small, words in the dense regions of the embedding space are prone to getting substituted with dissimilar words (that are further away), compared to the words in the sparse region. This is because in areas where embedding vectors are densely located, the distance between two irrelevant words is commensurate to that between two words with similar meanings in a sparse region. Hence, adapting the word-level substitution to variations in the density of the embedding space can help boost the utility of models trained on perturbed datasets. To do this efficiently (and without any expensive computation of local sensitivity each time a substitution is made), we propose a novel mechanism based on a truncated Gumbel distribution and prove that it admits metric DP. Instead of sampling based on the distance from the original word, this approach samples k candidate substitutions following the Truncated Poisson distribution and then makes a distance-based calibrated random choice from the $k-1$-nearest neighbors of the original word in the embedding space (see Algorithm 1). We describe this mechanism in more detail and prove its formal privacy guarantees in the Appendix.

Dataset	IMDb	SNLI
Task type	binary	three-class
Training set size	20,000	550,152
Testing set size	1000	10,000
Total word count	11,856,015	4,614,822
Vocabulary size	145,901	49,895
Sentence length	263.46±195.29	8.25±3.20

Table 1: Summary of dataset properties.

Learning with Adversarial Examples. Motivated by the success of augmented training approaches when text perturbations happen in the form of extraneous text insertion (Jia and Liang, 2017), paraphrasing (Iyyer et al., 2018; Ribeiro, Singh, and Guestrin, 2018), character-level noise (Belinkov and Bisk, 2017; Ebrahimi et al., 2017), we also investigate the effectiveness of adding adversarial examples generated by metric DP mechanisms to the training set for retraining. Retaining the label of each sample, we perturb the text four times, during which every word is perturbed by either the existing multivariate Laplace Mechanism or the proposed truncated Gumbel Mechanism.

Experimental Results. We evaluate the proposed privacy mechanism, adversarial attacks and the defense approach by aiming to answer the following:

How will different adversarial training approaches, i.e., IBP with certified robustness, and the proposed augmented training, perform when testing on adversarial examples derived from metric-DP mechanisms?

Tasks and Datasets: We evaluate the robustness of models on two text classification tasks: sentiment analysis on the IMDb movie review dataset from (Maas et al., 2011); and textual entailment on premise-hypothesis relation dataset SNLI (Bowman et al., 2015). We use 300-dimensional GLOVE vectors for word embedding. The statistics of the two datasets are listed in Table. 1.

Sentiment Analysis: In IMDb, each movie review has a positive or negative label. We implemented the CNN architecture that achieved the best adversarial attack and certified accuracy in (Jia et al., 2019).

Textual Entailment: In SNLI, each sample is composed of two sentences: one as the premise and the other as the hypothesis. The classification task is to define the relationship as an entailment, contradiction, or neutral. Following the implementation in (Alzantot et al., 2018), only words in hypothesis are allowed to be substituted. Similarly, we adopted the architecture that outperformed others in (Jia et al., 2019) for evaluating different adversarial training approaches.

Compared Approaches. We compare robustness of the following two training approaches when adversarial examples are generated using metric-DP perturbation.

Certifiably Robust Trained Approach: Interval Bound Propagation (IBP) was leveraged to minimize the upper bound on the worst-case loss that any combination of word substitutions can induce. Specifically, an upper and lower bound on the activation of a neuron in each layer is computed based on the bounds of neurons in previous layers that connect to it. Bounds for the input layer is computed based on the smallest axis-aligned box that contains all the possible word substitutions, while the upper bound on the loss in the final layer is combined with the normal cross entropy loss to optimize the classification performance on the actual word and any other substitutions. The allowed substitutions are based on (Alzantot et al., 2018).

Augmented Training: We add the adversarial examples generated by metric DP mechanisms (perturbing each sample four times) into the training set and retrain the model.

Adversarial Attack Methodology Following (Alzantot et al., 2018), a population-based genetic attacker is implemented to search for perturbations that lead to misclassification from the model. Given an original or modified sentence, the attacker randomly substitutes a word from the sentence with a new one based on the perturbation mechanism satisfying metric DP. After multiple substitutions, the attacker obtains a population of new sentences together with their fitness scores (negatively proportional to the probability predicted for the correct label). If the new sentence with the highest fitness score successfully fools the model, then the attacker moves forward to the next sentence and starts a new round of testing. Otherwise, the attacker will perform crossover and mutation operations: sample two new sentences as parents from the population according to their fitness score, and then generate the child sentence by taking the word from either parent randomly. Another round of perturbation over the child sentence is then performed to further increase sentence diversity. The model is certified robust to after providing correct predictions over a predefined numbers of attacks.

Evaluation Metrics Based on attributes of the testing set, the following two metrics are evaluated: Clean Accuracy: the percentage of correct predictions when testing on the original samples; and, Adversarial Accuracy: percent of correct predictions when testing on perturbed samples.

Privacy Parameter. To make the correlation between privacy and noise more intuitive, and to aid comparison between the Laplace and Truncated Gumbel mechanism, we opted to surface the privacy parameter as a *noise scale*. The larger the noise scale, the stronger the privacy guarantees (as opposed to the inverse relation between privacy and ϵ). For more details on how the ϵ parameters for both Laplace and Truncated Gumbel map to the *noise scale*, see the Appendix.

Model Robustness Against Adversarial Samples. We list performance of the two adversarial training approaches when samples are perturbed by the multivariate Laplace mechanism in Table 3 and the truncated Gumbel mechanism in Table 2. In Table 3, clean accuracy of the proposed augmented training approach is approximately 8.74% higher than that of the certifiably robust trained approach IBP for any noise scale selection on IMDb and 3.33% higher when the noise injected is < 0.05 on SNLI. Retraining with adversarial examples helps maintain the similar level of clean accuracy as the normal training approach, which is consistent with observations in literature (Jia and Liang, 2017; Iyyer et al., 2018; Ribeiro, Singh, and Guestrin, 2018; Belinkov and Bisk, 2017; Ebrahimi et al., 2017). When evaluating the

Table 2: Performance of adversarial training approaches using the **Truncated Gumbel Perturbation Mechanism**.

Noise Scale			0.574	0.341	0.262	0.181	0.102	0.075	0.06	0.049	0.042
IMDb	Clean	IBP	81.00	81.00	81.00	81.00	81.00	81.00	81.00	81.00	81.00
		Aug	**89.60**	**88.10**	**90.00**	**88.30**	**89.20**	**89.00**	**89.40**	**89.80**	**89.70**
	Adv	IBP	34.60	**47.40**	58.60	70.90	79.90	80.80	80.90	80.90	81.00
		Aug	**34.90**	43.30	**60.20**	**71.80**	**86.20**	**88.80**	**89.30**	**89.70**	**89.70**
SNLI	Clean	IBP	79.19	79.19	79.19	79.19	79.19	79.19	79.19	79.19	79.19
		Aug	**79.92**	**81.32**	**81.74**	**81.77**	**82.20**	**82.18**	**81.86**	**81.65**	**81.96**
	Adv	IBP	11.49	12.98	14.95	**24.01**	**58.78**	74.51	78.18	78.88	79.12
		Aug	**17.34**	**16.57**	**17.05**	23.96	58.58	**76.54**	**80.62**	**81.41**	**81.90**

Table 3: Performance of adversarial training approaches using the **Multi-variate Laplace Mechanism**.

Noise Scale			1	0.2	0.11	0.05	0.025	0.0167	0.0125	0.01	0.005
IMDB	Clean	IBP	81.00	81.00	81.00	81.00	81.00	81.00	81.00	81.00	81.00
		Aug	**88.22**	**88.20**	**87.34**	**87.38**	**88.60**	**88.74**	**88.12**	**88.46**	**87.76**
	Adv	IBP	0.30	0.50	1.20	4.90	**38.60**	**68.30**	**78.50**	**80.30**	81.00
		Aug	**10.80**	**8.50**	**10.20**	**6.90**	9.50	17.70	32.10	53.00	**88.30**
SNLI	Clean	IBP	**79.19**	**79.19**	**79.19**	**79.19**	79.19	79.19	79.19	79.19	79.19
		Aug	76.68	77.28	77.07	78.08	**81.38**	**81.79**	**81.75**	**81.91**	**82.00**
	Adv	IBP	1.84	1.90	2.21	3.70	**9.22**	**24.19**	46.62	64.92	79.16
		Aug	**2.44**	**2.61**	**3.01**	**4.20**	9.14	24.08	**46.94**	**66.54**	**81.94**

model's robustness against word perturbations from the multivariate Laplace mechanism, the augmented training outperforms the IBP approach only when the noise scale value is smaller than some threshold, e.g., < 0.01 on IMDb and < 0.025 on SNLI. This is expected as the augmented training cannot protect against all attacks especially when large amounts of noise results in any word substitution without considering semantic-preserving. In this case, the model can hardly learn the hidden relationship between the corrupted new texts and the original text label.

Given better semantic-preserving capability inherent in the proposed truncated Gumbel mechanism, the augmented training approach outperforms the certifiably robust trained IBP method in both clean and adversarial accuracy almost for any tested noise scale value tested. In Table 2, improvement of clean accuracy by the augmented training approach over IBP is 9.87% on IMDb and 3.77% on SNLI when the noise scale is 0.075. At the same time, better performance against adversarial attacks is achieved by the augmented training approach: 9.90% higher adversarial accuracy on IMDb and 2.72% on SNLI.

One possible explanation of the inferior adversarial accuracy achieved by the certified defense approach IBP may be attributed to the training procedure, which is based on the word substitutions that preserve semantic meanings (Alzantot et al., 2018). However, the testing adversarial examples are generated by randomized perturbations from metric DP mechanisms, where the semantic meaning is not always preserved, but dynamically determined by the noise scale.

Discussion and Conclusion. We study the performance of different adversarial training approaches against adversarial examples generated by metric DP mechanisms. To better preserve semantic meanings during word perturbations, we propose a novel Truncated Gumbel mechanism, which formally satisfies metric differential privacy. Empirically, our experiments demonstrate the advantage of this mechanism over the multivariate Laplace mechanism due to its smaller range of substitution candidates. In two text classification tasks, retraining with adversarial examples performs better than certified defence in both clean and adversarial accuracy.

We believe the following aspects are interesting for future work: 1) robustness of other adversarial training approaches based on the metric DP-inspired adversarial examples, e.g., surrogate-loss minimization; 2) generalization capability of the well-trained augmented training approach, e.g., performance against other types of adversarial examples; 3) privacy preservation performance of the proposed truncated gumbel mechanism, e.g., performance of membership inference attacks (MIA) on perturbed texts.

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Andrés, M. E.; Bordenabe, N. E.; Chatzikokolakis, K.; and Palamidessi, C. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 901–914.

Belinkov, Y., and Bisk, Y. 2017. Synthetic and natural noise

both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Chatzikokolakis, K.; Andrés, M. E.; Bordenabe, N. E.; and Palamidessi, C. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, 82–102. Springer.

Chatzikokolakis, K.; Palamidessi, C.; and Stronati, M. 2015. Constructing elastic distinguishability metrics for location privacy. *Proceedings on Privacy Enhancing Technologies* 2015(2):156–170.

Durfee, D., and Rogers, R. M. 2019. Practical differentially private top-k selection with pay-what-you-get composition. In *Advances in Neural Information Processing Systems*, 3532–3542.

Dvijotham, K.; Gowal, S.; Stanforth, R.; Arandjelovic, R.; O'Donoghue, B.; Uesato, J.; and Kohli, P. 2018. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

Ebrahimi, J.; Rao, A.; Lowd, D.; and Dou, D. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Fernandes, N.; Dras, M.; and McIver, A. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, 123–148. Springer, Cham.

Feyisetan, O.; Balle, B.; Drake, T.; and Diethe, T. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 178–186.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T.; and Kohli, P. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32(11):1238–1274.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, 656–672. IEEE.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

Matyasko, A., and Chau, L.-P. 2017. Margin maximization for robust classification using deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 300–307. IEEE.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 856–865.

Wu, X.; Li, F.; Kumar, A.; Chaudhuri, K.; Jha, S.; and Naughton, J. 2017. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, 1307–1322.

Xu, Z.; Aggarwal, A.; Feyisetan, O.; and Teissier, N. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. In *Proceedings of the Workshop on PrivateNLP at the 2020 conference on empirical methods in natural language processing (EMNLP)*.

Privacy Proof for Truncated Gumbel Mechanism

Restatement of Theorem 1. *The truncated Gumbel mechanism, defined in Algorithm 1, is ϵd_χ-private with respect to the Euclidean metric, for any given privacy parameter $\epsilon > 0$.*

Proof. We first show for any pairs of substitutable words w and w',

$$\frac{\Pr[M(w) = u_i | K = n]}{\Pr(M(w') = u_i | K = n]} \leq \exp\left[\frac{2}{b} e^{\frac{2}{b}\Delta} d(w, w')\right],$$

where $n = |\mathcal{W}|$ and $d(w, w') = \|\phi(w) - \phi(w')\|_2$. Conditional on $K = n$,

$$\Pr(M(w) = u_i | K = n) = \Pr(d_i + g_i < \min_{j \neq i} d_j + g_j).$$

Since $g_1, \ldots, g_n$ are $i.i.d.$ random variables, we argue for each i independently. Fix $g_{-i} = [g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n]$ as a random draw from $n-1$ independent Gumbel distributions. Define $g^* = \sup g : d_i + g < \min_{j \neq i} d_j + g_j$. Then $g_i < \min_{j \neq i}(d_j + g_j) - d_i$ if and only if $g_i \leq g^*$, which means $M(w) = u_i$ if and only if $g_i \leq g^*$. Now consider another substitutable word w' with a corresponding distance vector $\mathbf{d}' = [d'_1, \ldots, d'_n]$. By triangle inequality, we have

$$|d_i - d'_i| \leq d(w, w'), \text{ for } i = 1, \ldots, n.$$

Therefore,

$$\begin{aligned}
&\Pr(M(w') = u_i | K = n) \\
=&\Pr(d'_i + g_i < \min_{j \neq i}(d'_j + g_j)) \\
=&\Pr(g_i < \min_{j \neq i}(d'_j + g_j) - d'_i) \\
=&\Pr(g_i < \min_{j \neq i}(d_j + g_j) - d_i + 2d(w, w')) \\
=&\Pr(g_i \leq g^* + 2d(w, w')).
\end{aligned}$$

Since g_i follows a Truncated Gumbel distribution, and the normalizing constant gets cancelled out in the probability ratio below, we have

$$\begin{aligned}
&\frac{\Pr(M(w) = u_i | K = n)}{\Pr(M(w') = u_i | K = n)} \\
\geq&\frac{\Pr(g_i \leq g^*)}{\Pr(g_i \leq g^* + 2d(w, w'))} \\
=&\frac{\exp(-e^{-\frac{1}{b}g^*})}{\exp(-e^{-\frac{1}{b}g^* - \frac{2}{b}d(w,w')})} \\
=&\exp[-e^{-\frac{1}{b}g^*}(1 - e^{-\frac{2}{b}d(w,w')})],
\end{aligned}$$

which is increasing in g^* as $1 - e^{-\frac{2}{b}d(w,w')} > 0$. Since $g^* \geq -2\Delta$, and then

$$\begin{aligned}
&\frac{\Pr(M(w) = u_i | K = n)}{\Pr(M(w') = u_i | K = n)} \\
\geq&\exp(-e^{-\frac{1}{b}(-2\Delta)}(1 - e^{-\frac{2}{b}d(w,w')})) \\
\geq&\exp\left[-e^{\frac{2}{b}\Delta} \cdot \frac{2}{b} d(w, w')\right].
\end{aligned}$$

By symmetry of w and w', we also have

$$\frac{\Pr(M(w) = u_i | K = n)}{\Pr(M(w') = u_i | K = n)} \leq \exp\left[\frac{2}{b} e^{\frac{2}{b}\Delta} d(w, w')\right].$$

Recall that $K \sim \mathsf{TruncatedPoisson}\,(\lambda; 1, n)$. We want to show an upper bound for $\frac{\Pr(M(w)=u_i)}{\Pr(M(w')=u_i)}$, which is

$$\begin{aligned}
&\frac{\Pr(M(w) = u_i)}{\Pr(M(w') = u_i)} \\
=&\frac{\sum_{k=1}^n \Pr(M(w) = u_i | K = k) \Pr(K = k)}{\sum_{k=1}^n \Pr(M(w') = u_i | K = k) \Pr(K = k)} \\
\leq&\frac{\sum_{k=1}^n \Pr(M(w) = u_i | K = k) \Pr(K = k)}{\Pr(M(w') = u_i | K = n) \Pr(K = n)} \\
\leq&\frac{n - 1 + \Pr(M(w) = u_i | K = n) \Pr(K = n)}{\Pr(M(w') = u_i | K = n) \Pr(K = n)},
\end{aligned}$$

Since

$$\begin{aligned}
\Pr(M(w) = u_i | K = n) =& \exp(-e^{-\frac{1}{b}g^*}) \\
\geq& \exp(-e^{\frac{2\Delta}{b}}),
\end{aligned}$$

and $\Pr(K = n) \geq e^{-\lambda}$,

$$\begin{aligned}
&\frac{\Pr(M(w) = u_i)}{\Pr(M(w') = u_i)} \\
\leq&\exp\left(\frac{2}{b} e^{\frac{2}{b}\Delta} d(w, w')\right)\left(1 + \frac{n-1}{\exp(-e^{\frac{2\Delta}{b}} - \lambda)}\right) \\
=&\left(1 + (n-1)e^{e^{\frac{2\Delta}{b}} + \lambda}\right)\exp\left(\frac{2}{b} e^{\frac{2}{b}\Delta} d(w, w')\right) \\
\leq&2n\exp(e^{\frac{2\Delta}{b}} + \lambda)\exp\left(\frac{2}{b} e^{\frac{2}{b}\Delta} d(w, w')\right).
\end{aligned}$$

In order to guarantee ϵ d_χ-privacy, we solve for b using

$$e^{\epsilon d(w,w')} \geq 2n\exp(e^{\frac{2\Delta}{b}} + \lambda)\exp\left(\frac{2}{b} e^{\frac{2}{b}\Delta} d(w, w')\right).$$

Taking logarithm on both sides,

$$\epsilon \geq \frac{1}{d(w, w')} \log_e\left(2n\exp(e^{\frac{2\Delta}{b}} + \lambda)\right) + \frac{2}{b} e^{\frac{2}{b}\Delta},$$

so we need to find an upper bound for the right-hand side of the equation as a function of b.

$$\begin{aligned}
&\frac{1}{d(w, w')} \log_e\left(2n\exp(e^{\frac{2\Delta}{b}} + \lambda)\right) + \frac{2}{b} e^{\frac{2}{b}\Delta} \\
\leq&\frac{1}{\Delta_0}\left(2 + \log n + e^{\frac{2\Delta}{b}} + \lambda\right) + \frac{2}{b} e^{\frac{2}{b}\Delta} \\
=&\frac{2 + \log n + \lambda}{\Delta_0} + \left(\frac{1}{\Delta_0} + \frac{2}{b}\right) e^{\frac{2}{b}\Delta},
\end{aligned}$$

which is decreasing in b. When $b \leq \Delta_0$,

$$\begin{aligned}
&\frac{2 + \log n + \lambda}{\Delta_0} + \left(\frac{1}{\Delta_0} + \frac{2}{b}\right) e^{\frac{2}{b}\Delta} \\
\leq&\frac{2 + \log n + \lambda}{\Delta_0} + \frac{3}{b} e^{\frac{2}{b}\Delta},
\end{aligned}$$

it is sufficient to set

$$b = \frac{2\Delta}{W\left(\frac{2\Delta}{3}\left(\epsilon - \frac{2+\log n+\lambda}{\Delta_0}\right)\right)},$$

where W is Lambert-W function. When $b > \Delta_0$,

$$\frac{2+\log n+\lambda}{\Delta_0} + \left(\frac{1}{\Delta_0} + \frac{2}{b}\right)e^{\frac{2}{b}\Delta} \leq \frac{2+\log n+\lambda}{\Delta_0} + \frac{3}{\Delta_0}e^{\frac{2}{b}\Delta},$$

it is sufficient to set

$$b = \frac{2\Delta}{\log_e\left(\frac{\Delta_0}{3}\left(\epsilon - \frac{2+\log n+\lambda}{\Delta_0}\right)\right)}.$$

Thus, a sufficient condition for

$$\epsilon \geq \frac{1}{d(w,w')}\log_e\left(2n\exp(e^{\frac{2\Delta}{b}}+\lambda)\right) + \frac{2}{b}e^{\frac{2}{b}\Delta},$$

is to set b to be

$$\max\left(\frac{2\Delta}{W\left(\frac{2\Delta}{3}\left(\epsilon - \frac{2+\log n+\lambda}{\Delta_0}\right)\right)}, \frac{2\Delta}{\log_e\left(\frac{\Delta_0}{3}\left(\epsilon - \frac{2+\log n+\lambda}{\Delta_0}\right)\right)}\right).$$

Now that we have proved the proposed mechanism M is $\epsilon\ d_\chi$-private with respect to Euclidean metric d on a string of one word, we have for any pair of inputs $w, w' \in \mathcal{W}^\ell$ and any output $u \in \mathcal{W}^\ell$,

$$\frac{\Pr(M(w)=u)}{\Pr(M(w')=u)} = \prod_{i=1}^{\ell}\left(\frac{\Pr(M(w_i)=u_i)}{\Pr(M(w_i')=u_i)}\right) \leq \prod_{i=1}^{\ell}\exp(\epsilon d(w_i,w_i')) = \exp(\epsilon d(w,w')),$$

where $d(w,w') = \sum_{i=1}^{\ell} d(w_i,w_i')$. □

For Algorithm 1, we set $\lambda = \log|\mathcal{W}|$, so that the value of b used is the following:

$$b = \max\left(\frac{2\Delta}{W\left(\frac{2\Delta}{3}\left(\epsilon - \frac{2+2\log|\mathcal{W}|}{\Delta_0}\right)\right)}, \frac{2\Delta}{\log_e\left(\frac{\Delta_0}{3}\left(\epsilon - \frac{2+2\log|\mathcal{W}|}{\Delta_0}\right)\right)}\right)$$

For this value of b to be defined, we must ensure that ϵ is set in a way that the logarithm and Lambert-W function in the denominator has a positive argument. This holds whenever ϵ is larger than $2\left(1+\log|\mathcal{W}|\right)/\Delta_0$. However, since the nearest neighbor search in the embedding space is scale-invariant, we can scale Δ_0 to make ϵ arbitrarily small.

Fraction of Modified Words

Lemma 1. *For given $\epsilon > 0$, string $x = w_1 \dots w_\ell$ and any fixed k, the expected fraction of words that get modified using Algorithm 1 is at least $(1-p)$, where $p = \exp\left(-e^{-\frac{2\Delta}{b}}\right)$. In particular, $\mathbb{E}(N_w) \leq p|\mathcal{W}|$.*

Proof. Fix a word $w_i \in x$. Since $u_1 = w_i$, observe that we can write the probability that it does not get modified as $\Pr(\widetilde{w_i} = u_1) = \Pr\left(g_1 < \min_{j\geq 2}(d_j + g_j)\right)$. Let $g_1^* = \sup g : g < \min_{j\geq 2}(d_j+g_j)$. Then, similar to the proof of Theorem 1, $g_1 < \min_{j\geq 2}(d_j+g_j)$ if and only if $g_1 \leq g_1^*$. This gives $\Pr(\widetilde{w_i} = u_1) = \Pr(g_1 \leq g_1^*) = \exp\left(-e^{-g_1^*/b}\right)$. Since $g_1^* \leq 2\Delta$, we can write $\Pr(\widetilde{w_i} = u_1) \leq \exp\left(-e^{-\frac{2\Delta}{b}}\right)$.

Thus, the expected fraction of words in x that do not get modified is at most p, where $p = \exp\left(-\exp\left(-\frac{2\Delta}{b}\right)\right)$. From this, we compute the expected fraction of words that get modified as at least $(1-p)$, as desired. The bound on $\mathbb{E}(N_w)$ follows from a simple union bound over all the words in the vocabulary. □

Note that $\frac{\partial p}{\partial b} = \frac{\partial}{\partial b}\exp\left(-e^{-\frac{2\Delta}{b}}\right) < 0$, and hence, p is a decreasing function in b, implying that as the privacy increases (b increases), the value of $\mathbb{E}(N_w)$ decreases, as expected.

Utility Analysis vs. Sparsity of the Embedding Space

For a fixedd ϵ, the Truncated Gumbel mechanism keeps w unchanged when the noise added to w is smaller than any other perturbed candidate. If $p_{\mathsf{Gum}}(w)$ is the probability that w does not change under this perturbation, then we can write the following:

$$\begin{aligned} p_{\mathsf{Gum}}(w) &\geq \Pr\left(g_1 < \delta(w) + g_2\right)\Pr(K \geq 2) \\ &= \Pr\left(g_1 - g_2 < \delta(w)\right)\Pr(K \geq 2) \end{aligned}$$

Since the difference of two i.i.d. Gumbel random variables follows a Logistic distribution, we obtain the following (by letting $G_b \sim \mathsf{Logistic}(0,b)$):

$$\begin{aligned} p_{\mathsf{Gum}}(w) &\geq \Pr\left(G_b < \delta(w)\right)\Pr(K \geq 2) \\ &= \left(\frac{1}{1+e^{-\delta(w)/b}}\right)\Pr(K \geq 2) \\ &\geq e^{-e^{-\delta(w)/b}}\Pr(K \geq 2), \end{aligned}$$

where, the last inequality follows since $1 + x \leq e^x$. Thus, even when $\delta(w)$ approaches 0 (denser regions), there is at least $p_{\mathsf{Gum}}(w)|_{\delta(w)\to 0} \geq \frac{\Pr(K\geq 2)}{e} = \frac{1}{e}\left(1 - \frac{\log|\mathcal{W}|}{e^{|\mathcal{W}|}}\right) \xrightarrow{|\mathcal{W}|\to\infty} 36.7\%$ probability that w remains unchanged. This helps preserve utility by ensuring that the modified word is likely to be closer to the original word since there is a significant probability mass around the original word (specially as $|\mathcal{W}|$ increases).