Probabilistic Categorical Adversarial Attack and Adversarial Training

Han Xu [1] **Pengfei He** [1] **Jie Ren** [1] **Yuxuan Wan** [1] **Zitao Liu** [2] **Hui Liu** [1] **Jiliang Tang** [1]

Abstract

The studies on adversarial attacks and defenses have greatly improved the robustness of Deep Neural Networks (DNNs). Most advanced approaches have been overwhelmingly designed for continuous data such as images. However, these achievements are still hard to be generalized to categorical data. To bridge this gap, we propose a novel framework, *Probabilistic Categorical Adversarial Attack (or PCAA)*. It transfers the discrete optimization problem of finding categorical adversarial examples to a continuous problem that can be solved via gradient-based methods. We analyze the optimality (attack success rate) and time complexity of PCAA to demonstrate its significant advantage over current search-based attacks. More importantly, through extensive empirical studies, we demonstrate that the well-established defenses for continuous data, such as adversarial training and TRADES, can be easily accommodated to defend DNNs for categorical data.

1. Introduction

Adversarial examples (Goodfellow et al., 2015) have raised great concerns for the applications of Deep Neural Networks (DNNs) in many security-critical domains (Cui et al., 2019; Stringhini et al., 2010; Cao & Tay, 2001). Recent years have witnessed an increasing number of adversarial attack and defense methods (Goodfellow et al., 2015; Madry et al., 2018; Ilyas et al., 2019). These studies have not only greatly deepened our understanding on the vulnerabilities of DNNs but also tremendously advanced the robustness of DNNs. Until now, the majority of existing accomplishments have been achieved in continuous data such as images, where gradient-based approaches can be leveraged. However, there are also many machine learning tasks where the input data is categorical. For example, data in ML-based intrusion detection systems (Khraisat et al., 2019) contains records of the type of system operations; in financial transaction systems, data includes categorical features such as the region and card information of transactions; and in NLP tasks, the words in a sentence can only be chosen from a given vocabulary, which is categorical. To generate categorical adversarial examples, there are recent search-based approaches such as (Yang et al., 2020; Lei et al., 2019; Bao et al., 2022). For example, the method (Yang et al., 2020) first finds top-K features of a given sample that have the maximal influence on the model output, and then a greedy search is applied to obtain the optimal perturbation in these K features. However, these search-based attack methods usually suffer from a poor trade-off between efficiency and optimality (attack success rate) to find strong adversarial examples. Moreover, if these attack methods are applied in defenses like adversarial training (Madry et al., 2018), they can only search for adversarial examples for each training sample at each time, instead of efficiently producing adversarial examples in batches. In a nutshell, these drawbacks of existing categorical attack methods dramatically prohibit the possibility of applying recent advances of attack and defense established in continuous data to categorical data.

Therefore, a natural question is *can we generalize the well-studied methods of continuous data to categorical data?* We face tremendous challenges to answer this question. First, the input data space are categorical, thus the gradient methods of adversarial attacks and defenses (Goodfellow et al., 2015; Madry et al., 2018) for continuous data are not directly applicable. Second, most attacks for categorical data desire to constrain the number of perturbed features (Yang et al., 2020; Bao et al., 2022), which is different from the commonly considered l_2, l_∞ norm constraints in continuous data. To address these challenges, we propose a novel framework: **P**robabilistic **C**ategorical **A**dversarial **A**ttack (PCAA). Overall, it transforms the discrete optimization problem of finding categorical adversarial examples into a continuous problem by estimating the probabilistic distribution of categorical adversarial examples. In detail, given a clean data sample, we assume that (each feature of) its adversarial example follows a categorical distribution, and satisfies: **(1)** the samples following this distribution have

[1] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA [2] Guangdong Institute of Smart Education, Jinan University, Guangzhou, China. Correspondence to: Zitao Liu <liuzitao@jnu.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

a high expected loss value and **(2)** the samples only have a few features which are different from the original clean sample (with high probability). Based on this property, we are able to leverage existing gradient-based algorithms from continuous data such as (Goodfellow et al., 2015; Madry et al., 2018) to figure out the adversarial examples for categorical data. In such a way, we can successfully obtain the categorical adversarial examples by optimizing the adversarial distribution and taking samples from this distribution. Meanwhile, our attack can also be easily incorporated with existing powerful defenses for continuous data such as adversarial training (Madry et al., 2018) and TRADES (Tu et al., 2019). Empirically, we verify that our attack can achieve a better optimality vs. efficiency trade-off to find strong adversarial examples, and demonstrate the advantages of our defense over other categorical defenses. To summarize, the major contributions of this paper are:

- We propose an efficient and effective framework PCAA to bridge the gap between categorical data and continuous data, which allows us to generate adversarial examples for categorical data by leveraging methods from continuous data.
- Equipped with PCAA , existing defenses in continuous data, such as adversarial training and TRADES, can be easily adapted to categorical data. This contribution enables us to generalize new advances in defenses from continuous data to categorical data.
- We empirically validate the great benefit of PCAA from perspectives of both attack and defense.

2. Related Work

In this section, we provide a brief review of existing methodologies of adversarial attacks and defenses for continuous and categorical data, which highlights the gap between the major methodologies in these two types of data.

2.1. Attacks and Defenses on Continuous Data

The adversarial attacks and defenses in the image domain have been extensively studied (Madry et al., 2018; Ilyas et al., 2019; Xu et al., 2020). Most frequently studied attack methods such as FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018) and C&W Attack (Carlini & Wagner, 2017) can only be conducted in the continuous domain, since they require calculating the gradients of model outputs on the input data. As countermeasures to resist adversarial examples in the image domain, most defense strategies are also based on the assumption that the input data space is continuous. For example, adversarial training methods (Madry et al., 2018; Zhang et al., 2019) train the DNNs on the adversarial examples generated by PGD. A SOTA certified defense method Randomized Smoothing (Cohen et al., 2019), calculates the certifiable bounds by adding Gaussian noise to the input samples. However, these methods are hard to be directly applicable to categorical data as the input data space is discrete.

2.2. Attacks and Defenses on Categorical Data

To generate adversarial examples for categorical data, most existing methods apply a search-based approach. For example, the works (Yang et al., 2020; Bao et al., 2022; Lei et al., 2019) first find top-K features of a given sample that have the maximal influence on the model output, and then a greedy search or brutal search is applied to obtain the optimal combination of perturbation in these K features. In NLP tasks, there are also many attack methods proposed to find adversarial “sentences” or “articles”, which follow a similar approach as general search-based categorical attacks. For example, Ebrahimi et al. (2017) proposes to search important characters in the text based on the gradient, and then apply greedy search to find the optimal character flipping. Samanta & Mehta (2017) generate the adversarial embedding in the word embedding space, then search for the closest meaningful adversarial example that is legitimate in the text domain. These attack methods are very different from those in continuous domain.

To defend against categorical attacks, most defenses have been proposed for NLP tasks and exclusively rely on the property of word embedding: similar words have a close distance in the embedding space. Miyato et al. (2016); Barham & Feizi (2019) conduct adversarial training on embedding space, where the adversarial examples are within l_2-ball around the embedding of clean samples. Dong et al. (2021) also proposes an adversarial training method - ASCC, which conducts adversarial training in the space which is composed of convex hulls of adversarial word vectors. However, these methods can hardly be applied to defend DNNs for general categorical data beyond NLP tasks. Different from existing defense methods in NLP tasks, in this paper, our proposed defense does not rely on the property of word embedding and achieves a similar defense performance to these NLP defenses. Moreover, our defense can be applied in general categorical ML tasks beyond NLP.

3. Probabilistic Categorical Adversarial Attack

In this section, we first introduce the necessary notations and definitions of our studied problem. Then, we provide the details of our proposed attack framework PCAA.

3.1. Problem Setup

In this work, we consider a classifier f that predicts labels $y \in \mathcal{Y}$ based on categorical inputs $x \in \mathcal{X}$. Each input sample x contains n categorical features, and each feature x_i can be perturbed to take a value from d allowed categories. Namely, we define the space of all allowed perturbed samples of x to be $\mathcal{S}(x)$. Meanwhile, to keep the perturbation "un-noticeable", we follow the setups of existing works (Yang et al., 2020; Bao et al., 2022; Wang et al., 2020), to limit the number of perturbed features, which is the l_0 distance of clean sample x and the adversarial example x'. It is because constraining l_0 distance is most intuitive and has a broad interest in categorical data, e.g. only a few nucleotides are mutated in genetics data. Therefore, we formally define the objective of our considered attack to be: given the budget size ϵ, we aim to find an adversarial example x', which maximizes the model's loss value, while has a small l_0 distance to x:

$$\max_{x' \in \mathcal{S}(x)} \mathcal{L}(f(x'), y) \ \text{ s.t. } \ \|x' - x\|_0 \leq \epsilon. \qquad (1)$$

Notably, the objective in Eq.(1) is general and it can be applied to find adversarial examples in various applications by accommodating the space $\mathcal{S}(x)$. For example, in NLP tasks such as sentiment analysis, we can define the space $\mathcal{S}(x)$ to be the set of sentences where some words in x are changed to their synonyms. In this way, we can keep the semantic meaning of x during attacking. More details about how to get $\mathcal{S}(x)$ in NLP tasks are given in Section 5.2.

3.2. The Objective of PCAA

To solve the problem in Eq.(1), there are existing search-based methods (Lei et al., 2019; Yang et al., 2020) to search the adversarial examples, via either a greedy search method or brutal search method. However, both of these two search methods can suffer from poor optimality vs. efficiency trade-off during attacking. For example, if one conducts a brutal search (Lei et al., 2019) to traverse the whole discrete space, it must cause an extremely high computational cost. Meanwhile, greedy search narrows down the search space so that it usually finds weak adversarial examples (which cannot maximize the loss in Eq.(1)). Moreover, the poor optimality vs. efficiency trade-off will make these attacks impossible to be incorporated to the most studied defense methods (in the continuous domain), such as adversarial training.

To address these problems, we are motivated to leverage gradient-based methods to conduct adversarial attacks in the categorical domain. In general, we first define a ***continuous probabilistic space*** where the adversarial examples are sampled from, and we devise a new objective in Eq.(2) to approximately solve Eq.(1). In specific, following the illustration in Figure 1, we assume that each feature of (adversarial) categorical data x'_i follows a categorical distribution: *Categorical*(π_i), where $\pi_i \in \Pi_i = (0,1)^d$. Each element $\pi_{i,j}$ represents the probability that the feature i takes the category value j. In the remaining of the paper, we will use π_i to denote the categorical distribution *Categorical*(π_i) without the loss of generality. Then, the input sample x's distribution is the joint distribution of all π_i, which is denoted as $\pi = [\pi_0; \pi_1; ...; \pi_n] \in \Pi \subset \mathbb{R}^{n \times d}$.

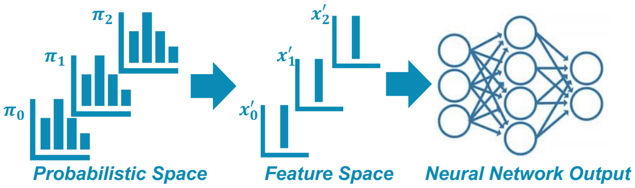


Figure 1. An illustration of PCAA when $n = 3$ and $d = 4$. Each feature x'_i of the adversarial example x' is sampled from probabilistic distribution π_i before feed into the model.

Then, we define a new continuous optimization problem to find a probability distribution π in the space of Π:

$$\max_{\pi \in \Pi} \mathbb{E}_{x' \sim \pi} \mathcal{L}(f(x'), y) \ \text{ s.t. } \ \Pr_{x' \sim \pi}(\|x' - x\|_0 \geq \epsilon) \leq \delta \quad (2)$$

where ϵ denotes the perturbation budget size and δ is the tail probability constraint. By solving the problem in Eq.(2), we aim to find a distribution with parameter π such that: **(1)** ***on average***, the generated samples x' from distribution π have a high loss value; and **(2)** ***with low probability***, the sampled x' has a l_0 distance to the clean sample x larger than ϵ. Thus, the generated samples x' are likely to mislead the model prediction while preserving most features of x. As shown in Figure 1, the probabilistic distribution π to Eq.(2) is first used to sample adversarial examples x', and then, the model makes predictions on the x'.

It worth mentioning that, in our attack, each feature x'_i of the adversarial example x' is sampled independently. However, it does not mean that the features themselves in the data distribution are independent to each other. Therefore, our framework is general and applicable to various data types and model architectures, including sequential data such as sentences in NLP areas, or DNA sequences.

3.3. An Efficient Algorithm of PCAA

Solving the problem in Eq.(2) is not trivial because the probability and the l_0 term are not differentiable. Thus, we provide a feasible algorithm to solve Eq.(2), by substituting the constraint in Eq.(2) to a differentiable term. In detail, we substitute the l_0 distance between x' and x by calculating the sum of Cross Entropy Loss between π_i and x_i, which is $\mathcal{L}_{CE}(\pi_i, x_i)$, for all features $i \in |x|$. It is because $\mathcal{L}_{CE}(\pi_i, x_i)$ measures the probability that the categorical variables x'_i following the distribution π_i is different from x_i. Thus, we use the sum of Cross Entropy $\sum_{i \in |n|} \mathcal{L}_{CE}(\pi_i, x_i)$ to approximate the total number of changed features in x', which is the l_0 difference $\|x' - x\|_0$. In our algorithm, we penalize the searched π when the term $\sum_{i\in|n|}\mathcal{L}_{CE}(\pi_i, x_i)$ exceeds a positive value ζ as:

$$\max_{\pi\in\Pi}\mathbb{E}_{x'\sim\pi}\mathcal{L}(f(x'), y)\ \text{ s.t. }\ \sum_{i\in|n|}\mathcal{L}_{CE}(x_i, \pi_i) - \zeta \leq 0$$

As a result, we equivalently limit the probability that the generated samples x' have the number of perturbed features larger than ϵ. Moreover, since the Cross-Entropy Loss is differentiable in terms of π, we further transform the problem to its Lagrangian form as:

$$\max_{\pi\in\Pi}\left(\mathbb{E}_{x'\sim\pi}\mathcal{L}(f(x'), y)\right) - \lambda\left[\sum_{i\in|n|}\mathcal{L}_{CE}(x_i, \pi_i) - \zeta\right]^+ \tag{3}$$

where λ is the penalty coefficient, and $[\cdot]^+$ is $\max(\cdot, 0)$. Next, we will show how to solve the maximization problem above by applying gradient methods.

Back propagation through Gumbel-Softmax. Note that the gradient of the expected loss function with respect to π cannot be directly calculated in Eq.(3), so we apply the Gumbel-Softmax estimator (Jang et al., 2017). In practice, we consider an unnormalized categorical distribution $\pi_i \in (0, C]^d$, where $C > 0$ is a large constant so that the searching space is sufficiently large. The distribution generates sample vectors x'_i as follows:

$$x'_{ij} = \frac{\exp((\log\pi_{ij} + g_j)/\tau)}{\sum_{j=1}^{d}\exp((\log\pi_{ij} + g_j)/\tau)},\quad \text{for } j = 1, ..., d \tag{4}$$

where g_j denotes i.i.d samples from the Gumbel$(0, 1)$ distribution, and τ is the softmax temperature. This reparameterization process facilitates us to calculate the gradient of the expected loss in terms of π. Therefore, we can derive the estimation of gradients for the expected loss:

$$\begin{aligned}\frac{\partial\mathbb{E}_{x'\sim\pi}\mathcal{L}(f(x'), y)}{\partial\pi} &\approx \frac{\partial}{\partial\pi}\mathbb{E}_g\mathcal{L}(f(x'(\pi, g)), y)\\ = \mathbb{E}_g\left[\frac{\partial\mathcal{L}}{\partial x'}\frac{\partial x'}{\partial\pi}\right] &\approx \frac{1}{n_g}\sum_{i=1}^{n_g}\left[\frac{\partial\mathcal{L}}{\partial x'}\frac{\partial x'(\pi, g_i)}{\partial\pi}\right]\end{aligned} \tag{5}$$

where n_g is the number of i.i.d samples from g. In Eq.(5), the first approximation is from the reparameterization of a sample x'; the second equality comes from exchanging the order of expectation and derivative, and the third approximation is to approximate the expectation of gradients by calculating the average of gradients. Finally, we derive the practical solution to solve Eq.(3), by leveraging the gradient ascent algorithm, such as (Madry et al., 2018). In Algorithm 1, we provide the details of our proposed attack method. Specifically, during each iteration, we first estimate the gradient of expected loss (line 3), and then update the unnormalized distribution π by gradient ascent (line 4). Finally, we clip π back to its domain $(0, C]^d$ (line 5).

Algorithm 1 Probabilistic Categorical Adversarial Attack

```
input Data $\mathcal{D}$, budget $\epsilon$, number of samples $n_g$, penalty
    coefficient $\lambda$, max iteration $I$, learning rate $\gamma$
output Adversarial Distribution $\pi$
1: Initialize distribution $\pi^0$
2: for $t \leq I$ do
3:    Estimate expected gradient using Eq.5:
      $\nabla_\pi\mathbb{E}_\pi\mathcal{L} \approx \frac{1}{n_g}\sum_{i=1}^{n_g}\left[\frac{\partial\mathcal{L}}{\partial x'}\frac{\partial x'(\pi^t, g_i)}{\partial\pi}\right]$
4:    Gradient ascent:
      $\widetilde{\pi}^{t+1} = \pi^t + \gamma\left(\nabla_\pi\mathbb{E}_\pi\mathcal{L} - \lambda\nabla_\pi[\mathcal{L}_{CE}(\pi_t, x) - \zeta]^+\right)$
5:    Clip to $(0, C]^d$: $\pi^{t+1} = \max(\widetilde{\pi}^{t+1}, C)$
6: end for
```

3.4. Time Complexity Analysis

In this subsection, we compare the time complexity of PCAA with four representative attack methods (Lei et al., 2019; Yang et al., 2020). Notably, they are existing search-based methods to find adversarial examples for categorical data. Each of them consists of 2 stages: (1) the first stage is to search the top-K features that are most influential to the model output, which is determined by either manipulating the features and checking the loss change *(loss-guided)* or the gradient scale *(gradient-guided)*; (2) the second stage applies either a *brutal search* or a *greedy search* to find the optimal perturbation on the selected features. In Table 1, we summarize the main stages for different attack methods, and we name them as Search Attack **(SA)**, Greedy Attack **(GA)**, Gradient-guided Search Attack **(GSA)** and Gradient-guided Greedy Attack **(GGA)**. More details can be found in Appendix A.1. In Table 1, we assume that the whole dataset has N data points, each data point has n features, each feature has d categories and the budget of the allowed perturbation is ϵ. In the following time complexity analysis, one feedforward / backpropagate step is considered as one computational unit. Results are summarized in Table 1 where C_1 is a constant related to the number of samples n_g and max iteration I shown in Algorithm 1, and detailed time complexity analysis can be found in Appendix A.1

Table 1. Time complexity analysis.

Method	*Stage 1*	*Stage 2*	*Time complexity*
SA	loss-guided	brutal	$N\cdot\mathcal{O}(nd + d^\epsilon)$
GA	loss-guided	greedy	$N\cdot\mathcal{O}(nd + \epsilon d)$
GSA	gradient-guided	brutal	$N\cdot\mathcal{O}(1 + d^\epsilon)$
GGA	gradient-guided	greedy	$N\cdot\mathcal{O}(1 + \epsilon d)$
PCAA	-	-	$C_1 N\cdot\mathcal{O}(1)$

From the analysis above, SA and GSA suffer from the exponential increase of time complexity when the number of feature categories d and budget size ϵ is increasing. GA and GGA accelerate the algorithm and achieve better time efficiency. However, they can sacrifice the performance as they greatly narrow down the search space (see Table 2 in Section 5.1). In Section 5.1, we further empirically show that PCAA can achieve significantly better optimality than GA and GGA, as well as significantly lower computational cost than SA and GSA. Thanks to the advantage in efficiency and optimality, PCAA can fast generate strong adversarial examples. Moreover, because PCAA is a gradient-based method, the adversarial examples can be generated by batches. As a result, it can be easily incorporated into powerful defenses which are originally designed for continuous data. In the following section, we will present PADVT , as an example to show PCAA 's potential to be applied in popular adversarial defenses such as adversarial training and TRADES.

4. Probabilistic Adversarial Training

In this section, we provide an exemplar case to transfer one representative defense, PGD adversarial training (Madry et al., 2018) to categorical defense. Note that we also extend another effective defense TRADES (Zhang et al., 2019) and the detail is shown in Appendix A.3. It is also worth mentioning that other types of defenses such as certified defenses (Cohen et al., 2019) also have the potential to be transferred to the categorical data via PCAA and we leave this exploration as future investigations.

Based on PGD adversarial training for continuous data, we propose Probabilistic Adversarial Training (PADVT) based on PCAA to train robust models for categorical data. Recalling the formulation of PCAA in Eq.(3), and denoting the parameters for classifier f as θ, the training objective for PADVT is formulated as:

$$\min_{\theta}\left[\max_{\pi}\mathbb{E}_{x'\sim\pi}\left(\mathcal{L}(f(x';\theta),y)-\lambda\left[\sum_{i\in|n|}\mathcal{L}_{CE}(x_i,\pi_i)-\zeta\right]^{+}\right)\right]$$

Since our objective involves a penalty coefficient, we adopt the strategy in (Yurochkin et al., 2020) to update λ during training. Specifically, we adaptively choose λ according to $\mathcal{L}_{CE}(x,\pi)-\zeta$ from the last iteration: when the value is large, we increase λ to strengthen the constraints and vice versa. The implementation of PADVT is illustrated in Algorithm 2. Specifically, it first initializes model parameters (line 1); then during each iteration (from line 2 to line 9), it samples a mini-batch of data (line 3) and obtains an adversarial distribution for each data point through Algorithm 1 (line 4 to 5), afterward n_{adv} adversarial examples are sampled (line 6) and used to update θ through Adam (Kingma & Ba, 2015) (line 8) and penalty coefficient λ (line 9). The process will continue until the training process converges.

Algorithm 2 Probabilistic Adversarial Training (PADVT)

```
input data $\mathcal{D}$, parameters of clean model $\theta$, budget $\epsilon$, parameters of Algorithm 1, $n_{adv}$, initial penalty coefficient $\lambda^0$, penalty coefficient step size $\alpha$, parameters of Adam optimizer, number of iterations $I$
output parameters $\theta$ of the robust model
1: Initialize the network with a pre-trained robust model
2: repeat
3:   Sample mini-batch $B = \{x^1, ..., x^m\}$
4:   for $i = 1, ..., m$ (in parallel) do
5:     Apply Algorithm (1) to $x^i$ to obtain adversarial distribution $\pi^i$
6:     Sample $n_{adv}$ examples $\{x'^i_1, ..., X'^i_{n_{adv}}\}$ from $\pi^i$ using Gumbel Softmax
7:   end for
8:   Update $\theta$ to minimize the average adversarial loss
9:   $\lambda = (\lambda - \alpha(\zeta - \frac{1}{m}\sum_{i\in[m]}\sum_{j\in[n]}\mathcal{L}_{\text{CE}}(x^i_j, \pi^i_j)))^+$
10: until Training converged
```

5. Experiment

In this section, we conduct experiments to validate the effectiveness and efficiency of PCAA and PADVT . In Section 5.1, we demonstrate that PCAA achieves a better balance between attack success rate and time efficiency. In Section 5.2, we empirically validate that PADVT achieves good robustness against categorical attacks. Our code is available at `https://anonymous.4open.science/r/categorical-attack-0B9B`.

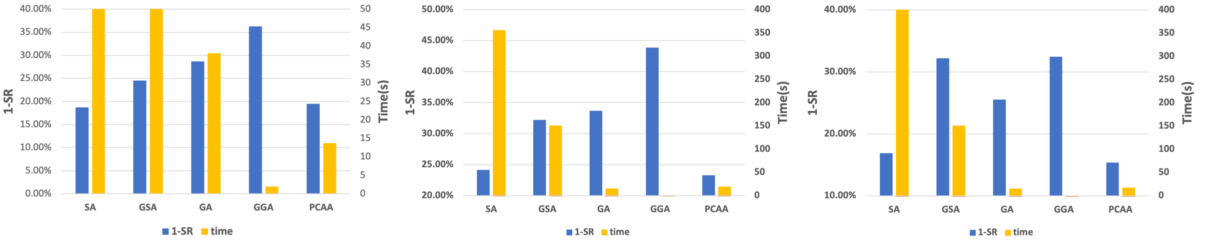
5.1. Categorical Adversarial Attacks

Experimental Setup. In this evaluation, we focus on three categorical datasets for various applications.**(1) Intrusion Prevention System (IPS) (Wang et al., 2020).** IPS dataset has 242,467 instances, with each input consisting of 20 features and each feature has 1,103 categorical values. The output space has three labels. A standard LSTM based classifier(Bao et al., 2022) is trained for IPS dataset. **(2) AG's News corpus.** This dataset consists of titles and description fields of news articles. The tokens of each sentence correspond to the categorical features, and the substitution set (of size 70) corresponds to the categorical values. A character-based CNN(Zhang et al., 2015) is trained on this dataset. **(3) Splice-junction Gene Sequences (Splice) (Noordewier et al., 1990).** Splice dataset has 3,190 instances. Each one is a gene fragment of 60 features with 5 categorical values. The output space has three labels and the model is LSTM.

Baseline Attacks. We compare PCAA with the following search-based attacks including SA, GA, GSA and GGA, which are discussed in Section 3.4. The details of these attacks can also be found in Appendix A.1.

Table 2. Attacking performance on IPS, AG's news, and Splice datasets. "SR." represents the attack success rate; "T." denotes the average running time in seconds; and "-" indicates the running time over 10 hours. Each result runs 5 times, 95% confidence intervals are shown.

Dataset	**Attack Method**	$\epsilon = 1$		$\epsilon = 2$		$\epsilon = 3$		$\epsilon = 4$		$\epsilon = 5$	
		SR.(↑)	T.(↓)	SR.(↑)	T.(↓)	SR.(↑)	T.(↓)	SR.(↑)	T.(↓)	SR.(↑)	T.(↓)
IPS	**SA**	66.11±0.03	38.5	**81.24±0.01**	2028	–	–	–	–	–	–
	GA	66.11±0.03	35.4	71.32±0.02	38.1	79.44±0.06	39.9	85.28±0.07	41.5	91.15±0.11	43.2
	GSA	41.53±0.04	**2.06**	75.47±0.03	1022	–	–	–	–	–	–
	GGA	41.53±0.04	**2.06**	63.72±0.06	**2.01**	70.76±0.05	**2.94**	75.89±0.08	**3.74**	82.43±0.10	**4.56**
	PCAA	**67.56±0.05**	14.04	80.51±0.06	13.75	**88.37±0.07**	13.02	**93.67±0.04**	12.21	**96.63±0.12**	14.59
AG	**SA**	41.22±0.01	15.3	**67.38±0.02**	21.9	75.87±0.04	356	83.10±0.02	19160	–	–
	GA	41.22±0.01	15.3	60.71±0.05	15.4	66.33±0.04	15.5	74.47±0.07	15.7	86.63±0.12	15.9
	GSA	32.39±0.03	**0.352**	59.21±0.02	3.24	67.79±0.03	151	79.22±0.05	7551	–	–
	GGA	32.39±0.03	**0.352**	41.29±0.07	**0.393**	56.11±0.06	**0.511**	67.53±0.10	**0.613**	72.28±0.07	**0.856**
	PCAA	**46.31±0.07**	16.03	67.27±0.08	15.79	**76.71±0.06**	19.21	**84.65±0.09**	17.83	**90.21±0.11**	17.35
Splice	**SA**	**72.11±0.01**	0.905	79.02±0.02	1.02	**86.59±0.01**	1.36	90.11±0.02	2.73	92.58±0.03	8.29
	GA	**72.11±0.01**	0.905	74.42±0.03	0.911	78.18±0.06	0.915	80.61±0.04	0.922	83.74±0.05	0.928
	GSA	61.71±0.02	**0.028**	68.28±0.03	0.083	72.82±0.02	0.251	77.11±0.04	0.917	82.53±0.04	3.56
	GGA	61.71±0.02	**0.028**	65.26±0.08	**0.031**	70.31±0.05	**0.0337**	74.84±0.07	**0.035**	80.49±0.08	**0.037**
	PCAA	72.05±0.03	3.27	**79.33±0.06**	2.82	86.12±0.07	3.18	**90.33±0.06**	2.56	**92.90±0.08**	3.02



(a) IPS with $\epsilon = 2$

(b) AG's news with $\epsilon = 3$

(c) AG's news with $\epsilon = 4$

Figure 2. An illustration of attack success rate and time efficiency trade-offs. The blue bar represents $1 - SR$ under different attacks while the yellow bar denotes the average running time. For both metrics, smaller values indicate stronger attacks.

Implementation details. For each dataset, we evaluate the performance in terms of the attack success rate (SR.) and the average running time (T.) under various budget sizes ϵ ranging from 1 to 5. Remind that in PCAA in Eq.(3), the threshold ζ significantly influences the effectiveness of our method. Therefore, in Table 2, we iteratively conduct PCAA with different choices of ζ from a pre-defined set. Given each ζ, we make 100 samplings from the probabilistic distribution. In this process, once a successful adversarial example (satisfying the l_0 budget constraint) is generated, we claim it to be a successful attack.

Performance Comparison. The experimental results on IPS, AG's news, and Splice datasets are demonstrated in Table 2. The results clearly show that our PCAA reaches the best balance between optimality and efficiency.

(1) PCAA vs. SA / GSA. Both SA and GSA apply brutal search and this leads to terrible efficiency in practice, especially when the budget and the dimension of input space are large (in datasets such as IPS and AG). On IPS dataset, in which each data point consists of more than 1,000 categorical features, when the budget is more than 2, SA and GSA become infeasible leading to their incapability of either practical attack or defense. Compared to SA or GSA, the time complexity of PCAA does not increase with the increase of perturbation budget ϵ. Moreover, PCAA always has a better (or at least a similar) successful attack rate than SA or GSA.

(2) PCAA vs. GA / GGA. These methods accelerate the search process(second stage) by leveraging greedy algorithms. They achieve good efficiency on all 3 datasets, especially on datasets with a small number of categorical features such as the Splice dataset. However, they sacrifice the performance significantly and usually obtain a success rate of 10% less than other methods. The lack of optimality prevents these methods from generating practical attacks and further usage for defenses. Our PCAA does not have this concern as it outperforms them by significant margins, e.g., over 12% higher than GA/GGA in success rate, while still having an acceptable running time.

To further demonstrate that PCAA achieves a better balance, we visualize results on some datasets with different budgets based on Table 2 in Figure 2, where we plot

($1-$ **success rate)** and **time** for each method. From the figure, all baseline methods have high levels on at least one metric, while PCAA can keep both metrics under low levels. Therefore, PCAA can overcome the drawbacks of baseline methods and achieve better efficiency vs. optimality trade-off.

5.2. Categorical Adversarial Defenses

Experimental Setup. For the defense evaluation, we focus on two datasets, AG's News Corpus and IMDB. It is because IPS and Splice have too few samples (no more than 1,000 for each dataset). In particular: **(1) AG's News corpus.** is the same dataset used in the attack evaluation and the model is also a character-based CNN. As character swapping does not require embeddings for each character, we can directly apply attacking methods on input space. Therefore, the robustness of the defense models are evaluated using six attacks, i.e., Hot-Flip (Ebrahimi et al., 2017), SA, GA, GSA, GGA, and PCAA. **(2) IMDB reviews dataset** (Maas et al., 2011). Under this dataset, we focus on a word-level classification task and we study two model architectures, namely Bi-LSTM and CNN, trained for prediction. To evaluate the robustness, four attacks are deployed, including a genetic attack (Alzantot et al., 2018) (which is an attack method proposed to generate adversarial examples in embedding space), as well as SA, GSA, and PCAA . Note that for text data, we also consider preserving semantic meanings and grammatical correctness during defenses, and only perturb words with synonyms and correct grammatical forms.

Baseline Defenses. We compare our defense method PADVT with the following existing baseline defenses:

- **Standard training**. It minimizes the average cross-entropy loss on clean input.
- **Hot-flip** (Ebrahimi et al., 2017). It uses the gradient with respect to the one-hot input representation to find out which individual feature under perturbation has the highest estimated loss. It is initially proposed to model char flip in Char-CNN model, and we also apply it to word-level substitution, as in (Dong et al., 2021).
- **Adv l_2-ball** (Miyato et al., 2017). It uses an l_2 PGD adversarial attack inside the word embedding space for adversarial training.
- **ASCC-Defense** (Dong et al., 2021). A state-of-the-art defense method in text classification. It uses the worst perturbation over a sparse convex hull in the word embedding space for adversarial training.

Legible attacks on NLP. It is worth noting that when we apply our method to the text datasets IMDB, we consider the additional requirements of maintaining semantic meaning and grammatical correctness for NLP tasks. Following existing textual data attacking methods (Dong et al., 2021), we only switch some words with their synonyms while keeping correct grammatical forms and perturb each word separately. Specifically, we follow the same way of (Dong et al., 2021) which constructs the feasible perturbation space $\mathcal{S}(x)$ (see Eq.(1)) to maintain the semantic meaning and avoid grammatical errors, and apply the perturbation space into our framework. In Appendix A.4, we provide some real examples to show how words are replaced with their synonyms while keeping correct grammatic forms. We also provide adversarial texts for human evaluation to further confirm that this strategy meets the aforementioned requirements.

Performance Comparison. The experimental results on AG's news and IMDB are shown respectively in Fig. 4 and Fig. 3. The Y-axis represents the error rate and X-axis represents different attacks where each bar inside the group of an attack denotes one defense method. On AG's news dataset, our defense method achieves leading robustness on Char-CNN over all attacks with significant margins, surpassing Hot-Flip-defense by 10%. On the IMDB dataset, and we have similar observations to these on AG's news dataset. Our PADVT shows competitive adversarial robustness as ASCC defense. Notably, ASCC is a defense method that conducts adversarial training on word-embedding space. It relies on the key assumption that similar words have a close distance in embedding space. However, our method does not rely on this assumption, which may result in the performance being competitive (slightly worse) than ASCC. For all other defenses, PADVT outperforms them across different architectures significantly. Note that PADVT is based on PGD adversarial training. We also adapt TRADES to categorical data based on PCAA and details are shown in Appendix A.3.

5.3. Ablation Study

Concentration of PCAA . To further understand the behavior of our attack algorithm, in this subsection, we ask the question: *what is the variance of our optimized probability distribution π^* (from solving Eq.(2)?* Intuitively, we desire the distribution π to have a smaller variance, so that we don't need too many times sampling to obtain the optimal adversarial examples. To confirm this point, we conduct an ablation study based on an experiment on IPS dataset to visualize the distribution π, which is optimized via PCAA under various budget sizes. In Fig. 5, we choose three budget sizes $\epsilon = 1, 3, 5$ and randomly choose 3 features to present the adversarial categorical distributions, where the Y-axis represents the magnitude of unnormalized probabilities for each level within the feature. Notably, in Fig. 5, the left and middle two columns correspond to the feature distribution where the most probable category is the same as original category, and the right column are the feature that where

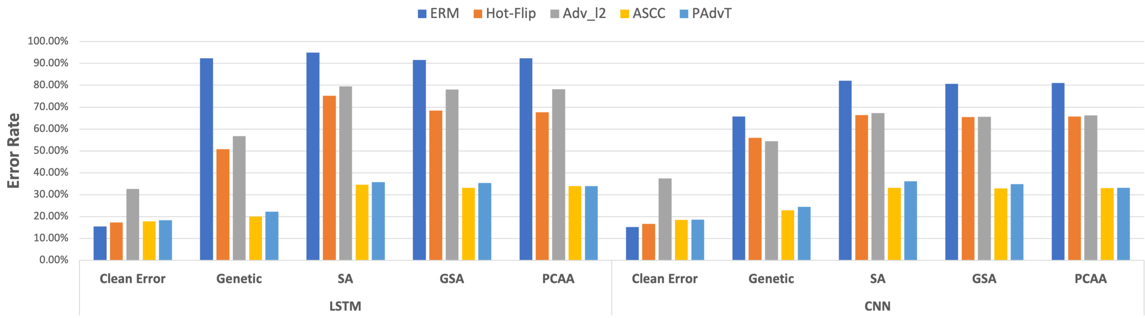


Figure 3. PAdvT and baseline defense performance under different attacks on IMDB dataset.

Table 3. Ablation study: impact of the budget regularization term ζ on PAdvT

	Clean Err		**Genetic SR**		**SA SR**		**Gradient Search SR**		**PCAA SR**	
	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN
$\zeta = 0.1$	16.1	16.9	37.5	29.6	41.3	43.5	39.5	41.4	40.2	42.2
$\zeta = 0.2$	17.2	17.3	34.1	32.7	39.9	41.3	38.6	39.7	39.1	40.6
$\zeta = 0.32$	17.9	18.1	30.1	31.4	38.7	38.3	38.5	37.6	38.8	37.8
$\zeta = 0.4$	18.4	18.6	26.3	28.5	37.8	36.2	36.4	34.9	36.9	35.7

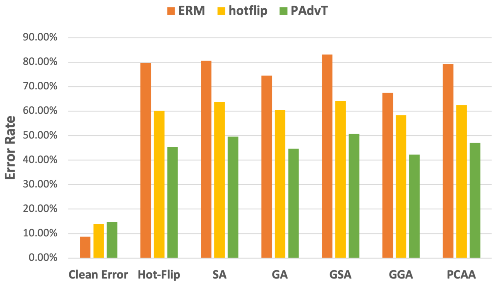


Figure 4. PAdvT and baseline defense performance under different attacks on AG's news dataset.

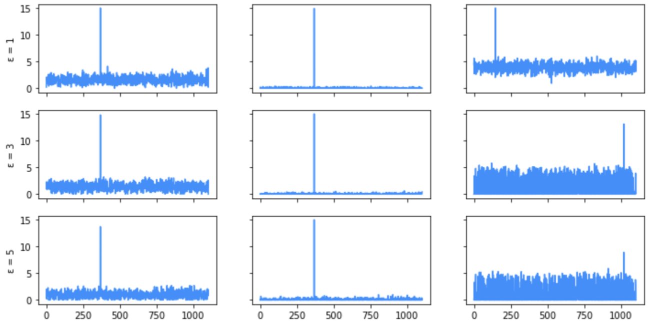


Figure 5. Visualization of Optimized Categorical Distribution for Various Features (IPS)

the most probable category is different. From the figure, we can see that for all features, there exists one category with a much higher probability compared to other categories. This fact indicates that during the sampling process of PCAA , the samples are highly likely to have the same category for a certain feature. As a result, we confirm that our sampled adversarial examples are well-concentrated.

The Impact of ζ on PAdvT . In our training objective in Eq.(4), ζ controls the budget size used for adversarial training and possibly affects the robustness of the model. We conduct an ablation study on the IMDB dataset to understand the impact of ζ. The results are demonstrated in Table 3. When ζ increases, the success rates of all attacks decrease, meaning that the robustness of the models is enhanced. However, large ζ will decrease the model accuracy. Thus, ζ controls the balance between the accuracy and the robustness of the model. When $\zeta = 0.4$, our algorithm reaches a good balance between accuracy and robustness.

6. Conclusion

In this paper, we propose a novel probabilistic framework, PCAA, to bridge the gap between categorical data and continuous data, which allows us to easily adapt gradient-based attacking methods in continuous data to categorical data. Our framework significantly improves the optimality-efficiency trade-off compared with search-based methods and shows promising empirical performance across different datasets. Furthermore, we adapt defenses in continuous data to categorical data through the proposed framework and achieve better robustness. Our future work will pursue transferring other advanced methods designed for the continuous domain, such as certified defenses (Cohen et al., 2019), to the categorical domain.

7. Acknowledgements

Han Xu, Pengfei He, Jie Ren and Jiliang Tang are supported by the National Science Foundation (NSF) under grant numbers CNS1815636, IIS1845081, IIS1928278, IIS1955285, IIS2212032, IIS2212144, IOS2107215, and IOS2035472, the Army Research Office (ARO) under grant number W911NF-21-1-0198, the Home Depot, Cisco Systems Inc, Amazon Faculty Award, and SNAP. Zitao Liu was supported in part by National Key R&D Program of China, under Grant No. 2022YFC3303600; and in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003).

References

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B., Srivastava, M. B., and Chang, K. Generating natural language adversarial examples. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2890–2896. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1316. URL https://doi.org/10.18653/v1/d18-1316.

Bao, H., Han, Y., Zhou, Y., Shen, Y., and Zhang, X. Towards understanding the robustness against evasion attack on categorical data. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=BmJV7kyAmg.

Barham, S. and Feizi, S. Interpretable adversarial training for text. *arXiv preprint arXiv:1905.12864*, 2019.

Cao, L. and Tay, F. E. H. Financial forecasting using support vector machines. *Neural Comput. Appl.*, 10(2):184–192, 2001. doi: 10.1007/s005210170010. URL https://doi.org/10.1007/s005210170010.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Cui, J., Liew, L. S., Sabaliauskaite, G., and Zhou, F. A review on safety failures, security attacks, and available countermeasures for autonomous vehicles. *Ad Hoc Networks*, 90, 2019. doi: 10.1016/j.adhoc.2018.12.006. URL https://doi.org/10.1016/j.adhoc.2018.12.006.

Dong, X., Luu, A. T., Ji, R., and Liu, H. Towards robustness against natural language word substitutions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=ks5nebunVn_.

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

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Khraisat, A., Gondal, I., Vamplew, P., and Kamruzzaman, J. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecur.*, 2(1):20, 2019. doi: 10.1186/s42400-019-0038-7. URL https://doi.org/10.1186/s42400-019-0038-7.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Lei, Q., Wu, L., Chen, P., Dimakis, A., Dhillon, I. S., and Witbrock, M. J. Discrete adversarial attacks and submodular optimization with applications to text classification. In Talwalkar, A., Smith, V., and Zaharia, M. (eds.), *Proceedings of Machine Learning and Systems 2019, MLSys 2019, Stanford, CA, USA, March 31 - April 2, 2019*. mlsys.org, 2019. URL https://proceedings.mlsys.org/book/284.pdf.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R. (eds.), *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pp. 142–150. The Association for Computer Linguistics, 2011. URL https://aclanthology.org/P11-1015/.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Miyato, T., Dai, A. M., and Goodfellow, I. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

Miyato, T., Dai, A. M., and Goodfellow, I. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1X3g2_xl.

Noordewier, M., Towell, G., and Shavlik, J. Training knowledge-based neural networks to recognize genes in dna sequences. *Advances in neural information processing systems*, 3, 1990.

Samanta, S. and Mehta, S. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*, 2017.

Stringhini, G., Kruegel, C., and Vigna, G. Detecting spammers on social networks. In Gates, C., Franz, M., and McDermott, J. P. (eds.), *Twenty-Sixth Annual Computer Security Applications Conference, ACSAC 2010, Austin, Texas, USA, 6-10 December 2010*, pp. 1–9. ACM, 2010. doi: 10.1145/1920261.1920263. URL https://doi.org/10.1145/1920261.1920263.

Tu, Z., Zhang, J., and Tao, D. Theoretical analysis of adversarial learning: A minimax approach. *Advances in Neural Information Processing Systems*, 32, 2019.

Wang, Y., Han, Y., Bao, H., Shen, Y., Ma, F., Li, J., and Zhang, X. Attackability characterization of adversarial evasion attack on discrete data. In Gupta, R., Liu, Y., Tang, J., and Prakash, B. A. (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 1415–1425. ACM, 2020. doi: 10.1145/3394486.3403194. URL https://doi.org/10.1145/3394486.3403194.

Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., and Jain, A. K. Adversarial attacks and defenses in images, graphs and text: A review. *Int. J. Autom. Comput.*, 17(2):151–178, 2020. doi: 10.1007/s11633-019-1211-x. URL https://doi.org/10.1007/s11633-019-1211-x.

Yang, P., Chen, J., Hsieh, C., Wang, J., and Jordan, M. I. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *J. Mach. Learn. Res.*, 21:43:1–43:36, 2020. URL http://jmlr.org/papers/v21/19-569.html.

Yurochkin, M., Bower, A., and Sun, Y. Training individually fair ML models with sensitive subspace robustness. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1gdkxHFDH.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *CoRR*, abs/1901.08573, 2019. URL http://arxiv.org/abs/1901.08573.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

A. Appendix

A.1. Detailed computation for time complexity

In this section, we provide the detailed computation of time complexity analysis mentioned in Section 3.4. Recall the assumptions that the whole dataset has N data points, each data point has n features, each feature has d categories and the budget of the allowed perturbation is ϵ. In the following time complexity analysis, one feedforward / backpropagate step is considered as one computational unit.

PCAA . In PCAA , the time complexity is only from gradient ascent. Here, we assume that we sample n_g times when estimating the expected gradient, and the maximum number of iterations is I in Algorithm 1. We compute gradient n_g times during one iteration, which consists of one feedforward and one backpropagate step. Thus, the time complexity is:

$$N \cdot n_g \cdot \mathcal{O}(1) \cdot I = C_1 N \cdot \mathcal{O}(1) \tag{6}$$

where C_1 is some constant related to n_g and I.

Search Attack (SA). SA consists of two stages. The first stage involves traversing all features. For the i_{th} feature, it replaces the original category with all other $d-1$ categories respectively, and records the change of the model loss for each category. The largest change is treated as the impact score for the i_{th} feature. Then it selects the top ϵ features with the highest impact scores to perturb. In the second stage, it finds the combination with the greatest loss among all possible combinations of categories for selected features. Each loss calculation above involves one feedforward step and totally there are $nd + d^\epsilon$ loss calculations. Therefore, the time complexity for SA is

$$N \cdot [\mathcal{O}(nd) + \mathcal{O}(d^\epsilon)] = N \cdot \mathcal{O}(nd + d^\epsilon)$$

Greedy attack (GA) (Yang et al., 2020). This method is a modified version of SA. The first stage is similar to that of SA, while the second stage searches for the best perturbation feature by feature via greedy search. For the i_{th} selected feature, it replaces the original category with one that results in the largest loss and then searches the next selected feature until all selected features are traversed. Each loss calculation above involves one feedforward step and totally there are $nd + \epsilon d$ loss calculations. It has the complexity:

$$N \cdot [\mathcal{O}(nd) + \mathcal{O}(\epsilon d)] = N \cdot \mathcal{O}(nd + \epsilon d)$$

Gradient-guided SA (GSA) (Lei et al., 2019). To determine which features to perturb, this method utilizes gradient information in the first stage. It computes the gradient of the loss function w.r.t the original input and treats the gradient of each feature as the impact score. Those ϵ features with the greatest impact scores are selected to be perturbed. In the second stage, it follows the same strategy as that of SA. The gradient calculation involves one feedforward and one backpropagate step, and the loss calculation involves one feedforward step per feature. Therefore the time complexity is:

$$N \cdot [\mathcal{O}(1) + \mathcal{O}(d^\epsilon)] = N \cdot \mathcal{O}(1 + d^\epsilon)$$

Gradient-guided GA (GGA). On the basis of GSA, it remains the same first stage and modifies the second stage by adopting the same strategy as that in the second stage of GA. Thus its time complexity is:

$$N \cdot [\mathcal{O}(1) + \mathcal{O}(\epsilon d)] = N \cdot \mathcal{O}(1 + \epsilon d).$$

A.2. Additional experimental results

To better compare the performance of different defenses, we provide exact results (error rates) corresponding to Fig. 4 and Fig. 3 and show them in Table 4 and Table 5, respectively. Specifically, we run each experiment 5 times and compute the 95% confidence interval.

We also run PAdvT with a mixture of adversarial examples and clean samples on the IMDB dataset. Results are shown in Table 6 where values represent error rates. It is noticeable that clean samples will slightly improve the clean performance and lead to a small decrease in robustness.

A.3. A categorical defense based on TRADES

To better show the capability of our probabilistic framework, we apply another effective continuous defense method, TRADES (Zhang et al., 2019), on the categorical dataset AG's news through our framework, and the results are shown in

Table 4. PAdvT and baseline defense performances under different attacks on IMDB dataset

	Clean Err		**Genetic SR**		**GS SR**		**Gradient Search SR**		**PCAA SR**	
	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN
ERM	15.50±0.005	15.23±0.007	92.62±0.022	65.68±0.031	94.86±0.030	82.02±0.049	91.61±0.011	80.54±0.030	92.26±0.087	81.42±0.073
Hotflip	17.37±0.049	16.57±0.051	50.63±0.030	55.93±0.025	75.24±0.015	66.35±0.012	67.96±0.021	65.47±0.022	68.08±0.069	65.90±0.074
Adv_l2	32.53±0.034	37.38±0.043	56.59±0.036	54.35±0.033	79.69±0.030	67.00±0.034	78.34±0.031	65.81±0.038	78.58±0.068	66.23±0.077
ASCC	17.76±0.036	18.37±0.030	20.05±0.047	22.52±0.046	34.67±0.047	33.28±0.045	33.46±0.054	32.61±0.061	33.97±0.101	33.01±0.082
PAdvT	18.57±0.033	18.85±0.049	22.32±0.065	24.50±0.049	37.30±0.063	36.36±0.061	35.60±0.067	34.85±0.053	33.90±0.104	33.20±0.093

Table 5. PAdvT and baseline defense performances under different attacks on AG's news dataset

	Clean Err	Hotflip	PCAA	GS	GGS	GA	GGA
ERM	8.70±0.009	79.72±0.015	80.75±0.066	83.09±0.015	79.28±0.010	74.43±0.010	67.53±0.016
Hotflip	13.99±0.017	60.07±0.013	63.47±0.057	64.28±0.018	62.41±0.019	60.51±0.017	58.33±0.013
PAdvT	14.62±0.028	45.38±0.037	49.50±0.081	50.65±0.035	47.14±0.041	44.71±0.040	42.18±0.045

Table 6. Comparison of PAdvT on IMDB dataset with/without mixture of clean samples.

	Clean Err	Genetic	GS	GGS	PCAA
IMDB LSTM(mix)	18.27	26.22	37.67	35.79	36.05
IMDB LSTM	18.57	26.01	37.3	35.60	35.46
IMDB CNN(mix)	18.69	28.51	36.47	34.96	35.72
IMDB CNN	18.85	28.35	36.36	34.85	35.47

Table 7. PAdvT and TRADES on AG's news

	clean error	hotflip	PCAA	SA	GSA	GA	GGA
ERM	8.70	79.72	80.75	83.09	79.28	74.43	67.53
PAdvT	14.62	45.38	49.50	50.65	47.14	44.71	42.18
TRADES$(1/\lambda = 1)$	13.89	46.21	50.83	51.32	48.45	45.98	43.17
TRADES$(1/\lambda = 5)$	14.31	45.57	49.76	50.89	47.66	45.12	42.31

Table 7. In detail, we modify Algorithm 2 to implement this defense, and replace the average adversarial loss in line 8 with the TRADES loss on probabilistic distribution π, i.e

$$\mathcal{L}(f(x,\theta),Y) + \max_{\pi} \mathbb{E}_{x' \sim \pi} \mathcal{L}(f(x',\theta), f(x,\theta))/\lambda$$

According to the results, our framework can be easily leveraged for TRADES and achieves good performance, and can reduce error under different attacks. We conduct this defense with different choices of regularizer parameter $1/\lambda$, and the results show that the combination of PCAA and TRADES can achieve both high robustness and high accuracy on categorical data, indicating the capability of our framework.

A.4. Case studies on IMDB

To better illustrate how our method can maintain original meanings and grammaticals when applied to NLP tasks, we provide some case studies on IMDB dataset. First, we present 2 cases including original texts and replacement for the perturbed words. The first case shows that theses words will be replaced with their synonyms, and the second case indicates that the grammatical form of replaced words will be consistent with the original words. Original words are marked in blue, while their replacements are listed in red.

- **Case 1**. Synonyms.

  The cast is excellent({admirable, distinguished, exquisite, finest, first-rate, good, magnificent ,Outstanding, skillful, sterling, superb, marvelous, best, attractive, great, exceptional, accomplished, fine, exemplary, first-class}), the acting good, the plot interesting, the evolvement full of suspense, but it is hard to cram all those elements into a film that is barely 80 minutes long. If more time was taken to develop the plot and subplots, it would have a much better effect.

Another 30 minutes of substance would have made this a very good({acceptable, exceptional, favorable, great, right, marvelous, satisfying, superb ,valuable, wonderful, ace, bad, capital, nice, pleasing, excellent, positive, satisfactory, excellent, fine}) film rather than just a good one.

- **Case 2**. Grammatricals.

  There is great detail in a 'bug 's life'. Everything is covered. The film looks({glances, peeks, reviews, stares, views, expects, casts, gazes, inspects, leers, notices, observes, regards, sight, watches, sees, glimpses, reads, cares, notes}) great and the animation is sometimes jaw dropping. The film isn't too terribly original.

Moreover, human evaluation is usually needed in NLP tasks, so we provide some adversarial texts generated by PCAA and show in Tables 8 and 9. In detail, we run PCAA attack on IMDB dataset over two victim models LSTM and word-CNN. The candidate sets are pre-specified synonym sets. Similariy, adversarial words are marked in red while original words are in blue. It is obvious that these replacements do not hurt the semantic meaning but can fool the classifiers.

Table 8: IMDB Adversarial Examples from PCAA on LSTM

Class	Perturbed Class	Perturbed Texts
Negative	Positive	I watched this film for 45 minutes and counted 9 mullets. That's a mullet every 5 minutes. Seriously, though this film is residing evidence(living proof) that formula works if it ain't broke, it don't need fit in a streetwise yet vulnerable heroine, a hardened ex-cop martial arts master with a heart of gold and a serial killer with 'issues' pure magic.
Negative	Positive	Claustrophobic camera angles that do not aid(help) the movie. Too long face only shots, where you most of the time get the hunch(feeling) that the lower half of the film is missing that the screen is cut off because there seems to be important actions going on, but you can not see them. There is anyway already too much confusion in the movie, so these viewing angles make it worse and do not contribute to artful visuals. I like artfully made movies and unconventional camera work. I can handle deep and slow movies but this one is trying too hard to be something artful and fails, in my opinion, painfully. Nothing to get attached to any of the characters because they are not worked out well enough to work out characters. More is needed than just minute long face shots. At least with this set of script director actors, I wonder whether some of the not so decent(good) acting is due to the script and director or due to the actors. I will stay away from films both written and directed by le you for sure in the future. What an annoying film even for person(someone) who would be interested in that part of history and for someone who spent time in Shanghai.
Positive	Negative	I really liked this version of 'vanishing point' as opposed to the 1971 version. I finds(found) the 1971 version quite boring if I can get up in the middle of a movie a few times as I did with the 1971 version, then to me it is not all that great. Of course, this could be due to the fact that I was only nine at the time the 1971 version was brought out. However, I have noticed(seen) many remakes everytime(where) I have liked the original and older one better. I found that the plot of the 1997 version was more understandable and had basically kept true to the original without undermining the meaning of the 1971 version. In my opinion I felt the 1997 version had more excitement and wasn't so blasé boring.

Positive	Negative	The cast is marvellous(excellent), the acting good, the plot interesting, the evolvement full of suspense, but it is hard to cram all those elements into a film that is barely 80 minutes long. If more time was taken to develop the plot and subplots, it would have a much better effect. Another 30 minutes of substance would have made this a very right(good) film rather than just a good one.
Positive	Negative	There is great detail in a 'bug 's life'. Everything is covered. The film expects(looks) great and the animation is sometimes jaw dropping. The film isn't too terribly original. It 's basically a modern take on kurosawa 's seven samurai only with bugs, I enjoyed the character interaction however, and the naughty boys(bad guys) in this film actually seemed bad. It seems that Disney usually makes their bad guys carbon copy cut outs, the grasshoppers are menacing and hopper the lead bad guy was a brilliant creation. Check this one out.

Table 9: IMDB Adversarial Examples from PCAA on word-CNN

Class	Perturbed Class	Perturbed Texts
Positive	Negative	I am a college student studying A levels and need help and comments from anyone who has any views at all about the theme of mothers in film. In The Mother, whether you have gone through something similar or just want to comment and help me research more about this film, any comment would much greatly appreciated. The comments will be used alone(solely) for exam purposes and will be included in my written exam. So if you have any views at all I'm convinced(sure) I can put them to use and you could help me get an A. I am also studying about a boy and tadpole. So if you have seen these films as well, I would appreciate it if you could leave comments on here on that page. Thank you.
Negative	Positive	This movie is so horrendous(awful). It is hard to find the right words to describe it. At first the story is so ridiculous. A narrow minded human can write a better plot. The actors are boring and untalented. Perhaps they were compelled to play in this dorky(cheesy) film. The camera receptions of the national forest are the only good in this whole movie. I should feel ashame because I paid for this lousy picture. Hopefully nobody makes a sequel or make a similar film with such a worse storyline.
Positive	Negative	This movie is wonderful, the writing, directing, acting, all are marvelous(fantastic). Very witty and clever script quality performances by actors. Ally Sheedy is strong and dynamic and delightfully quirky really original and heart warmingly unpredicatable. The scenes are alive with fresh energy and really talented generating(production)

Positive	Negative	This may not be war peace but the two academy noms wouldn't have been forthcoming. If it weren't for the genius of James Wong Howe, this is one of the few films I've fallen in love with as a infant(child) and gone back to without dissatisfaction. Whether you have any interest in what it offers fictively or not, BBC is a visual feast. I'm not saying it's his best work. I'm no expert there for sure but the look of this movie is astounding(amazing). I love everything about it, Elsa Lanchester, the cat, the crazy hoodoo, the retro downtown Ness, but the way it was put on film is breathtaking. I even like the inconsistencies pointed out on this page aforementioned(above) and the special effects that seem backward. Now it all creates a really consistent world.
Positive	Negative	Bette Midler is again divine raunchily hilarious(humorous) in love with burlesque, capable of bringing you down to tears either with old jokes, with new dresses or merely with old songs, with more power punch than ever. All in all, sung(singing) new ballads power, singing the good old perennial ones such as the rose 'stay with me' and yes even 'wind beneath my wings'. The best way to appreciate the Divine Miss M has always been libe since this is the next best thing to it. I strongly recommended to all with a mixture of adult extensive(wide) eyed enchantment and appreciation and a child 's mischievous wish for pushing all boundaries.