Provably Unlearnable Data Examples

Derui Wang*†, Minhui Xue*†, Bo Li‡, Seyit Camtepe*† and Liming Zhu*
*CSIRO's Data61, Australia
†Cyber Security Cooperative Research Centre, Australia
‡University of Chicago, USA
*{derek.wang, jason.xue, seyit.camtepe, liming.zhu}@data61.csiro.au, ‡bol@uchicago.edu

***Abstract*—The exploitation of publicly accessible data has led to escalating concerns regarding data privacy and intellectual property (IP) breaches in the age of artificial intelligence. To safeguard both data privacy and IP-related domain knowledge, efforts have been undertaken to render shared data unlearnable for unauthorized models in the wild. Existing methods apply empirically optimized perturbations to the data in the hope of disrupting the correlation between the inputs and the corresponding labels such that the data samples are converted into Unlearnable Examples (UEs). Nevertheless, the absence of mechanisms to verify the robustness of UEs against uncertainty in unauthorized models and their training procedures engenders several under-explored challenges. First, it is hard to quantify the unlearnability of UEs against unauthorized adversaries from different runs of training, leaving the soundness of the defense in obscurity. Particularly, as a prevailing evaluation metric, empirical test accuracy faces generalization errors and may not plausibly represent the quality of UEs. This also leaves room for attackers, as there is no rigid guarantee of the maximal test accuracy achievable by attackers. Furthermore, we find that a simple recovery attack can restore the clean-task performance of the classifiers trained on UEs by slightly perturbing the learned weights. To mitigate the aforementioned problems, in this paper, we propose a mechanism for certifying the so-called (q, η)-Learnability of an unlearnable dataset via parametric smoothing. A lower certified (q, η)-Learnability indicates a more robust and effective protection over the dataset. Concretely, we 1) improve the tightness of certified (q, η)-Learnability and 2) design Provably Unlearnable Examples (PUEs) which have reduced (q, η)-Learnability. According to experimental results, PUEs demonstrate both decreased certified (q, η)-Learnability and enhanced empirical robustness compared to existing UEs. Compared to the competitors on classifiers with uncertainty in parameters, PUEs reduce at most 18.9% of certified (q, η)-Learnability on ImageNet and 54.4% of the empirical test accuracy score on CIFAR-100. Our source code is available at https://github.com/NeuralSec/certified-data-learnability.**

I. Introduction

Data privacy and intellectual property (IP) breaches have become major concerns in recent legislations and regulations serving for building responsible Artificial Intelligence (AI), such as GDPR [1], CCPA [2], the European Union AI Act [3], and the recent US Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI [4]. These regulations have emphasized that consumers have the right to limit the use and disclosure of data collected from them. Despite the laws and regulations, technologies for enhancing data privacy and protecting IP face increasingly severe challenges due to rapidly evolving machine learning algorithms. Particularly, data published by individual users on content-sharing social media platforms is exposed to the threat of domain exploitation by unauthorized machine learning models. For instance, a pretrained Stable Diffusion can be fine-tuned on a small set of paintings to generate new images mimicking the style of the paintings [5], [6]. As another example, classifiers trained on publicly accessible data can be used to launch membership inference attacks exposing confidential personal information [7]. As data becomes the new oil, these threats highlight the imminent need to implement effective measures for Data Availability Control (DAC).

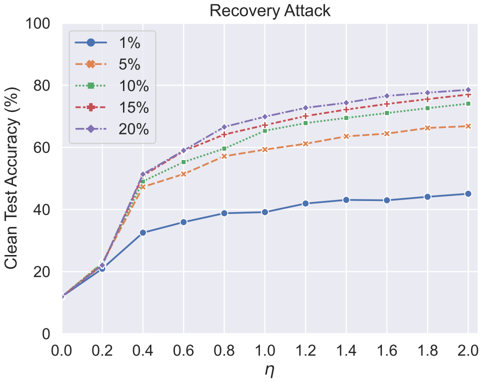

Fig. 1: Recovery attacks using a small portion (1%-20%) of the CIFAR10 training set. Points on the curves trace the clean testing accuracy of classifiers whose weights are perturbed away from the poisoned classifier. For each fine-tuned classifier, the ℓ_2 norm of the weight perturbation is capped by η. This special adversary reveals that current UEs are not robust against uncertainty in classifier parameters, and their reliability cannot be guaranteed. Therefore, a mechanism for certifying UE performance is pivotal.

As a response, a series of attempts have been made to make the published data unlearnable. One pragmatic and effective way to protect image datasets against unauthorized classifiers is to perturb the pixels before publishing the images. Existing defenses apply anti-learning perturbations that minimize the training loss with respect to either correct or target labels on the input data. [8], [9]. In lieu of learning functions mapping the input data to the corresponding labels, the defenses encourage unauthorized models to learn a strong correlation (*i.e.,* a shortcut) between the perturbations and the labels. As a consequence, despite that the trained models can achieve low

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.240886
www.ndss-symposium.org

error rates on the data with the anti-learning perturbations, they generalize poorly to datasets from the clean data distribution. This line of work is also referred to as Perturbative Availability Poison (PAP) in the literature [10], [11], [8], [12].

Motivation. Existing PAP methods empirically search for the anti-learning perturbations based on a finite amount of training samples and models [9], [13], [11], [8], [14], [15], [16]. Henceforth, the generated perturbations may face the cross-model generalization problem brought by unknown illegal classifiers in the wild. Moreover, there is no rigid guarantee of the maximally attainable learning results for adversaries on the UEs. The widely adopted empirical test accuracy, as a metric, is prone to uncertainty in the data, model selection, and training process. This indicates that empirical test accuracy is not a sound metric for evaluating UEs. UEs resulting in lower test accuracy scores on certain models and datasets do not necessarily evince their superiority over others and may leave room for attackers. Let us consider a type of recovery attack in which an attacker can fine-tune a poisoned classifier on a small set of clean data to restore its clean-task performance without significantly changing the classifier parameters. This kind of adversary may exist in federated learning systems based on weight aggregation with Byzantine robustness [17]. For example, an adversarial client training its unauthorized classifier on UEs can perform a recovery attack and upload the parameters to the server without alerting outlier detectors, rendering the global classifier to violate corresponding DAC protocols. We randomly collect 1%-20% of CIFAR10 training samples to fine-tune a ResNet-18 trained on the error-minimizing PAP [9]. We use projected stochastic gradient descent (SGD) in the fine-tuning so the ℓ_2 norm of the weight differences is restricted within a small range. Surprisingly, we find that the accuracy can be restored from around 0.1 to near 0.8 by seemingly minuscule modifications in the weights using 5%-10% of CIFAR10 samples (Figure 1). This result reflects that the robustness of UEs towards weight perturbations and uncertainties in classifier parameters is neither rigidly gauged nor carefully pondered.

The aforementioned problems hinder the assessment of UEs and constitute additional risks to be exploited by deliberate adaptive adversaries. Therefore, it is crucial to develop a mechanism to ensure that a set of UEs can take effect without being ravaged by the problems of training stochasticity, cross-model generalization, and adaptive attackers. The mechanism can resort to *finding classifiers that can achieve the best possible clean-task accuracy when trained on UEs and restraining this best-case accuracy.* We are thus motivated to ask the question:

> *Is it possible to go beyond heuristics and provide an upper bound towards the clean-data performance of undivulged models trained on a set of UEs?*

The question is daunting to be thoroughly answered. However, we will show in this work that, it is possible to derive a clean-data accuracy upper bound guaranteed with a high probability, for unauthorized classifiers under some constraints. On the other hand, this threshold also serves as an upper bound of clean accuracy when facing the recovery attack. It can henceforth be a tool for gauging the effectiveness and robustness of UEs.

Our method and its broader impact. In this paper, we propose a mechanism for certifying the best clean-task accuracy retained by unauthorized classifiers trained on a set of UEs. The mechanism can be applied to any empirically made UEs to derive a *certified* (q, η)*-Learnability*. (q, η)-Learnability gauges the effectiveness and robustness of UEs towards arbitrary classifiers sampled from a certain parameter space. Note that the certification pertains to the parameter space rather than the input space, alleviating the need for knowledge of the training strategies or learning algorithms used by the attackers. We also proposed certifying a generalization (q, η)-Learnability, which removes the requirement for holding private test datasets in the evaluation of UEs. To make the certification more sound, the paper also proposes a simple method to expand the certifiable parameter space and reduce the gap between the certified (q, η)-Learnability and the True Learnability (see Definition 1). Moreover, we produce *Provably Unlearnable Examples* (PUEs) which achieve lower (q, η)-Learnability scores and be more robust than existing PAP methods.

The (q, η)-Learnability metric serves as a rigid guarantee of data availability for machine learning models. This guarantee can operate in parallel with other techniques, such as deep watermarking [18] and differential privacy (DP) [19], to deter illegal data exploitation and support responsible AI regulations and standards (e.g., CCPA, GDPR, and Australia's AI Safety Standard [20]). Instead of post-event ownership auditing through watermarks and protection of sensitive information with DP, (q, η)-Learnability proactively controls access to domain knowledge through data, safeguarding both data IP and privacy. The contributions of this paper are as follows:

- We formally derive the *certified* (q, η)*-Learnability* of an unlearnable dataset. The certified (q, η)-Learnability serves as a guaranteed upper bound of the clean test accuracy that can be achieved by particular classifiers trained on the unlearnable dataset. To the best of our knowledge, this is the *first* attempt towards guaranteed effectiveness of PAPs.
- As the *first* step towards DAC with provable guarantees, we propose a simple method to narrow the gap between certified (q, η)-Learnability and True Learnability. Moreover, we design PUEs that can suppress the certified (q, η)-Learnability.
- PUEs not only achieved lower certified (q, η)-Learnability but also demonstrated better empirical robustness in scenarios where unauthorized classifiers cannot be certified.
- Our source code provides a tool for measuring the provable effectiveness of UEs and producing PUEs.

II. Background

In this section, we briefly introduce the concepts of UEs and smoothed classifiers.

A. Unlearnable Examples

Let a set $\mathbb{D}_s := \{(x_i, y_i) | (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{N}$ be a source dataset containing N labeled samples. $\mathcal{X}$ and $\mathcal{Y}$ are the I-dimensional input space and the K-dimensional label space, respectively. $\mathcal{X} \times \mathcal{Y}$ is the Cartesian product of the two spaces. The aim of the adversary is to train a classifier $f : \mathcal{X} \subset \mathbb{R}^I \to \mathcal{Y} \subset \mathbb{R}^K$ by empirically minimizing a training loss $\mathcal{L}(\cdot)$ over $\mathbb{D}_s$ as follows:

$$\min_f \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i). \tag{1}$$

Suppose $\mathcal{D}$ is the clean distribution from which $\mathbb{D}_s$ is drawn. The trained f is supposed to perform well on samples from $\mathcal{D}$. Once $\mathcal{D}$ is a data distribution from which other datasets with privacy concerns or registered intellectual property are drawn, it is crucial to make sure $\mathbb{D}_s$ cannot be exploited in the unauthorized training of models which will later be applied to infer other private datasets drawn from $\mathcal{D}$. Instead of maximizing the training error, the common goal of the defender is to find a perturbation $\delta \in \mathcal{X}$, such that

$$\begin{aligned} &\min_{\hat{f}} \mathbb{E}_{(x,y)\sim\mathbb{D}_s} \min_{\delta} \mathcal{L}(\hat{f}(x+\delta),\, y) \\ &\quad s.t.\ \|\delta\|_p \leq \xi, \end{aligned} \tag{2}$$

where $\hat{f}$ is a surrogate model used for searching δ and $\|\delta\|_p$ is the ℓ_p-norm of the perturbation bounded by ξ. The ℓ_∞ norm is commonly used in the previous literature. Such $x+\delta$ is a UE and δ is the PAP noise inserting a noise-label shortcut to minimize the training loss. However, since the loss is empirically calculated, there is no rigid guarantee on the performance of the noise δ.

B. Smoothed Classifiers

A smoothed classifier $g(x)$ predicts by returning the majority vote of predictions from a K-way base classifier $f_\theta : \mathbb{R}^I \to \{1, 2, ..., K\}$ parameterized by θ over random noises applied to either the input $x \in \mathbb{R}^I$ or the parameters during the test time. That is,

$$\begin{aligned} g(x) &= \arg\max_{y\in\mathcal{Y}} \Pr_{\epsilon\sim\pi(x)}[f_\theta(x+\epsilon) = y], \quad or \\ g(x) &= \arg\max_{y\in\mathcal{Y}} \Pr_{\epsilon\sim\pi(\theta)}[f_{\theta+\epsilon}(x) = y]. \end{aligned} \tag{3}$$

$\pi(x)$ is the distribution of the noise (*i.e.,* smoothing distribution) centered at x (or θ) and y is a predicted label from the label space $\mathcal{Y}$. The output from the smoothed classifier can be statistically bounded to offer certifiable robustness with respect to changes in the smoothed variables. Cohen et al. derived a tight ℓ_2 robustness bound by smoothing classifier inputs with noise drawn from Gaussian distributions and certifying based on Neyman-Pearson Lemma [21]. Subsequently, the certified robustness is extended to various smoothing noises and certification frameworks [22], [23], [24], [25], [26], [27], [28], [29], [30]. Such smoothed classifier $g(x)$ can also be extended to randomized learning functions which in addition smooth the training dataset $\mathbb{D}_s$, resulting

$$g(\mathbb{D}_s, x) = \arg\max_{y\in\mathcal{Y}} \Pr_{\hat{\mathbb{D}}_s\sim\pi(\mathbb{D}_s)}[f_\theta(\hat{\mathbb{D}}_s,\ x) = y]. \tag{4}$$

Once we get the background information of UEs and randomized classifiers, we will move to define the research problem and the threat model of this paper in the following section.

III. Problem Statement

Problem definition. Inheriting from the previous definitions, consider a to-be-published dataset $\mathbb{D}_s$ (e.g., portraits to be posted on Facebook) with N data points. Additionally, there is a private dataset $\mathbb{D}$ (e.g., patient records with photos in a hospital database) sharing the same distribution $\mathcal{D}$ with $\mathbb{D}_s$. The defender's goal is to render $\mathbb{D}_s$ unlearnable against unknown learning algorithms and classifiers (e.g., deep neural networks) by perturbing it into $\mathbb{D}_s \oplus \delta$, where δ is the anti-learning perturbation and $\oplus$ is the perturbing operator. We specifically consider additive perturbations $\delta = \{\delta_i\}_{i=1}^{|\mathbb{D}_s|}$ applied to each sample of $\mathbb{D}_s$ in this paper. Thus, $\oplus$ represents the element-wise addition. "Unlearnable" here means that a classifier trained on $\mathbb{D}_s \oplus \delta$ would exhibit a high classification error on $\mathbb{D}$. Furthermore, the defender wants to verify how robust $\mathbb{D}_s \oplus \delta$ is by having an upper bound on the accuracy that can be achieved by unknown classifiers stochastically trained on $\mathbb{D}_s \oplus \delta$. More profoundly, the defender wants to determine the best possible accuracy that classifiers trained on $\mathbb{D}_s \oplus \delta$ can achieve on $\mathbb{D}$. The defender does not desire an elevated best possible accuracy since it implies $\mathbb{D}_s \oplus \delta$ is exploitable. Therefore, the research goals are as follows.

- A certification mechanism should be developed to verify the maximum accuracy achievable by any classifiers and learning algorithms on $\mathbb{D}$ through training on $\mathbb{D}_s \oplus \delta$.
- A better noise δ should be designed such that $\mathbb{D}_s \oplus \delta$ can effectively decrease the verified best accuracy.

To better formulate the problem, we first formally define the *True Learnability* of $\mathbb{D}_s \oplus \delta$ for classification tasks.

Definition 1 (*True Learnability of a perturbed dataset*)**.** *Given a δ-perturbed dataset $\mathbb{D}_s \oplus \delta$, let $f_{\hat{\theta}}$ be a hypothesis with parameters $\hat{\theta} \in \mathbb{R}^d$ selected by arbitrary learning algorithm $\Gamma(\cdot)$, such that $\hat{\theta} = \Gamma(\mathbb{D}_s \oplus \delta)$. $\mathbb{D}$ and $\mathbb{D}_s$ come from the same data distribution $\mathcal{D}$. Suppose there are sample pairs $(x, y) \in \mathbb{D}$. The learnability of the perturbed training dataset $\mathbb{D}_s \oplus \delta$ is*

$$\mathbf{L}(\Theta; \mathbb{D}_s \oplus \delta) := \max_{\hat{\theta}\in\Theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \mathbb{1}[f_{\hat{\theta}}(x) = y], \tag{5}$$

where $\mathbb{1}[\cdot]$ is the indicator function and Θ is the space of all possible parameters that can be selected.

Herein, $\hat{\theta}$ can be viewed as a function of $D_s \oplus \delta$. In other words, $\mathbf{L}(\Theta; \mathbb{D}_s \oplus \delta)$ is the *best possible* testing accuracy a classifier from the space Θ can achieve when evaluated on $\mathbb{D}$. Note that the learnability defined here differs from PAC learnability [31]. $\mathbf{L}(\cdot)$ deterministically reflects the lowest generalization error any classifier in the space Θ can achieve. A smaller $\mathbf{L}(\cdot)$ indicates that the learned classifier is harder to be generalized to the test-time dataset. After having the definition of learnability, we can formulate the aim of an anti-learning perturbation δ as follows.

$$\begin{aligned} &\min_{\delta} \mathbf{L}(\Theta; \mathbb{D}_s \oplus \delta), \\ &s.t.\ \forall\ \delta_i \in \delta,\ \|\delta_i\|_p \leq \xi. \end{aligned} \tag{6}$$

Herein, ξ is a fixed ℓ_p perturbation budget for noises added to each data point. An optimal anti-learning noise is the minimizer of the learnability.

In order to find the δ, we need to first compute $\mathbf{L}(\Theta; \mathbb{D}_s \oplus \delta)$. Solving $\mathbf{L}(\Theta; \mathbb{D}_s \oplus \delta)$ in an unknown space Θ is intractable. However, we will show that a learnability score guaranteed with a high probability can be found if we restrict the space of possible hypothesis by only considering $\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta} + \upsilon, \sigma^2 I),\ \|\upsilon\| \leq \eta\}$, where $\|\cdot\|$ is the ℓ_2 norm, υ is a parametric perturbation, η and σ are two constants, and $\hat{\theta}$ is a set of surrogate parameters selected by the defender based on the dataset $\mathbb{D}_s \oplus \delta$ to form the hypothesis $f_{\hat{\theta}}$. If a set of

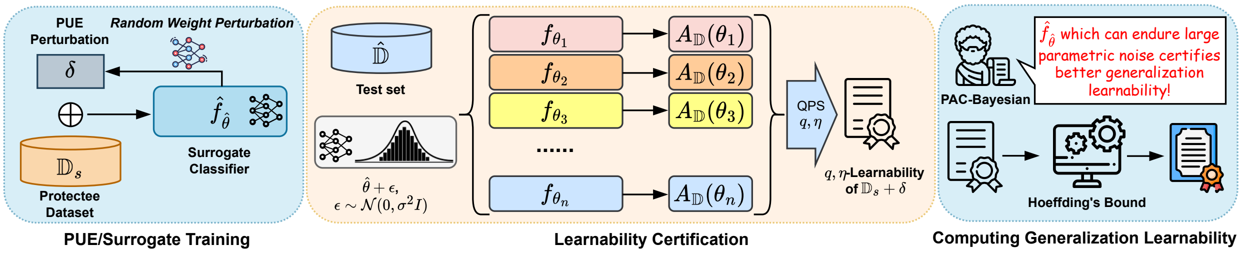
PUE Perturbation
Random Weight Perturbation
δ
$\oplus$
$\hat{f}_{\hat{\theta}}$
Surrogate Classifier
$\mathbb{D}_s$
Protectee Dataset
$\hat{\mathbb{D}}$
Test set
$\hat{\theta} + \epsilon$,
$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$
f_{θ_1}
f_{θ_2}
f_{θ_3}
......
f_{θ_n}
$A_{\mathbb{D}}(\theta_1)$
$A_{\mathbb{D}}(\theta_2)$
$A_{\mathbb{D}}(\theta_3)$
$A_{\mathbb{D}}(\theta_n)$
QPS q, η
q, η-Learnability of $\mathbb{D}_s + \delta$
PAC-Bayesian
$\hat{f}_{\hat{\theta}}$ which can endure large parametric noise certifies better generalization learnability!
Hoeffding's Bound
PUE/Surrogate Training
Learnability Certification
Computing Generalization Learnability

Fig. 2: An overview of the certification and PUE crafting framework. A dataset $\mathbb{D}_s$ is perturbed into $\mathbb{D}_s \oplus \delta$ before being released to the public. The (q, η)-Learnability of $\mathbb{D}_s \oplus \delta$ can be certified to ensure that any unauthorized classifier trained on $\mathbb{D}_s \oplus \delta$ has a provable upper bound on its performance on any test set $\hat{\mathbb{D}}$ (or $\mathbb{D}$) in the same domain with $\mathbb{D}_s \oplus \delta$, as long as the parameters of the unauthorized classifier are within a certified parameter set. Generalization learnability can be computed using Hoeffding's bound, and PAC-Bayesian theory suggests that a certification surrogate with low prediction variance under large parametric noise can improve certified learnability. Optimized PUEs lead to lower (q, η)-Learnability.

PAP noises can reduce such a learnability score, they could achieve theoretically more robust protection of datasets against the uncertainty in classifiers and learning algorithms.

In essence, the problem to be addressed in this paper is two-fold. First, we certify the learnability of a dataset towards arbitrary classifiers trained on it as long as the trained parameters are from $\hat{\Theta}$. This setting non-trivially appears in data exploitation cases through *model retraining*, *transfer learning*, *model fine-tuning*, and *recovery attacks*. Second, we strive to reduce the learnability of the protected dataset by designing a proper δ. We will present in the following part a threat model of the paper.

Threat model. The producer of UEs (*i.e.,* defender) has white-box access to a surrogate classifier $f_{\hat{\theta}}$ and a protectee dataset $\mathbb{D}_s \sim \mathcal{D}$. Additionally, either a clean test dataset $\mathbb{D}$ or the domain $\mathcal{D}$ of $\mathbb{D}_s$ is revealed to the defender. Note that it is *not* necessary for the defender to access $\mathbb{D}$ if it is privately held by a third party. Instead, the defender can sample a dataset $\hat{\mathbb{D}}$ from $\mathcal{D}$ to compute a certified generalization learnability which serves as a generalization upper bound on the certified learnability. The defender can alter the parameters of $f_{\hat{\theta}}$ and modify $\mathbb{D}_s$ at will. The defender crafts an unlearnable version $\mathbb{D}_s \oplus \delta$ of $\mathbb{D}_s$ and releases the unlearnable dataset rather than the unprotected one to the public. Besides, defenders can either train surrogate classifiers on generated $\mathbb{D}_s \oplus \delta$ or simultaneously with δ on $\mathbb{D}_s$. However, the defender has zero access to any potential unauthorized classifiers and is unaware of the training procedure for the unauthorized classifiers.

As the adversaries in our threat model, the unauthorized classifiers can obtain labeled $\mathbb{D}_s \oplus \delta$ as their training dataset. The trained unauthorized classifiers will be used to infer the labels of samples in $\mathcal{D}$. Importantly, the exact anti-learning perturbation δ is hidden from the adversaries. Our defense tackles three levels of adversaries.

- **General Adversary (GA)**: Unauthorized classifier whose trained parameters are in $\mathbb{R}^d$, where d is the parameter size.
- **Certifiable Adversary (CA)**: Unauthorized classifier whose trained parameters fall in a certified parameter set $\hat{\Theta}$.
- **Special Adversary (SA)**: Recovery attacker which changes the model parameters within an ℓ_2 norm bound.

Among these adversaries, GA encompasses almost all attackers in real-world scenarios. CA denotes the subset of attackers from GA that can be provably mitigated by the defense. SA serves as a tool to gauge the robustness of different UEs against GAs who cannot be provably invalidated, as well as to assess the tightness of our defense against CAs. The defender aims to expand $\hat{\Theta}$ to improve the proportion of CAs within GAs. We will introduce in Section VI a method for augmenting $\hat{\Theta}$.

IV. Overview

We will provide a brief overview over the crux of learnability certification and PUE generation in this section. Our first desideratum is measuring how well an unlearnable dataset can sabotage those illegal but unknown classifiers by certifying a possibly best learning outcome on this dataset. In a nutshell, a surrogate classifier selected based on the unlearnable set $\mathbb{D}_s \oplus \delta$ can be viewed as a hypothesis of the data distribution. Intuitively, if the distribution of the unlearnable set is highly disjoint from that of clean data, the classifier would demonstrate a significant generalization error on the clean test dataset. In contrast, if the unlearnable set is suboptimal, the classifier might still be able to generalize to the clean dataset and thus induce a lower generalization error. Herein, we convert the certification problem to certifying a surrogate classifier appropriately modeling the distribution of UEs.

Second, we design a simple method for selecting the surrogate classifier used in the certification. Importantly, we emphasize that the two entities, the UE and the surrogate classifier, serve different objectives. On the side of UEs, their purpose is upfront — to minimize the learnability by lowering the clean testing accuracy of any classifiers trained on them. Instead, the surrogate should *1) resemble the distributional characteristics of UEs* and *2) have a high chance of discovering better clean test accuracy when its parameters are perturbed.* In lieu of directly using classifiers trained by UEs as surrogates, we design a better way to construct surrogates meeting these two criteria.

In the following sections, we divide our method into two major parts. In the first part, we will introduce how to certify the learnability of a dataset through a surrogate classifier. Subsequently, we will show how PAP noises and a surrogate can be constructed to make more sound learnability certification in the UE regime. At the same time, an overview of our framework is summarized in Figure 2.

V. Learnability Certification

In this section, we will first describe the learnability certification process in detail. Certifying the learnability of a dataset over an unknown space of classifiers is a daunting task. However, it is possible to certify the learnability of a dataset towards classifiers whose parameters are in a particular set.

A. Quantile Parametric Smoothing

To effectively analyze classifiers in the space Θ, we first introduce a Quantile Parametric Smoothing (QPS) function in the following definition.

Definition 2 (*Quantile Parametric Smoothing function*)**.** *Given a dataset $\mathbb{D}$ from the space $\mathcal{X} \times \mathcal{Y}$, an $\mathbb{D}$-parameterized function $A_{\mathbb{D}} : \Theta \rightarrow [0, 1]$ with an input $\theta \in \Theta$, and a parametric smoothing noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ under a standard deviation of σ, a Quantile Parametric Smoothing function $h_q(\theta)$ is defined as:*

$$h_q(\theta) = \inf \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\theta + \epsilon) \leq t] \geq q\}, \tag{7}$$

where $q \in [0, 1]$ is a probability.

The definition connects to percentile smoothing used in certified robustness [32], [33]. In lieu of randomizing the inputs of classifiers, QPS takes the parameters (*i.e.,* weights) of classifiers as its variables and randomizes the parameters.

The QPS function has several nice properties for the task of learnability certification. First, it returns the accuracy score in the q-th quantile of the samples/population. By tuning the value q, QPS can naturally suit the need for computing the possibly highest accuracy that can be obtained by hypotheses, such that the learnability can be approximated. Second, QPS considers all hypotheses whose parameters are defined on the support of the Gaussian smoothing distribution (*i.e.,* the set $\mathbb{R}^d$ of real numbers, where d is the dimensionality of θ). This set of hypotheses covers almost all real-world classifiers with parameters as real numbers. However, most of the Gaussian probability mass concentrates around the mean, making it nearly impossible in practice to sample parameters far away from θ. We are thus motivated to consider the QPS function with varying mean values such that parameters from a more extensive space can be sampled.

B. Certified Learnability

To apply the QPS function to obtain a learnability certification, we explicitly define $A_{\mathbb{D}}$ as a performance metric in this section. Without loss of generality, such $A_{\mathbb{D}}(\hat{\theta})$ can be the top-1 accuracy function of a classifier $f_{\hat{\theta}}$ mapping its input x to a label y, where $(x, y) \in \mathbb{D}$. That is

$$A_{\mathbb{D}}(\hat{\theta}) := \mathbb{E}_{(x,y)\sim\mathbb{D}} \mathbb{1}[f_{\hat{\theta}}(x) = y]. \tag{8}$$

Note that the parameters of the classifier are the variables of $A_{\mathbb{D}}$. Subsequently, we can obtain $h_q(\hat{\theta})$ by substituting such $A_{\mathbb{D}}(\hat{\theta})$ into Equation 7. An empirical value of $h_q(\hat{\theta})$ can be approximately calculated by evaluating $A_{\mathbb{D}}(\hat{\theta}+\epsilon)$ multiple times given sampled ϵ values. Suppose $f_{\hat{\theta}}$ is a surrogate classifier trained on an unlearnable dataset $\mathbb{D}_s \oplus \delta$ by the defender, $h_q(\hat{\theta})$ indicates that, with probability at least q, there exists an accuracy upper bound for all classifiers $f_{\hat{\theta}+\epsilon}$ whose parameters are from the support of $\mathcal{N}(\hat{\theta}, \sigma^2 I)$.

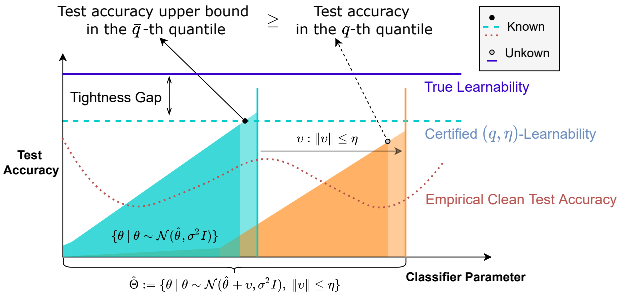
Test accuracy upper bound in the $\bar{q}$-th quantile
$\geq$
Test accuracy in the q-th quantile
● Known
Unkown
True Learnability
Tightness Gap
Certified (q, η)-Learnability
Test Accuracy
$\upsilon : \|\upsilon\| \leq \eta$
Empirical Clean Test Accuracy
$\{\theta \mid \theta \sim \mathcal{N}(\hat{\theta}, \sigma^2 I)\}$
$\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta} + \upsilon, \sigma^2 I),\ \|\upsilon\| \leq \eta\}$
Classifier Parameter

Fig. 3: An illustration of the certified (q, η)-Learnability.

We further move to the case in which the parameters are sampled from Gaussians with varying means. Specifically, we sample parameters from $\mathcal{N}(\hat{\theta}+\upsilon, \sigma^2 I)$, where $\|\upsilon\| \leq \eta$. When η gets larger, we become more capable of sampling classifiers with diverse parameters. Given a sufficiently large η, if there still exists an upper bound of the accuracy scores, then this upper bound can be used as a certificate of the best clean test accuracy. In other words, we want to certify an upper bound of $h_q(\hat{\theta} + \upsilon)$, $\forall \upsilon : \|\upsilon\| \leq \eta$. Such upper bound is called the *certified* (q, η)*-Learnability* of $\mathbb{D}_s \oplus \delta$. η is a *certified parametric radius of mean* describing the parameter space in which (q, η)-Learnability of $\mathbb{D}_s \oplus \delta$ can be guaranteed with probability at least q. For simplicity, we refer to η as the *certified parametric radius* in the following paper. To this point, the task of learnability certification becomes clear.

Next, we aim to construct such certified (q, η)-Learnability with a closed form so that it can be computed. Importantly, the certification of (q, η)-Learnability is provided through the following theorem:

Theorem 1 (*Perturbation bound on QPS*)**.** *Let $\Gamma : \mathcal{X} \times \mathcal{Y} \rightarrow \hat{\theta} \in \Theta$ be a learning function selecting $\hat{\theta}$ from the parameter space Θ based on a dataset defined in $\mathcal{X} \times \mathcal{Y}$. Given an target dataset $\mathbb{D}$ and a quantile smoothed function $h_q(\hat{\theta})$ centered at a Gaussian $\mathcal{N}(\hat{\theta}, \sigma^2 I)$, then there exists an upper bound for $h_q(\hat{\theta} + \upsilon)$. Specifically,*

$$h_q(\hat{\theta} + \upsilon) \leq \inf \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\hat{\theta} + \epsilon) \leq t] \geq \bar{q}\},\ \forall\ \|\upsilon\| \leq \eta, \tag{9}$$

where $\bar{q} := \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$. $\Phi(\cdot)$ is the standard Gaussian CDF and $\Phi^{-1}(\cdot)$ is the inverse of the CDF. $\|\upsilon\|$ is the ℓ_2 norm of the parameter shift υ from $\hat{\theta}$.

We defer the proof of Theorem 1 to Appendix B. The theorem states that the return of the QPS function is bounded with respect to the perturbation υ in its input. Recall that $h_q(\hat{\theta}+\upsilon)$ is actually an upper bound of the clean test accuracy for classifiers whose parameters $\theta \sim \mathcal{N}(\hat{\theta} + \upsilon, \sigma^2 I)$, with probability at least q. Therefore, the right-hand side of Inequality 9 actually guarantees the best possible clean test accuracy for classifiers with parameters from the subspace $\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta} + \upsilon, \sigma^2 I),\ \|\upsilon\| \leq \eta\}$. We name such $\hat{\Theta}$ as a *certified parameter set*. Though it seems that $\hat{\Theta}$ covers all parameters in $\mathbb{R}^d$, we only consider those θ having a sufficiently large probability of being sampled.

Based on Theorem 1, we have the formal definition of (q, η)-Learnability as follows.

Definition 3 (*(q, η)-Learnability*)**.** *Suppose a learning function Γ selects $\hat{\theta}$ based on an unlearnable dataset $\mathbb{D}_s \oplus \delta$. The certified (q, η)-Learnability of $\mathbb{D}_s \oplus \delta$ is*

$$l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta) = \inf \{t | \Pr_{\epsilon}[A_{\mathbb{D}}(\hat{\theta} + \epsilon) \leq t] \geq \overline{q}\}, \quad (10)$$

where $\overline{q} = \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$. For any θ^* drawn from the certified parameter set $\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta} + \upsilon, \sigma^2 I), \ \|\upsilon\| \leq \eta\}$, there is $A_{\mathbb{D}}(\theta^*) \leq l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$ with probability no less than q.

It should be noted that, though $\mathbb{D}_s \oplus \delta$ does not explicitly appear in the function, $\hat{\theta}$ is selected by a learning function $\hat{\theta} = \Gamma(\mathbb{D}_s \oplus \delta)$ given $\mathbb{D}_s \oplus \delta$. Henceforth, $\hat{\theta}$ can be viewed as a function of $\mathbb{D}_s \oplus \delta$ in this case. In practice, we can sample $A_{\mathbb{D}}(\hat{\theta} + \epsilon)$ using Monte Carlo and obtain the accuracy in the empirical $\overline{q}$-th quantile. Next, a Binomial confidence interval upper bound of the accuracy can be adopted as $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$ with a confidence level of $1 - \alpha$. Detailed steps of computing the confidence interval upper bound are in Appendix A.

Since $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$ is an upper bound only for accuracy obtained within $\hat{\Theta}$, there may exist a gap between such $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$ and the True Learnability. Specifically,

$$\triangle l = \mathbf{L}(\Theta; \mathbb{D}_s \oplus \delta) - l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta). \quad (11)$$

The gap $\triangle l \in [0, 1]$ represents the *tightness* of the certified (q, η)-Learnability. In order to make sense of using the certified (q, η)-Learnability as a measurement for UEs, the tightness gap should be minimized. Considering the concentration of the Gaussian probability mass, $\hat{\Theta}$ with a large η can have higher probability measures at locations far from $\hat{\theta}$, which helps reduce the tightness gap. A sketch describing the certified (q, η)-Learnability of a one-parameter classifier is illustrated in Figure 3. We will introduce the practical certification algorithm in the next section.

C. Certification Algorithm

We summarize the certification algorithm in this section. We first train a surrogate classifier $f_{\hat{\theta}}$ of which the parameters $\hat{\theta}$ will be randomized for n times over a Gaussian distribution to calculate the QPS function. For each randomization, the resulting classifier with parameters $\hat{\theta} + \epsilon$ is evaluated on the clean test set $\mathbb{D}$ such that $A_{\mathbb{D}}(\hat{\theta} + \epsilon)$ is obtained. Next, $\overline{q}$ can be theoretically computed based on Theorem 1. Finally, we compute the confidence interval of the $A_{\mathbb{D}}(\hat{\theta} + \epsilon)$ in the $\overline{q}$-th quantile to obtain the upper bound of the interval as $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$. The detailed certification process is depicted in Algorithm 1 and Algorithm 2. In addition, a meticulous description is in Appendix A.

There are a few important things to be noticed. A higher q is recommended for obtaining a learnability upper bound with high probability, and a larger η helps establish a more inclusive certified parameter set. In practice, both parameters can be optimized through grid search. Moreover, the σ of the parametric smoothing noise should be selected with caution. There is an intuition behind randomizing the surrogate weights and computing the QPS function — *some of the randomized classifiers can outperform the surrogate classifier on the clean test set*. Notwithstanding, we find that, when an improperly large σ is employed, the number of randomized classifiers having a higher accuracy than that of the surrogate drops significantly. The possible reason for this phenomenon could be that the shortcuts added to the UEs cannot fully conceal the mapping information between input samples and their labels. Hence, the surrogate trained on the unlearnable dataset, though not adequately, is somehow fit to the clean data distribution as well. Large noises added to the parameters thus not only shift the surrogate away from the distribution of UEs but also move it off the clean data manifold. This assumption is further verified by the recovery attack, which discovers classifiers with satisfactory performance on clean data in the vicinity of the poisoned classifier. Therefore, we will further introduce the methods for constructing sound surrogates for tighter certification.

Algorithm 1: Quantile Upper Bound

```
func QUpperBound
Input: noise draws $n$, $\alpha$, $\sigma$, $\eta$, quantile $q$.
Output: Index of the value in the $q$-th quantile
$\overline{q} \leftarrow \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$
$\underline{k}, \ \overline{k} \ \leftarrow \ \lceil n * \overline{q} \rceil, \ n$
$k^* \ \leftarrow \ 0$
for $k \in \{\underline{k}, \underline{k} + 1, ..., \overline{k}\}$ do
    if Binomial$(n, k, \overline{q}) > 1 - \alpha$ then
        $k^* \ \leftarrow k$
    else
        $Continue$
if $k^* \neq 0$ then
    Output: $k^*$
else
    Abstain
func Binomial
Input: Sampling number $n$, $k$, $\overline{q}$.
Conf $\leftarrow \ \sum_{i=1}^{k} \binom{n}{k} (\overline{q})^i (1 - \overline{q})^{n-i}$
Output: Conf
```

Algorithm 2: (q, η)-Learnability Certification

```
Input: Accuracy function $A_{\mathbb{D}}$, surrogate weights $\hat{\theta}$, test set $\mathbb{D}$, $n$, $\sigma$,
       $q$, $\eta$.
Output: $(q, \eta)$-Learnability
Initialize $a$
for $i \in 1, ..., n$ do
    $\theta \ \leftarrow \ \hat{\theta} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I)$
    Evaluate $A_{\mathbb{D}}(\theta)$ on $\mathbb{D}$
    Append $A_{\mathbb{D}}(\theta)$ to $a$
$a \ \leftarrow \ Sort(a)$
$k \ \leftarrow \ $QUpperBound$(n, \alpha, \sigma, \eta, q)$
$t \ \leftarrow \ a_k$
Output: $t$
```

D. Generalization of the Certification

In this section, we generalize the certification to the case where the defender has no access to the test dataset $\mathbb{D}$ and discuss more properties of the generalization certification.

As a solution, the defender can sample a test set $\hat{\mathbb{D}}$ from the accessible domain $\mathcal{D}$ and certify a *generalization (q, η)-Learnability* which bounds the learnability on test sets drawn from $\mathcal{D}$. By Hoeffding's Inequality, $A_{\mathcal{D}}(\hat{\theta}) \leq A_{\hat{\mathbb{D}}}(\hat{\theta}) + \sqrt{\frac{1}{2N} \log(\frac{2}{\beta})}$ exists with probability at least $1 - \beta$. The right-hand side of the inequality can be used in the QPS function instead of $A_{\mathbb{D}}(\hat{\theta})$. Combining with the union bound, a certified generalization (q,η)-Learnability can be acquired as $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta) + \sqrt{\frac{1}{2N}\log(\frac{2n}{\beta})}$, where n is the number of noise draws. However, Hoeffding's Inequality does not account for factors that improve generalization learnability. Therefore, we explore alternative approaches to establish the relationship between generalization learnability and the key components involved in the certification process. Notably, classifiers sampled around a surrogate relate to PAC-Bayesian predictors (*e.g.,* Gibbs classifiers) whose parameters are randomly drawn from a distribution [34]. We can thus bound the expected generalization accuracy under random weight perturbation through the following corollary.

Corollary 1 (*Expected generalization accuracy under parametric smoothing noise*)**.** *Let $A_{\hat{\mathbb{D}}} : \Theta \times \mathcal{X} \times \mathcal{Y} \rightarrow [0,1]$ be an accuracy function of a hypothesis parameterized by $\hat{\theta} \in \Theta$ evaluated on a dataset $\hat{\mathbb{D}} \sim \mathcal{D}$. When $\hat{\theta}$ is under a random perturbation υ, with probability at least $1-\alpha$, we have:*

$$\mathbb{E}_\epsilon[A_{\mathcal{D}}(\hat{\theta}+\epsilon)] \geq \mathbb{E}_\epsilon[A_{\hat{\mathbb{D}}}(\hat{\theta}+\epsilon)] - \sqrt{\frac{\frac{\|\hat{\theta}\|^2}{\sigma^2} + \ln\frac{N}{\alpha}}{2(N-1)}}, \qquad (12)$$

where N is the size of $\hat{\mathbb{D}}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

The corollary indicates that the expected generalization accuracy over the clean data distribution is lower bounded by the expected empirical accuracy minus a function of $\hat{\theta}$, σ and N. Although the corollary bounds the expectation rather than the q-th quantile, it connotes that a $f_{\hat{\theta}}$ constantly producing correct predictions on $\hat{\mathbb{D}}$ also has a high generalization accuracy, which evinces the inferiority of the UEs resulting $\hat{\theta}$. Furthermore, if the expectation of $A_{\hat{\mathbb{D}}}(\hat{\theta}+\epsilon)$ is robust to varying σ, the expected generalization accuracy would grow with increasing σ. For appropriately skewed accuracy distributions, this result implies that an $f_{\hat{\theta}}$ which endures large parametric noise can certify (q,η)-Learnability that generalizes better to $\mathcal{D}$.

Remark. We defined the QPS function of accuracy scores obtained by a set of classifiers randomized from a surrogate classifier fit to UEs $\mathbb{D}_s \oplus \delta$. The QPS function returns the value in the q-th quantile of all accuracy scores. The output of the QPS function can be further upper bounded by Theorem 1. The upper bound of the accuracy in the $\overline{q}$-th quantile is treated as the (q,η)-Learnability of $\mathbb{D}_s \oplus \delta$. The key message of the certified (q,η)-Learnability is that — if an unauthorized classifier has its trained parameters in the certified parameter set $\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta}+\upsilon, \sigma^2 I),\ \|\upsilon\| \leq \eta\}$, then its clean test accuracy will not exceed the (q,η)-Learnability, with probability at least q. Moreover, if $\mathbb{D}$ is hidden from the defender, a generalization (q,η)-Learnability can be certified by sampling a test set from $\mathcal{D}$ by the defender.

VI. Provably Unlearnable Examples

The core tasks of this section are 1) selecting surrogate classifiers that help certify a tighter (q,η)-Learnability and 2) generating PUEs to suppress (q,η)-Learnability against CAs while empirically tackling GAs.

A. Desiderata for Surrogates and PUEs

Recall that we rely on a surrogate classifier fit to the UEs to sample the accuracy scores $A_{\mathbb{D}}(\hat{\theta}+\epsilon)$. *The aims of the surrogate are to 1) produce as tight as possible certified (q,η)-Learnability scores and 2) capture the distributional characteristics of the UEs it is trained on.* Through parametric randomization, a plausible surrogate should discover as many as possible neighboring classifiers that have higher clean test accuracy. The best case of unauthorized classifiers can thus have a higher chance of being sampled. This means the noise added to the surrogate parameters should be large enough such that the probability measure of these potential best cases during sampling can be significant. Conversely, an immediate requirement for this type of surrogate classifier is that it should be able to tolerate large parametric noises while making predictions. Furthermore, the second target can be fulfilled when the surrogate has a low classification error on the set of UEs.

As the surrogate for certification, $f_{\hat{\theta}}$ should exhibit a low classification error on $\mathbb{D}_s \oplus \delta$, signifying its successful capture of the UE distribution. Second, the recovery attack reveals that the manifold of UEs can be close to that of clean samples. Similar to the proof in Appendix F of the paper by Cohen et al. [21], augmenting $f_{\hat{\theta}}$ with parametric noise maximizes the log-likelihood of correct predictions. This augmentation can tighten the certified (q,η)-Learnability.

Suppose there is a tightly certified (q,η)-Learnability, the task of making PUEs is straightforward — we will find better perturbations δ such that the surrogate $\hat{f}_\theta$ trained on $\mathbb{D}_s \oplus \delta$ can produce a lower $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$. At a high level, the key intuition behind crafting PUEs that can suppress its (q,η)-Learnability is described as follows:

A set of plausible UEs should lead to a trained classifier having not only a low clean test accuracy by itself but also low accuracy scores by the population of randomized classifiers in its vicinity.

Thereafter, we will show in the next section that the design goals of the surrogate and UEs are attainable under a unified training framework.

B. Random Weight Perturbation

To meet the first goal, we seek certification surrogates that can generate more correct predictions through randomizations under the parametric noise. Motivated by the findings in recovery attacks, such surrogates can be trained by using train-time random parametric noise as an augmentation. There are two possible ways to train surrogates. First, an offline surrogate can be trained on already made $\mathbb{D}_s \oplus \delta$. Otherwise, in an online setting, a surrogate can be trained together with δ.

In the online settings, we can optimize the PAP noise δ with the surrogate under random weight perturbations. It is hard to exactly minimize $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$. However, recall that when the ratio of small-valued $A_{\mathbb{D}}(\theta+\epsilon)$ grows, a lower value can be obtained in the fixed q-th quantile of accuracy scores. Therefore, PAP noises simultaneously updated with a surrogate over random weight perturbations may curtail the

(q, η)-Learnability of the resulting UEs. Given a dataset $\mathbb{D}_s$, we train additive error-minimizing noises and a surrogate classifier $f_{\hat{\theta}}$ under random weight perturbations. The objective of the optimization problem can be formulated as follows:

$$\begin{aligned} &\min_{\hat{\theta}} \mathbb{E}_{(x,y \sim \mathbb{D}_s)} \min_{\delta' \subseteq \delta} \mathbb{E}_{\epsilon \sim \pi(0)} \mathcal{L}(f_{\hat{\theta}+\epsilon}(x+\delta'), y), \\ &\quad s.t. \ \forall\ \delta' \subseteq \delta,\ \|\delta'\| \leq \xi, \end{aligned} \tag{13}$$

where δ' is the corresponding noise added to x and $\pi(0)$ is a train-time noise distribution with a mean of 0. In practice, we sample a set of U random weight perturbations with a zero-mean Gaussian distribution, to train δ and $\hat{\theta}$. In the offline setting, we omit the inner minimization objective on δ and only optimize $\hat{\theta}$. The training on the dataset $\mathbb{D}_s$ is thus to minimize the following empirical loss:

$$\min_{\hat{\theta}} \frac{1}{N} \sum_{j=1}^{N} \min_{\delta_j, \|\delta_j\| \leq \xi} \frac{1}{U} \sum_{i=1}^{U} \mathcal{L}[f_{\hat{\theta}+\epsilon_i}(x_j + \delta_j), y_j], \tag{14}$$

where (x_j, y_j) is sampled from $\mathbb{D}_s$ and $\epsilon_i \sim \mathcal{N}(0, {\sigma'}^2 I)$ is the i-th noise added to $\hat{\theta}$. In each step of training, we gradually increase the value of σ' with a step size of s to update $\hat{\theta}$ multiple times such that the classifier can better converge. The optimized perturbations δ are then superposed on $\mathbb{D}_s$ to make it unlearnable. The detailed algorithm for making the PUEs is depicted in Algorithm 3. More details of the optimization process are in Appendix A.

In the next section, we will evaluate the certified (q, η)-learnability and its tightness for PUEs, as well as other state-of-the-art UEs. Moreover, we will compare the empirical test accuracy and robustness of PUEs with those of competitors to demonstrate the effectiveness of PUEs against GAs.

Remark. PUEs aim to increase the classification error of a set of randomized classifiers. This ensures that the PUEs are effective towards not only the surrogate but also other classifiers who are on the support of the smoothing noise distribution. In contrast, the surrogate classifier trained on PUEs should stay immune to the weight perturbations caused by parametric smoothing such that the decrease in the certified learnability can be attributed to the effect of PUEs rather than the weight perturbations. We solve this two-layer problem as a task of bi-level minimization over the data distribution and random weight perturbations.

VII. Experiments

We present the experimental results of the certified (q, η)-Learnability and empirical performance of PUEs against CAs and GAs. Through the experiments, we aim to answer three questions.

- Can random weight perturbations increase the tightness of certified (q, η)-Learnability?
- Is PUE an effective way of producing unlearnable datasets suppressing certified (q, η)-Learnability?
- How robust is PUE against GAs who escape from the certified parameter set $\hat{\Theta}$?

Accordingly, we conduct experiments to measure both the certified (q, η)-Learnability and the empirical performance of PUE and its baselines.

Algorithm 3: Training of $f_{\hat{\theta}}$ and δ.

```
Input: Surrogate classifier $\hat{f}_\theta$ with parameters $\theta$, training dataset
       $\mathbb{D}_s$, smoothing distribution $\mathcal{N}(0, \sigma^2 I)$, $U_{train}$, $U_{perturb}$,
       noise STD cap $S$, step size $s$, training batch number $N$, train
       step number $M$, stop error rate $\tau$, validation set $\mathbb{D}'_s$.
Output: PUE noise $\delta$ and surrogate parameters $\hat{\theta}$.
Initialize:
$\delta \leftarrow [0]^{|\mathbb{D}_s| \times d}$
$error \leftarrow 1e^9$
while $error \geq \tau$ do
    for $i \in 1, ..., M$ do
        Load a mini-batch $(x_i, y_i)$ from $\mathbb{D}_s$
        $l \leftarrow 0$
        $\sigma' \leftarrow 0$
        while $\sigma' \leq S$ do
            $\sigma' \leftarrow \sigma' + s$
            for $j \in 1, ..., U_{train}$ do
                $\epsilon_j \sim \mathcal{N}(0, {\sigma'}^2 I)$
                $l \mathrel{+}= \mathcal{L}[f_{\theta+\epsilon_j}(x_i + \delta), y_i]$
            $\theta \leftarrow \textsc{Train\_step}(\theta, l/U_{train})$
    for $i \in 1, ..., N$ do
        Load a mini-batch $(x_i, y_i)$ from $\mathbb{D}_s$
        $l \leftarrow 0$
        $\sigma' \leftarrow 0$
        while $\sigma' \leq S$ do
            $\sigma' \leftarrow \sigma' + s$
            for $j \in 1, ..., U_{perturb}$ do
                $\epsilon_j \sim \mathcal{N}(0, {\sigma'}^2 I)$
                $l \mathrel{+}= \mathcal{L}[f_{\theta+\epsilon_j}(x_i + \delta), y_i]$
            $\delta \leftarrow \textsc{Opt\_step}(x_i, \delta, l/U_{perturb})$
    $error \leftarrow \textsc{Eval}(\theta, \mathbb{D}'_s + \delta)$
$\hat{\theta} \leftarrow \theta$
Output: $\delta$, $\hat{\theta}$
```

Baselines. We consider both *online surrogates* and *offline surrogates* in the certification against CAs. We compare PUE with a tailored version of Error-Minimizing Noise (EMN) [9] in their certified (q, η)-Learnability to show the effectiveness of our online surrogate and the robustness of PUE. Moreover, we train offline surrogates on off-the-shelf PUE, EMN, and OPS [35] to compare their certified (q, η)-Learnability. When it comes to GAs, we conduct a series of empirical comparisons with EMN and OPS to show their robustness and clean test accuracy scores in cases where unauthorized classifiers escape from the certified parameter set $\hat{\Theta}$. Though our method does not restrict the type of noise applied to make data unlearnable (*i.e.,* class-wise or sample-wise), we found that PUE surrogates trained with class-wise noises can converge rapidly while those trained by sample-wise noises do not converge very well. Therefore, in this section, we focus on class-wise noise in all experiments and comparisons.

Data and models. In our experiments, we adopt CIFAR-10, CIFAR100, and ImageNet with 100 randomly selected categories as the datasets in our evaluation. ResNet-18 is used as the surrogate model to generate UEs in all experiments. During certification experiments, we use ResNet-18 as the architecture of surrogates. We certify against 5000 ImageNet test set samples and the entire test datasets of CIFAR-10/CIFAR100. In empirical evaluations, we employ ResNet-18, ResNet-50, and DenseNet-121 as the architectures for unauthorized classifiers. We test the empirical robustness of PUE, EMN, and OPS on ResNet-18 using recovery attacks. The test accuracy scores of PUE, EMN, and OPS are measured by training ResNet-18, ResNet-50, and DenseNet-121 on them, respectively.

Training setups. In each iteration of PUE optimization and online/offline surrogate training, the standard deviation (STD) of the noises added to the parameters is gradually increased from 0 with a step size of 0.05 until the full noise level is reached. Furthermore, we start adding the weight perturbations after some warm-up steps to stabilize the training. Generally, we start perturbing weights after the classification error on UEs decreases below 50%. When generating UEs and fitting online surrogates, we set $M = 10$ on CIFAR10, $M = 20$ on CIFAR100, and $M = 100$ on ImageNet. No random data augmentation is used in the surrogate training processes. In empirical experiments, we follow the previous training routines [9], [35] and set the epoch to 60 when training classifiers or launching recovery attacks on CIFAR10. For the training and recovery attacks on CIFAR100 and ImageNet, the epoch number is set to 100. The weight decay rate is 5×10^{-4} on CIFAR10 and is 5×10^{-5} for CIFAR100/ImageNet. We apply an SGD with a momentum of 0.9 and a learning rate of 0.1 in all classifier/surrogate training runs.

Certification setups. We select 0.25 as the STD of the train-time Gaussian noise to obtain the surrogates. During the certification process, we set $q = 0.9$ and $\sigma \in \{0.25, 0.8\}$ for all the surrogates. There are reasons behind this setting. First, $\sigma = 0.25$ matches with that of the train-time noise, and can certify the highest (q, η)-Learnability at a fixed η (see Section VII-C). Second, $\sigma = 0.8$ is sufficient for certifying a parametric radius $\eta = 1.0$ within which the recovery attack can restore most of the clean accuracy (refer to Figure 5 in Section VII-B). For each surrogate classifier, we sample 1000 classifiers around it to compute the confidence interval of the QPS function under a confidence level of 0.99. For each surrogate, we certify under various values of η and record the certified learnability scores under each η. To compute the generalization (q, η)-Learnability, one can simply adding $\sqrt{\frac{1}{N}\log(\frac{2n}{\beta})}$ to the certified learnability scores given n, N and β. For instance, for $n = 1000$ and $\beta = 0.01$, sampling a dataset with size $N = 5000$ from the domain $\mathcal{D}$ as the test set to compute the generalization (q, η)-Learnability will attract a universal increase of 0.05 to the certified learnability scores in the tables. Each run of certification on CIFAR10/CIFAR100 takes around one GPU hour on Nvidia Tesla P100. Certifying an ImageNet classifier requires 26.4 GPU hours. We apply offsets to the certified results to mitigate the impact of accuracy differences in the surrogates. More details of hyper-parameters and certification settings can be found in Appendix C.

A. Learnability Certification

In this section, we study the certified (q, η)-Learnability of UEs for CAs and gauge the tightness of the certification using SAs (*i.e.,* recovery attacks).

Certification results. We compare certified learnability scores from two aspects. We first compare the certified learnability obtained based on surrogates trained on the same datasets but with different training strategies to show the effectiveness of random weight perturbations. Second, we compare the certified learnability of different PAP noises by using the same random weight perturbing strategy to train surrogates. This comparison aims to reveal the performance of PUE in reducing the certified learnability.

TABLE I: Certified (q, η)-Learnability under Different Training Methods (%, $\sigma = 0.25$)

Data	Method	$\eta \times 100$								
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0
CIFAR10	PUE-B	**10.62**	**10.67**	**10.71**	**11.07**	**11.86**	**12.50**	**13.20**	**14.67**	**15.75**
	EMN	5.69	5.70	5.74	5.91	6.24	6.43	6.96	8.61	10.27
CIFAR100	PUE-B	**1.32**	**1.32**	**1.33**	**1.37**	**1.41**	**1.47**	**1.53**	**1.59**	**1.68**
	EMN	0.43	0.43	0.44	0.47	0.52	0.59	0.70	0.77	0.89
ImageNet	PUE-B	**1.46**	**1.46**	**1.48**	**1.54**	**1.63**	**1.79**	**1.85**	**1.97**	**2.13**
	EMN	1.34	1.34	1.37	1.41	1.45	1.49	1.54	1.58	1.67

TABLE II: Certified (q, η)-Learnability under Different Training Methods (%, $\sigma = 0.8$)

Data	Method	η									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR10	PUE-B	**10.68**	**10.86**	**11.18**	**11.37**	**11.52**	**12.00**	**12.71**	**12.87**	**13.72**	**15.17**
	EMN	5.86	6.11	6.28	6.52	7.00	7.27	7.36	7.87	8.38	9.02
CIFAR100	PUE-B	**1.13**	**1.15**	**1.16**	**1.20**	**1.23**	**1.27**	**1.31**	**1.36**	**1.51**	**1.56**
	EMN	0.47	0.48	0.51	0.55	0.59	0.63	0.66	0.67	0.69	0.70
ImageNet	PUE-B	**1.19**	**1.24**	**1.26**	**1.28**	**1.32**	**1.37**	**1.45**	**1.50**	**1.61**	**1.80**
	EMN	1.12	1.14	1.16	1.20	1.25	1.29	1.40	1.40	1.48	1.48

To make the comparisons, we introduce a baseline surrogate, PUE-B, which employs random weight perturbations in the training of the surrogate parameters but does not perturb the weights when optimizing the PAP noise. Therefore, PUE-B differs from EMN only in the training method of the surrogate and can be used as the baseline in the evaluation of PUEs. On the other hand, the surrogate trained without random weight perturbation is denoted as EMN. The comparison is made under two different levels of smoothing noise ($\sigma = 0.25$ and $\sigma = 0.8$). The results are presented in Table I and Table II. According to Table I, compared to EMN, PUE-B can be certified at a higher $(0.9, \eta)$-Learnability across all datasets. It can also be observed that the gain in the $(0.9, \eta)$-Learnability of PUE-B increases when the certified radius rises. On CIFAR100 and ImageNet, the gain becomes more obvious. The observation supports the effectiveness of the random weight perturbation strategy in reducing the gap between the (q, η)-Learnability and the True Learnability.

Next, we compare the certified learnability of online surrogates trained by random weight perturbations on different UEs. We generated two versions of PUEs (*i.e.,* PUE-1 and PUE-10) by setting the value of $U_{perturb}$ to 1 and 10, respectively. The certification results based on the corresponding online surrogates are recorded in Table III and Table IV. Generally, in all the tables, PUE-10 achieves the lowest $(0.9, \eta)$-Learnability. When the smoothing noise level matches the noise level of random weight perturbation, PUE-10 can diminish the $(0.9, \eta)$-Learnability on all datasets. However, when the parametric smoothing noise gets larger to 0.8, PUE-10 has the best results at most of the certified radii across the three datasets but not for some cases at large η values. This should be the consequence of the large smoothing noise used, which will be verified by the ablation study in Section VII-C. Saliently, all of the results on PUEs outperform that of PUE-B, meaning that optimizing the PAP noise over random weight perturbations can effectively suppress the $(0.9, \eta)$-Learnability.

At last, we compare certification results based on offline

TABLE III: Certified (q, η)-Learnability under Different PAP Noises (%, $\sigma = 0.25$, online)

Data	Method	$\eta \times 100$								
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0
CIFAR10	PUE-10	**10.24**	**10.29**	**10.31**	**10.69**	**11.14**	**11.82**	**12.29**	**13.12**	**13.59**
	PUE-1	10.86	10.97	11.04	11.62	12.12	12.64	13.35	14.22	14.66
	PUE-B	10.62	10.67	10.71	11.07	11.86	12.50	13.20	14.67	15.75
CIFAR100	PUE-10	**1.29**	**1.29**	**1.31**	**1.35**	**1.41**	**1.43**	**1.46**	**1.51**	**1.65**
	PUE-1	1.35	1.36	1.36	1.42	1.48	1.53	1.57	1.69	1.87
	PUE-B	1.32	1.32	1.33	1.37	**1.41**	1.47	1.53	1.59	1.68
ImageNet	PUE-10	**1.45**	**1.45**	**1.45**	**1.50**	**1.61**	**1.68**	**1.76**	**1.84**	**1.95**
	PUE-1	1.58	1.58	1.58	1.67	1.73	1.81	1.95	2.00	2.19
	PUE-B	1.46	1.46	1.48	1.54	1.63	1.79	1.85	1.97	2.13

TABLE IV: Certified (q, η)-Learnability under Different PAP Noises (%, $\sigma = 0.8$, online)

Data	Method	η									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR10	PUE-10	**10.10**	**10.24**	**10.42**	**10.80**	**11.10**	**11.57**	**11.89**	**12.45**	**12.81**	13.72
	PUE-1	10.60	10.73	11.07	11.21	11.38	11.67	12.04	12.71	13.00	**13.59**
	PUE-B	10.68	10.86	11.18	11.37	11.52	12.00	12.71	12.87	13.72	15.17
CIFAR100	PUE-10	**1.11**	**1.14**	**1.16**	**1.20**	**1.23**	**1.25**	**1.27**	**1.35**	1.48	**1.52**
	PUE-1	1.13	1.15	1.18	1.21	1.24	1.27	1.31	**1.35**	**1.44**	1.60
	PUE-B	1.13	1.15	**1.16**	**1.20**	1.23	1.27	1.31	1.36	1.51	1.56
ImageNet	PUE-10	**1.17**	**1.19**	**1.24**	**1.26**	**1.28**	**1.32**	**1.37**	**1.45**	**1.54**	**1.61**
	PUE-1	1.22	1.24	1.30	1.35	1.37	1.41	1.45	1.52	**1.54**	1.68
	PUE-B	1.19	1.24	1.26	1.28	1.32	1.37	1.45	1.50	1.61	1.80

surrogates trained on CIFAR10 and CIFAR100 with PAP noises. We find surrogates trained on CIFAR100 OPS cannot be augmented properly within the same training epochs used for PUE and EMN. We thus only compare with OPS on CIFAR10. The certified learnability scores are presented in Table V and Table VI. We have similar observations on the certification results based on online surrogates and offline surrogates. According to the results, PUEs certified based on offline surrogates outperform both EMN and OPS when $\sigma = 0.25$. When $\sigma = 0.8$, OPS performs better on CIFAR10 at small η. Similarly, EMN obtains slightly lower certified learnability scores at $\eta = 0.7$. We suspect it is a consequence of the large smoothing noise, which leads to a loose certification. Additionally, the advantage of OPS can be a result of insufficient steps of training since classifiers usually require more epochs to converge on OPS. However, PUE still outperforms OPS when η exceeds 0.4.

Tightness of the certification. It is important to make sense of the gap between certified learnability and True Learnability. Since True Learnability is impossible to calculate, we gauge it by launching recovery attacks against surrogates used in the certification. Specifically, we use projected SGD to finetune the online surrogates of PUE-10 and EMN, respectively, on clean samples and clip their weights to make the after-trained weights stay within an ℓ_2 norm of η from the original weights. We use 20% of the CIFAR10/CIFAR100 training set in the fine-tuning to obtain the generalized clean test accuracy and use the CIFAR10/CIFAR100 test set to approximate the best accuracy. The learning rate of projected SGD is set to 0.01 in all attacks. The clean test accuracy scores recovered by the attack under different η are plotted in Figure 4. On the CIFAR10 EMN surrogate, the generalized test accuracy can be recovered from below 15% to near 60% within an η of 2.0 while the best possible accuracy is over 85%. In contrast, the recovery attack against the CIFAR10 PUE surrogate can only restore the accuracy from 10.68% to 12.98% and, in the best case, 15.31%. The best accuracy is close to the certified learnability scores, meaning tighter learnability is certified using the PUE surrogate.

TABLE V: Certified (q, η)-Learnability under Different PAP Noises (%, $\sigma = 0.25$, offline)

Data	Method	$\eta \times 100$								
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0
CIFAR10	PUE-10	**10.57**	**10.58**	**10.58**	**10.83**	**11.10**	**11.39**	**11.68**	**12.22**	**13.20**
	EMN	11.20	11.26	11.28	11.48	11.76	12.23	12.55	13.01	13.43
	OPS	10.59	10.61	10.64	10.96	11.37	11.73	11.98	12.54	13.38
CIFAR100	PUE-10	**1.13**	**1.13**	**1.13**	**1.16**	**1.21**	**1.27**	**1.31**	**1.35**	**1.47**
	EMN	1.16	1.16	1.17	1.20	1.25	1.31	1.37	1.43	1.60

TABLE VI: Certified (q, η)-Learnability under Different PAP Noises (%, $\sigma = 0.8$, offline)

Data	Method	η									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR10	PUE-10	**10.65**	10.86	11.06	**11.33**	**11.52**	**11.64**	**11.97**	**12.17**	**12.69**	**13.43**
	EMN	11.56	11.74	11.96	12.36	12.48	12.90	13.15	13.58	14.02	14.33
	OPS	**10.65**	**10.83**	**10.98**	11.36	11.55	11.72	12.02	12.24	13.59	13.65
CIFAR100	PUE-10	**1.13**	**1.15**	**1.17**	**1.20**	**1.25**	**1.30**	1.36	**1.38**	**1.39**	**1.42**
	EMN	1.14	1.17	1.22	1.26	1.29	1.34	**1.34**	**1.38**	1.43	1.46

Compared to CIFAR10, CIFAR100 surrogates demonstrate lower best accuracy and recovered accuracy, which could be imputed to two possible reasons. First, CIFAR100 is by nature more difficult to learn than CIFAR10. Second, the CIFAR100 surrogate is trained by more iterations on the UEs and is harder to recover. However, it can be seen that the PUE surrogate still outperforms the EMN one on the tightness of certification. The EMN surrogate has the best accuracy ranging from 1.03% to 13.69% when η increases, while the accuracy of the PUE surrogate can only be recovered to 1.40% in the best case. The comparison shows the superiority of the PUE surrogate in certifying tighter (q, η)-Learnability. We also examine the impact of surrogate architecture towards the tightness of the certified (q, η)-Learnability. Please refer to Appendix C for details. In addition, we check the validity of the certification under the stochasticity of different training runs to show that classifiers trained by adversaries can still fall inside the surrogate-certified $\hat{\Theta}$.

Validity of the certification. Recall that the validity of certified (q, η)-learnability is characterized by a certified parameter set. However, the training of classifiers faces stochasticity in weight initialization, mini-batch loading, stochastic gradient descent, etc. This implies that training on PUEs could lead to dissimilar model weights across different runs, so the PUEs that were certified by the surrogate classifier may not remain valid in subsequent training runs conducted by adversaries. To investigate the validity of PUEs for adversaries using mainstream training techniques, we train 10 ResNet-18 models on CIFAR10 PUEs with random weight initialization, batch loading orders, and data augmentations to check how diverse their trained weights can be. After training, we draw ten layers (nine convolution layers and one linear layer) and visualize the distribution of the converged weights. The visualization can be found along with our source code at https://github.com/Provably-Unlearnable-Examples/PUE. Furthermore, we calculate the pairwise difference for each parameter across the 10 classifiers and record the mean and STD of all the differences. According to the results, the weights from the 10 runs share similarities in their distributions, with a mean parameter difference of -4.92×10^{-6} and an STD of 0.01. Note that the certified

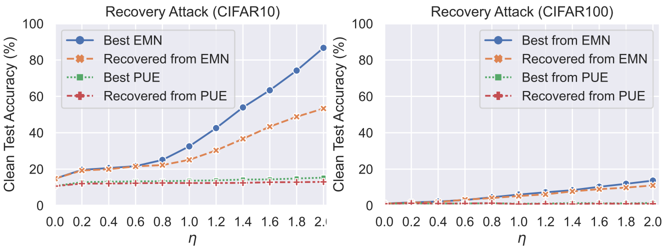

Fig. 4: The clean test accuracy scores of surrogate classifiers gauged on clean CIFAR10 (left) and CIFAR100 (right) test set after recovery attacks. The best accuracy scores approximate the True Learnability of parameters inside the hypersphere centered at $\hat{\theta}$ with a radius of η.

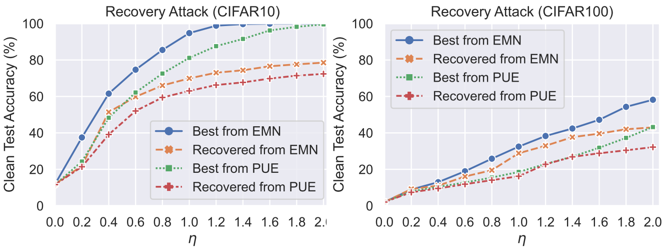

Fig. 5: The robustness of PUE and EMN against recovery attacks.

parameter set $\hat{\Theta}$ is actually an infinite Gaussian mixture. For example, with $\sigma = 0.25$ and $\eta = 1.0$, $\hat{\Theta}$ has sufficiently large probability mass within $[\hat{\theta} - 0.5, \hat{\theta} + 0.5]$ for $\hat{\theta}$ with a dimensionality of 11M (*i.e.,* the rough parameter count of ResNet-18) and beyond, which suggests that the certified parameter set can cover classifier weights from different training runs with stochasticity. Therefore, $\hat{\Theta}$ certified by a surrogate $\hat{\theta}$ can effectively capture classifiers trained separately by adversaries following certain standard training procedures.

B. Robustness of PUE against GAs

In this section, we aim to understand how robust PUEs are when the classifier trained on them is outside of the certified parameter set $\hat{\Theta}$ or evades the certified learnability scores with a probability of $1-q$. This setting aligns with real-world cases in which unauthorized classifiers are distinct from the surrogates used in the certification.

We particularly focus on the robustness of UEs by showing some hardness results of recovery attacks against classifiers trained on PUE, EMN, and OPS, following common training routines. Furthermore, we evaluate the performance of PUE in defending against various training approaches to verify its practicality. Specifically, we compare the empirical clean test accuracy of unauthorized classifiers trained on PUE, EMN and OPS using four representative training strategies (*i.e.,* MixUp, CutOut, Fast Autoaugment, and Adversarial Training). The results are deferred to Appendix D.

Hardness results of recovery attacks. Based on the same rules of gauging the surrogates using recovery attacks, we measure the performance of recovery attacks on classifiers trained on PUE, EMN, and OPS. The classifiers are trained on CIFAR10/CIFAR100 with PAP noises. We recover the clean accuracy using 20% of clean training data and calculate the best possible accuracy by training the poisoned classifiers on the clean test set. Note that adversaries possessing 20% of clean data have a strong capability to restore the performance of the poisoned classifier, evaluating PAP noises under this setting can better reveal their robustness. Moreover, data augmentations (*i.e.,* random crop, flip, and rotation) are employed in all recovery attacks to align with training strategies in real-world scenarios.

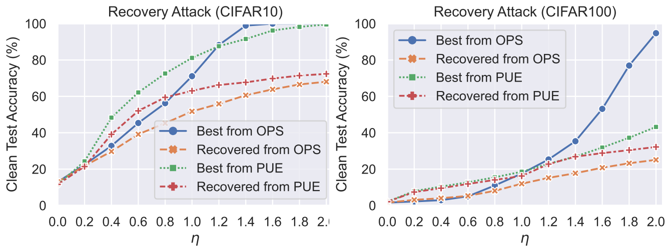

Fig. 6: The robustness of PUE and OPS against recovery attacks.

The comparison between EMN and PUE is presented in Figure 5. Compared to the results in Figure 4, the clean test accuracy of the normally trained classifiers can be better restored. On CIFAR10, the poisoned classifiers of PUE and EMN have initial accuracy scores of 11.62% and 11.89%, respectively. After the recovery, the EMN classifier has 78.55% accuracy and the PUE one has 72.37% accuracy. On the side of the best possible accuracy, EMN can achieve 100% and PUE reaches 99.55%. Similarly, the EMN classifier can be restored to 43.01% accuracy on CIFAR100, while the PUE classifier accuracy achieves 32.19%. However, the classifier trained on PUE demonstrates better robustness than the EMN classifier in both the best possible case and the recovery attack case. Most accuracy can be recovered within an η of 1.0, especially for CIFAR10. This is also why we certify at most an η of 1.0 in Section VII-A.

Aside from the above comparison, we notice several intriguing phenomena of PUE and OPS in Figure 6. OPS is a surrogate-free PAP noise. It searches for the pixel leading to the most significant change in the input data distribution, for each class in the dataset. Therefore, OPS is supposed to have higher robustness since it is model-agnostic and can be generalized to different classifiers based on experiments. It can be observed from the figure that, though OPS has a lower empirically recovered accuracy than PUE, its best recoverable accuracy gets higher than that of PUE rapidly when η grows. This implies that OPS has remarkable local robustness, possibly due to the effective manipulation of the input data distribution, which makes it hard to find proper clean training samples to reduce the clean generalization error of the trained classifier. However, when a set of plausible training samples is selected (*e.g.,* the Best from OPS case) for the recovery attack, OPS can be more vulnerable than other PAPs. Specifically, PUEs reduce at most 54.4% of the accuracy on CIFAR100. The weakness becomes more obvious when the perturbations/uncertainty in the weights increase beyond a certain threshold.

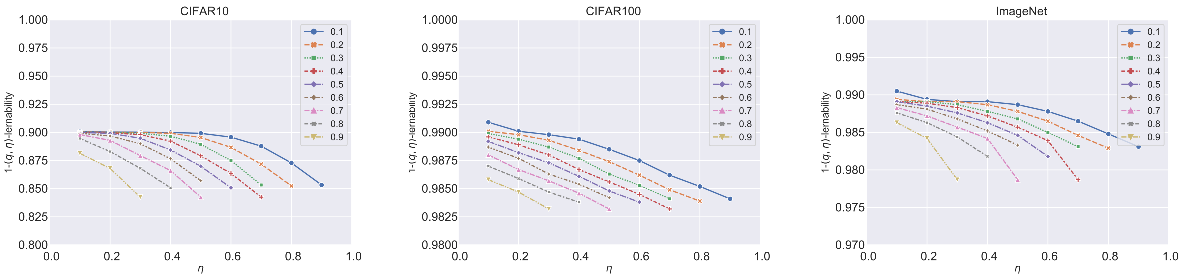

Fig. 7: The (q, η)-Learnability scores under different values of q are plotted on three data distributions (*i.e.,* CIFAR10, CIFAR100, and ImageNet). At the same η, a greater probability q can certify a higher (q, η)-Learnability.

The comparisons suggest that training GAs on PUE can lead to more robust poisoning results, especially against best-case recovery attacks. We attribute this to two reasons. First, theoretically, $\hat{\Theta}$ occupies a non-zero probability measure on every infinitesimal region on $\mathbb{R}^d$. Therefore, these empirically trained classifiers have non-zero probabilities of being included in $\hat{\Theta}$. In practice, a larger certified parametric radius η helps increase the probability measure at classifiers far from the surrogate. Second, the loss function 13 encourages PUE to reduce the accuracy of a collection of classifiers with Gaussian-modeled uncertainty in their parameters. Consequently, the collective robustness of PUE towards classifiers showing uncertainties in their parameters is enhanced.

C. Ablation Study

We conduct an ablation study on the trade-off between the STD σ of the smoothing noise and the learnability scores over different factors. First, we evaluate how the train-time noise during PUE surrogate optimization affects the trade-off. We train PUE-B surrogates on the CIFAR10/CIFAR100 datasets under different train-time noises and measure their certified learnability under smoothing noises with various σ values. We fix the value of q at 0.9 in all cases. The results on CIFAR10 and CIFAR100 are in Table VII and Table VIII of Appendix C, respectively. The certification returns ABSTAIN when a valid binomial confidence interval cannot be found. The abstentions are marked by "$-$" in the tables. The highest certified learnability under each η is highlighted.

The numbers in Table VII provide some intriguing observations about the trade-off. First, the largest certifiable parametric radius η positively correlates with σ. However, large σ values result in decreases in the (q, η)-Learnability scores at small certified radii, which indicates that the negative impact of the smoothing noise towards the tightness gap is significant in these cases. On the train-time noise side, generally, greater train-time noises make the resulting surrogates more endurable to large smoothing noises in the certification. Thus, the certification performs better in large-σ cases and produces better average certified learnability over different η values. Generally, when the train-time noise matches the σ value, the highest certified learnability scores can be found. For example, certifying using $\sigma = 0.25$ obtained the highest (q, η)-Learnability score when the STD of the train-time noise is 0.25. We observe similar trends in the CIFAR100 certification. This phenomenon suggests that the train-time noise should match the parametric smoothing noise to reduce the tightness gap.

Next, we evaluated the (q, η)-Learnability under different q values to see how the certified learnability can be affected. We select 0.25 as the STD for the train time noise and the parametric smoothing noise in this study. The learnability scores are illustrated in Figure 7. From the figure, certifying at a higher probability q compromises the maximal certifiable parametric radius across all datasets. However, the learnability scores certified at the same radius with larger q values are usually larger than those certified with small q values, which reveals tighter guarantees on the best cases unauthorized classifiers can achieve. At smaller parametric radii, the certified learnability is less distinguishable among different q values. Nevertheless, the discrepancy becomes obvious when η grows. In this paper, to ensure the (q, η)-Learnability can be certified at reasonable η values and be held with a sound probability, we select $q = 0.9$ in the certification experiments.

VIII. Discussion and Limitation

In the following part, we will discuss the limitations of our methods and explore potential avenues for improvement.

Certified parameter set. A crucial step for improving the certificate is to increase the certified parametric radius η such that the certificate can cover unauthorized classifiers with more diverse parameters. Based on the theory and the experiments, we can find a negative correlation between q and the maximum of certifiable η. This trade-off suggests that confident certifications can only be made at small η. While thoroughly breaking this trade-off seems impossible, it is possible to improve η by meticulously designing the parametric smoothing noise and the surrogate training method. In Theorem 1, the certification abstains if $\overline{q}$ is too large such that a valid Binomial confidence interval cannot be found. To avoid the abstention at large η, we should use a substantial σ since η scales with σ. Next, to make such a substantial σ functioning, we should have a surrogate classifier that can tolerate a significant level of smoothing noise while not degrading the tightness of the certification. Through experiments, we verify that larger parametric smoothing noise indeed leads to a broader certifiable parametric radius. Nevertheless, increasing the parametric radius η cannot be solely resorted to enlarging the smoothing noise since larger noise can also compromise the tightness of the (q, η)-Learnability. Specifically, too-large noises negatively impact the testing accuracy of the classifiers randomized from an insufficiently trained surrogate and make them fail to obtain plausible clean test accuracy scores. We propose a baseline solution in this paper to alleviate the negative effects of the parametric noise, but we reckon further endeavors can be made to increase η at high values of q.

Tightness gap of the certification. Given the noticeable connection between certified learnability and certified robustness, a possible way to further improve the tightness of the certification might be using a combination of adversarial training techniques, such as AWP [36] and SmoothAdv[37], to find the most critical perturbation direction for the weights while training the surrogate. In contrast, optimizing the PAP noise against these carefully augmented surrogates may lead to more robust UEs. Another possible pathway is to select better parametric smoothing noise. Though this simple uni-variance Gaussian noise used in this paper endorses a tidy form of (q, η)-Learnability, we suspect the tightness gap can be further reduced by introducing better smoothing noises. First, the probability density of Gaussian might not be optimal in discovering the best cases of classifiers. Intuitively, the optimal parametric smoothing noise should have a large probability of drawing parameters that can lead to the highest clean test accuracy. Second, the weights in different layers of a classifier may respond differently against the noise added to them, rendering uni-variance noises less efficient in surrogate augmentation and certification. A better option could be applying various levels of train-time noise in different layers to cope with the layer-wise noise sensitivities and using corresponding parametric smoothing noise in the certification.

Real-world practicality. Certified (q, η)-Learnability is a versatile framework that can be applied to different data modalities and tasks, as the certification holds in the parameter space and $A_{\mathbb{D}}(\hat{\theta})$ can represent any function that maps model parameters to a scalar value. As a proof of concept, we certify the unlearnable SST2 training set with Error-min-0 modifications [38] via an LSTM surrogate. A $(0.9, \eta)$-learnability of 0.61 can be certified at $\eta = 1$ while the accuracy at $\eta = 0$ is 0.52. Furthermore, the certification framework and PUE generation process may introduce additional computational overhead. We benchmark this overhead across various model architectures and datasets, as shown in Tables XIII and XIV. The overhead is confined to the certification and PUE generation stages and remains manageable. To reduce the cost, techniques such as perturbing only specific layers during PUE generation can be explored. Importantly, the application of PUE does not impact real-time data streaming, as PAP noise for each data category can be pre-generated and applied to corresponding streaming data. Finally, PUE requires a high perturbation rate, which means that a small proportion of PUEs in a clean dataset will not hinder the training process for authorized users. Legitimate users can also be provided with the PUE perturbations to restore data learnability. These features suggest that PUE is compatible with other defenses, such as model watermarking and unlearning, as the functionality of the watermark or retained set can be maintained.

IX. Related work

Perturbative availability poison and UEs. PAPs emerged as a response to unauthorized models trained to infer private datasets. There were early attempts that generate perturbations dynamically on surrogate models, which is computationally intensive [39], [10]. To alleviate the cost, error-minimizing perturbations were introduced to produce unlearnable data samples [9]. Subsequently, the training cost of the perturbation was further reduced by using pretrained surrogate classifiers or generalized neural tangent kernels [11], [8], [14], [15]. A series of works also resort to surrogate-free perturbations [40], [41], [42], [35]. In addition, enhancements were proposed to make transferable PAP noises or robust UEs against adversarial training [13], [43], [44]. We also notice some work making efforts to address the uncertainty of UEs [45], [16]. Particularly, the sharpness-aware method SAPA [16] leverages weight perturbations to approximate the worst-case model for the poison attack. However, the perturbation considered in SAPA is model- and data-dependent and may not be effective with variations in the type of loss function. Moreover, all the current methods fall short of offering a provable guarantee on the robustness of UEs against uncertainties in learning algorithms and adaptive adversaries in the field.

Certified robustness via randomized smoothing. Following previous certified robustness works using differential privacy and Rényi divergence [46], [47], randomized smoothing was first proposed by Cohen et al. as a tool for supplying a tight ℓ_2 robustness certificate for black-box functions, based on Neyman-Pearson Lemma (NPL) [21]. Though randomized smoothing places no restriction on its based classifiers, a line of work aimed at improving the base classifiers to obtain a better trade-off between the certified accuracy and the certified radius [37], [48], [49], [50], [51]. On the other hand, some trials have been made to design more capable smoothing noises, smoothing pipelines, and certification algorithms beyond NPL [22], [23], [24], [25], [26], [27], [28], [29]. Intriguingly, a series of works discovered that the maximum of certifiable ℓ_p $(p > 2)$ radius shrinks with the increasing input dimensionality [52], [30], [53], [54]. Moreover, randomized smoothing has been extended to poisoning defenses[55], [56], [57], [58], object detection [32], and watermark verification [33]. The works in this area focus on the certified robustness based on worst-case models rather than examining the certified learnability of datasets through the analysis of the best-case models trained on them.

X. Conclusion

In this paper, we propose a certification mechanism for deriving the possibly best clean data utility that unauthorized classifiers trained on UEs can achieve. The certified (q, η)-Learnability indicates that, with probability at least q, the unauthorized classifiers whose weights are from a specific parameter subspace have a guaranteed upper bound on their clean test accuracy. Moreover, we propose a way to craft provably unlearnable examples that obtain smaller (q, η)-Learnability scores, compared with the ones crafted by existing methods. Our certification mechanism takes the first step towards certifiably robust and provably effective UEs, which can provide a rigid guarantee on the protection level of PAP noises for data availability control. However, there exists a tightness gap between the certified (q, η)-Learnability and True Learnability. Such a tightness gap can be better alleviated in the future to make the certification more tenable. Crucially, such alleviation may also expand the space of certifiable parameters to cover greater parameter variance resulting from training techniques such as adversarial training. Finally, we reckon that future attempts can be made to extend PUE to sample-wise noises and reduce the training cost of making PUEs.

Acknowledgments

This work has been supported by the Cyber Security Research Centre Limited whose activities are partially funded by the Australian Government's Cooperative Research Centres Programme. Minhui Xue is supported in part by Australian Research Council (ARC) DP240103068 and in part by CSIRO – National Science Foundation (US) AI Research Collaboration Program.

References

[1] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council," 2016-05-04. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[2] C. S. Legislature, "California consumer privacy act of 2018," 2018. [Online]. Available: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

[3] E. Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021.

[4] T. W. House, "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence," 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[5] M. Heikkilä, "This artist is dominating ai-generated art, and he's not happy about it," 2022. [Online]. Available: https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/

[6] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022. [Online]. Available: arXivpreprintarXiv:2208.01618

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE SP*, 2017.

[8] L. Fowl, M. Goldblum, P.-y. Chiang, J. Geiping, W. Czaja, and T. Goldstein, "Adversarial examples make strong poisons," in *NeurIPS*, 2021.

[9] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable examples: Making personal data unexploitable," in *ICLR*, 2021.

[10] J. Feng, Q.-Z. Cai, and Z.-H. Zhou, "Learning to confuse: Generating training time adversarial data with auto-encoder," in *NeurIPS*, 2019.

[11] L. Tao, L. Feng, J. Yi, S.-J. Huang, and S. Chen, "Better safe than sorry: Preventing delusive adversaries with adversarial training," in *NeurIPS*, 2021.

[12] Z. Liu, Z. Zhao, and M. Larson, "Image shortcut squeezing: Countering perturbative availability poisons with compression," in *ICML*, 2023.

[13] S. Fu, F. He, Y. Liu, L. Shen, and D. Tao, "Robust unlearnable examples: Protecting data privacy against adversarial learning," in *ICLR*, 2022.

[14] C.-H. Yuan and S.-H. Wu, "Neural tangent generalization attacks," in *ICML*, 2021.

[15] S. Chen, G. Yuan, X. Cheng, Y. Gong, M. Qin, Y. Wang, and X. Huang, "Self-ensemble protection: Training checkpoints are good data protectors," in *ICLR*, 2023.

[16] P. He, H. Xu, J. Ren, Y. Cui, S. Zeng, H. Liu, C. Aggarwal, and J. Tang, "Sharpness-aware data poisoning attack," in *ICLR*, 2024.

[17] B. Zhu, L. Wang, Q. Pang, S. Wang, J. Jiao, D. Song, and M. I. Jordan, "Byzantine-robust federated learning with optimal statistical rates," in *AISTATS*, 2023.

[18] G. Wang, Z. Ma, C. Liu, X. Yang, H. Fang, W. Zhang, and N. Yu, "Must: Robust image watermarking for multi-source tracing," in *AAAI*, 2024.

[19] Y. Hu, F. Wu, Q. Li, Y. Long, G. Garrido, C. Ge, B. Ding, D. Forsyth, B. Li, and D. Song, "Sok: Privacy-preserving data synthesis," in *IEEE SP*, 2023.

[20] S. Australia, "As iso/iec 42001:2023," 2023.

[21] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019.

[22] M. Fischer, M. Baader, and M. Vechev, "Certified defense to image transformations via randomized smoothing," in *NeurIPS*, vol. 33, 2020, pp. 8404–8417.

[23] L. Li, M. Weber, X. Xu, L. Rimanic, B. Kailkhura, T. Xie, C. Zhang, and B. Li, "Tss: Transformation-specific smoothing for robustness certification," in *CCS*, 2021.

[24] Z. Hao, C. Ying, Y. Dong, H. Su, J. Song, and J. Zhu, "Gsmooth: Certified robustness against semantic transformations via generalized randomized smoothing," in *ICML*, 2022.

[25] P. Súkeník, A. Kuvshinov, and S. Günnemann, "Intriguing properties of input-dependent randomized smoothing," in *ICML*, 2022.

[26] A. Cullen, P. Montague, S. Liu, S. Erfani, and B. Rubinstein, "Double bubble, toil and trouble: Enhancing certified robustness through transitivity," in *NeurIPS*, 2022.

[27] K. D. Dvijotham, J. Hayes, B. Balle, Z. Kolter, C. Qin, A. Gyorgy, K. Xiao, S. Gowal, and P. Kohli, "A framework for robustness certification of smoothed classifiers using f-divergences," in *ICLR*, 2020.

[28] D. Zhang, M. Ye, C. Gong, Z. Zhu, and Q. Liu, "Black-box certification with randomized smoothing: A functional optimization based framework," in *NeurIPS*, 2020.

[29] H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, "Denoised smoothing: A provable defense for pretrained classifiers," in *NeurIPS*, 2020.

[30] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized smoothing of all shapes and sizes," in *ICML*, 2020.

[31] L. Valiant, *Probably Approximately Correct: Natureõs Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ), 2013.

[32] P.-y. Chiang, M. Curry, A. Abdelkader, A. Kumar, J. Dickerson, and T. Goldstein, "Detection as regression: Certified object detection with median smoothing," in *NeurIPS*, 2020.

[33] A. Bansal, P.-y. Chiang, M. J. Curry, R. Jain, C. Wigington, V. Manjunatha, J. P. Dickerson, and T. Goldstein, "Certified neural network watermarks with randomized smoothing," in *ICML*, 2022.

[34] D. McAllester, "Simplified pac-bayesian margin bounds," *Lecture Notes in Computer Science*, pp. 203–215, 2003.

[35] S. Wu, S. Chen, C. Xie, and X. Huang, "One-pixel shortcut: On the learning preference of deep neural networks," in *ICLR*, 2023.

[36] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," in *NeurIPS*, 2020.

[37] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," in *NeurIPS*, 2019.

[38] X. Li and M. Liu, "Make text unlearnable: Exploiting effective patterns to protect personal data," in *TrustNLP*, 2023.

[39] J. Shen, X. Zhu, and D. Ma, "Tensorclog: An imperceptible poisoning attack on deep neural network applications," *IEEE Access*, vol. 7, pp. 41 498–41 506, 2019.

[40] I. Evtimov, I. C. Covert, A. Kusupati, and T. Kohno, "Disrupting model training with adversarial shortcuts," in *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[41] D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu, "Availability attacks create shortcuts," in *KDD*, 2022.

[42] P. Sandoval-Segura, V. Singla, J. Geiping, M. Goldblum, T. Goldstein, and D. Jacobs, "Autoregressive perturbations for data poisoning," in *NeurIPS*, 2022.

[43] R. Wen, Z. Zhao, Z. Liu, M. Backes, T. Wang, and Y. Zhang, "Is adversarial training really a silver bullet for mitigating data poisoning?" in *ICLR*, 2022.

[44] J. Ren, H. Xu, Y. Wan, X. Ma, L. Sun, and J. Tang, "Transferable unlearnable examples," in *ICLR*, 2022.

[45] Y. Lu, G. Kamath, and Y. Yu, "Exploring the limits of model-targeted indiscriminate data poisoning attacks," in *ICML*, 2023.

[46] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *IEEE SP*, 2019.

[47] B. Li, C. Chen, W. Wang, and L. Carin, "Certified adversarial robustness with additive noise," in *NeurIPS*, 2019.

[48] R. Zhai, C. Dan, D. He, H. Zhang, B. Gong, P. Ravikumar, C.-J. Hsieh, and L. Wang, "Macer: Attack-free and scalable robust training via maximizing certified radius," in *ICLR*, 2020.

[49] J. Jeong and J. Shin, "Consistency regularization for certified robustness of smoothed classifiers," in *NeurIPS*, 2020.

[50] J. Jeong, S. Park, M. Kim, H.-C. Lee, D.-G. Kim, and J. Shin, "Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness," in *NeurIPS*, 2021.

[51] J. Jeong, S. Kim, and J. Shin, "Confidence-aware training of smoothed classifiers for certified robustness," in *AAAI*, 2023.

[52] A. Kumar, A. Levine, T. Goldstein, and S. Feizi, "Curse of dimensionality on randomized smoothing for certifiable robustness," in *ICML*, 2020.

[53] L. Li, J. Zhang, T. Xie, and B. Li, "Double sampling randomized smoothing," in *ICML*, 2022.

[54] Y. Shu, X. Xiao, D. Wang, Y. Cao, S. Chen, J. Xue, L. Li, and B. Li, "Effects of exponential Gaussian distribution on (Double sampling) randomized smoothing," in *ICML*, 2024.

[55] Y. Zhang, A. Albarghouthi, and L. D'Antoni, "Bagflip: A certified defense against data poisoning," in *NeurIPS*, 2022.

[56] M. Weber, X. Xu, B. Karlas, C. Zhang, and B. Li, "Rab: Provable robustness against backdoor attacks," in *IEEE SP*, 2022.

[57] A. Kumar, A. Levine, T. Goldstein, and S. Feizi, "Provable robustness against wasserstein distribution shifts via input randomization," in *ICLR*, 2023.

[58] Y. Zhang, A. Albarghouthi, and L. D'Antoni, "Pecan: A deterministic certified defense against backdoor attacks," 2023. [Online]. Available: arXivpreprintarXiv:2301.11824

[59] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[60] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[61] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017. [Online]. Available: arXivpreprintarXiv:1708.04552

[62] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," in *NeurIPS*, 2019.

APPENDIX

A. More Algorithmic Details

Computing (q, η)-Learnability. We introduce the detailed steps of computing (q, η)-Learnability. Please refer to Algorithm 1 and Algorithm 2 for the pseudocode.

First, according to Theorem 1, $\overline{q} = \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$ can be theoretically computed by supplying the values of q, η, and σ. Next, we sample a finite number of $A_{\mathbb{D}}(\hat{\theta} + \epsilon)$ using Monte Carlo. For simplicity, we denote the sampled and sorted $A_{\mathbb{D}}(\hat{\theta}+\epsilon)$ as a set $a = \{a_1, a_2, ..., a_n\}$, where $a_1 \leq a_2 \cdots \leq a_n$. Based on the order statistics a, the smallest accuracy $a_t \in a$ satisfying $\mathrm{Pr}_{a_i \in a}[a_i \leq a_t] \geq \overline{q}$ can be found as the accuracy in the empirical $\overline{q}$-th quantile (*i.e.,* a_t is the empirically calculated value of $h_{\overline{q}}(\hat{\theta})$). Note the (q, η)-Learnability should be calculated as the confidence interval upper bound of a_t. Calculating a one-sided confidence interval upper bound resorts to finding the k-th accuracy in a, such that by giving a confidence level $1 - \alpha$, there is $\Pr[a_t \leq a_k] \geq 1 - \alpha$. There is

$$\begin{aligned}\Pr[a_t \leq a_k] &= \sum_{i=1}^{k} \Pr[a_{i-1} \leq a_t \leq a_i] \\ &= \sum_{i=1}^{k} \binom{n}{i} (\overline{q})^i (1 - \overline{q})^{n-i}.\end{aligned}$$

Thereafter, we just find the smallest k letting

$$\sum_{i=1}^{k} \binom{n}{i} (\overline{q})^i (1 - \overline{q})^{n-i} \geq 1 - \alpha.$$

The k-th accuracy a_k is returned as the certified (q, η)-Learnability $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta)$.

Algorithms of surrogate training and PAP noise optimization. We attach the details of the surrogate training step and PAP noise optimization step in Algorithm 4 and Algorithm 5, respectively. The surrogate is updated by using gradient descent with a learning rate of r given $\frac{l}{U_{train}}$ from Algorithm 3. During optimizing the PAP noise δ, if using class-wise noise, the noise of each class is independently updated by optimizing using examples from the corresponding class. Specifically, suppose there are K classes of data in the mini-batch, for the i-th class, its corresponding noise $\delta[i]$ is updated as

$$\delta[i] \leftarrow Clip_{\pm\xi}\left(x[i] - r_p \cdot \mathrm{sign}(\frac{1}{U_{perturb}} \bigtriangledown_{\delta[i]} l[i])\right) - x[i],$$

where $x[i]$ are input samples belonging to the i-th class and $l[i] = \sum_{j=1}^{S/s} L[f_{\hat{\theta}+\epsilon_j}(x[i]), i]$ is the accumulated loss calculated from samples of the i-class in the mini-batch.

Algorithm 4: Training Step of Surrogate

```
func TRAIN_STEP
Input: $\theta$, $l/U_{train}$, learning rate $r$
$\theta \leftarrow \theta - \frac{r}{U_{train}} \cdot \bigtriangledown_{\theta} l$
Output: $\theta$
```

Algorithm 5: Optimization Step of PUE

```
func OPT_STEP
Input: $\delta$, $x$, $l/U_{perturb}$, learning rate $r_p$
$\delta \leftarrow Clip_{\pm\xi}\left(x - r_p \cdot \mathrm{sign}(\frac{1}{U_{perturb}} \bigtriangledown_{\delta}\, l)\right) - x$
Output: $\delta$
```

B. Proofs

Theorem 1 (*Perturbation bound on QPS*). *Let $\Gamma : \mathcal{X} \times \mathcal{Y} \rightarrow \hat{\theta} \in \Theta$ be a learning function selecting $\hat{\theta}$ from the parameter space Θ based on a dataset defined in $\mathcal{X} \times \mathcal{Y}$. Given an target dataset $\mathbb{D}$ and a quantile smoothed function $h_q(\hat{\theta})$ centered at a Gaussian $\mathcal{N}(\hat{\theta}, \sigma^2 I)$, then there exists an upper bound for $h_q(\hat{\theta} + \upsilon)$. Specifically,*

$$h_q(\hat{\theta} + \upsilon) \leq \inf\ \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\hat{\theta} + \epsilon) \leq t] \geq \overline{q}\},\ \forall\ \|\upsilon\| \leq \eta, \tag{9}$$

where $\overline{q} := \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$. $\Phi(\cdot)$ is the standard Gaussian CDF and $\Phi^{-1}(\cdot)$ is the inverse of the CDF. $\|\upsilon\|$ is the ℓ_2 norm of the parameter shift υ from $\hat{\theta}$.

Proof: The proof is similar to that of Lemma 2 in the paper of Chiang et al. [32]. For simplicity, notate the RHS of the inequality by $h_{\overline{q}}(\hat{\theta})$. Let

$$\lambda(\hat{\theta}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)} \mathbb{1}[A_{\mathbb{D}}(\hat{\theta} + \epsilon) \leq h_{\overline{q}}(\hat{\theta})],$$

there is $\lambda(\hat{\theta}) \in [0, 1]$. It is obvious that

$$\lambda(\hat{\theta}) = \Pr[A_{\mathbb{D}}(\hat{\theta} + \epsilon) \leq h_{\overline{q}}(\hat{\theta})].$$

TABLE VII: Trade-off between Train-Time/Certification Noise Level and Learnability (%) on CIFAR10

Train Noise	Certification Noise (σ)	$\eta \times 100$																						
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0	35.0	40.0	45.0	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0	95.0	100.0
0.10	0.10	**14.57**	**14.68**	**14.76**	**16.00**	**17.26**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	10.41	10.42	10.49	11.01	11.36	**12.00**	**12.66**	**13.57**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	10.49	10.52	10.53	10.68	11.00	11.23	11.63	11.75	12.16	**12.73**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	10.51	10.52	10.52	10.72	10.92	11.14	11.50	11.70	12.04	12.23	**12.77**	**12.88**	**13.49**	-	-	-	-	-	-	-	-	-	-
	0.50	10.67	10.69	10.69	10.81	11.12	11.24	11.62	12.01	**12.28**	12.44	12.60	12.75	12.89	13.07	13.11	13.11	-	-	-	-	-	-	-
	0.60	10.66	10.67	10.68	10.86	10.95	11.03	11.14	11.38	11.65	12.08	12.38	12.63	12.98	**13.07**	**13.45**	**13.46**	**13.75**	**14.80**	-	-	-	-	-
	0.70	10.51	10.51	10.53	10.60	10.74	10.82	10.95	11.12	11.19	11.32	11.45	11.63	11.70	11.80	11.99	12.28	12.57	12.77	**13.67**	**13.70**	**13.70**	-	-
	0.80	10.49	10.49	10.50	10.54	10.62	10.69	10.75	10.90	10.99	11.06	11.24	11.34	11.41	11.47	11.66	11.76	11.88	12.00	12.43	12.54	12.95	**12.95**	**13.18**
	0.90	10.50	10.50	10.53	10.56	10.62	10.70	10.78	10.91	10.98	11.06	11.18	11.29	11.35	11.56	11.77	11.94	12.06	12.29	12.33	12.58	12.58	12.81	13.01
	1.00	10.49	10.49	10.50	10.54	10.61	10.67	10.74	10.82	10.93	11.03	11.12	11.28	11.43	11.71	11.80	11.96	12.12	12.15	12.17	12.22	12.36	12.50	12.60
0.15	0.10	**13.75**	**13.85**	**14.04**	**15.17**	**16.59**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.15	10.53	10.60	10.62	11.26	11.99	**13.67**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	10.57	10.59	10.65	10.99	11.54	12.07	**13.14**	**15.20**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	10.49	10.49	10.54	10.73	11.03	11.39	11.82	13.00	**13.43**	**13.67**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	10.47	10.50	10.51	10.69	10.87	11.01	11.26	11.48	11.87	12.36	**12.55**	**13.10**	**13.92**	-	-	-	-	-	-	-	-	-	-
	0.50	10.44	10.44	10.47	10.58	10.69	10.82	11.09	11.39	11.61	12.28	12.34	12.61	12.78	**13.12**	**13.72**	**13.72**	-	-	-	-	-	-	-
	0.60	10.45	10.50	10.53	10.62	10.74	10.87	11.04	11.19	11.46	11.51	11,65	11.80	12.11	12.30	12.60	12.63	**13.55**	**14.10**	-	-	-	-	-
	0.70	10.40	10.40	10.42	10.49	10.57	10.66	10.78	10.96	11.05	11.20	11.32	11.43	11.63	11.78	12.04	12.39	12.58	12.72	**13.03**	**13.47**	**13.47**	-	-
	0.80	10.38	10.38	10.38	10.42	10.53	10.62	10.68	10.79	10.92	11.03	11.20	11.36	11.52	11.74	12.28	12.33	12.48	12.49	12.70	13.16	13.16	**13.16**	**13.74**
	0.90	10.42	10.42	10.43	10.51	10.56	10.62	10.71	10.78	10.92	10.97	11.03	11.16	11.19	11.25	11.39	11.67	11.81	12.06	12.12	12.17	12.20	12.38	12.78
	1.00	10.35	10.35	10.35	10.42	10.50	10.64	10.73	10.76	10.87	10.97	11.04	11.17	11.24	11.28	11.38	11.48	11.53	11.63	11.75	12.14	12.38	12.74	12.80
0.20	0.10	10.58	10.60	10.63	10.96	**13.70**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	**11.70**	**11.89**	**11.98**	**12.58**	13.38	**14.21**	**15.02**	**15.29**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	10.57	10.57	10.61	10.84	11.15	11.61	11.81	12.24	**12.77**	**12.95**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	10.42	10.42	10.44	10.65	10.90	11.10	11.22	11.31	11.81	11.96	**12.18**	**13.07**	**13.18**	-	-	-	-	-	-	-	-	-	-
	0.50	10.46	10.46	10.46	10.57	10.71	10.86	11.07	11.27	11.31	11.40	11.51	11.73	12.42	**12.59**	**14.54**	**14.54**	-	-	-	-	-	-	-
	0.60	10.42	10.44	10.46	10.55	10.64	10.77	10.89	10.96	11.09	11.18	11.34	11.46	11.68	12.02	12.33	12.38	**12.96**	**13.11**	-	-	-	-	-
	0.70	10.34	10.36	10.37	10.46	10.57	10.61	10.68	10.80	10.96	11.13	11.28	11.50	11.61	11.70	11.87	12.06	12.36	12.70	**13.00**	**14.18**	**14.18**	-	-
	0.80	10.40	10.40	10.44	10.47	10.61	10.73	10.76	10.91	10.97	11.08	11.22	11.30	11.47	11.61	12.04	12.17	12.20	12.57	12.74	13.08	13.19	13.19	13.23
	0.90	10.45	10.45	10.46	10.55	10.58	10.62	10.72	10.80	10.90	11.00	11.05	11.18	11.31	11.40	11.64	11.72	11.99	12.08	12.19	12.49	12.73	**13.30**	**13.48**
	1.00	10.43	10.43	10.44	10.49	10.54	10.61	10.67	10.78	10.92	10.99	11.04	11.08	11.23	11.34	11.36	11.44	11.54	11.73	11.81	11.99	12.44	12.61	12.76
0.25	0.10	9.96	9.96	9.96	10.00	10.98	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	10.02	10.24	10.27	10.58	10.98	11.85	12.69	14.49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.25	**10.62**	**10.67**	**10.71**	**11.07**	**11.86**	**12.05**	**13.20**	**14.67**	**15.75**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	10.35	10.35	10.40	10.69	10.98	11.32	11.98	12.57	12.95	**13.02**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	10.41	10.43	10.44	10.63	10.86	11.15	11.28	11.57	12.58	12.82	**13.03**	**13.18**	**14.21**	-	-	-	-	-	-	-	-	-	-
	0.50	10.40	10.41	10.42	10.53	10.73	10.88	11.17	11.45	11.60	11.76	11.92	12.22	12.73	**12.76**	**14.00**	**14.00**	-	-	-	-	-	-	-
	0.60	10.40	10.40	10.41	10.50	10.61	10.73	10.83	10.98	11.15	11.41	11.51	11.76	11.93	11.99	12.49	12.76	**13.08**	**15.25**	-	-	-	-	-
	0.70	10.47	10.50	10.52	10.57	10.72	10.83	10.96	11.02	11.14	11.19	11.31	11.55	11.77	11.90	12.13	12.19	12.38	12.55	**12.95**	**15.70**	**15.70**	-	-
	0.80	10.53	10.53	10.53	10.59	10.68	10.77	10.86	10.95	11.18	11.28	11.37	11.49	11.52	11.83	12.00	12.12	12.71	12.72	12.87	12.91	13.72	**13.72**	**15.17**
	0.90	10.52	10.52	10.53	10.59	10.66	10.77	10.88	10.96	11.09	11.27	11.40	11.47	11.51	11.52	11.85	12.21	12.23	12.51	12.90	13.10	13.36	13.62	13.96
	1.00	10.50	10.50	10.51	10.56	10.66	10.75	10.78	10.82	10.96	11.07	11.17	11.26	11.51	11.57	11.62	11.71	11.82	12.12	12.24	12.42	12.67	13.10	13.18

The following function

$$\begin{aligned}\Lambda(\hat{\theta}) &= \sigma\Phi^{-1}(\lambda(\hat{\theta}))\\ &= \sigma\Phi^{-1}(\Pr[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \le h_{\overline{q}}(\hat{\theta})])\end{aligned}$$

is 1-Lipschiz due to Lemma 2 of Salman et al. [37]. Therefore,

$$\begin{aligned}&\Phi^{-1}(\Pr[A_{\mathbb{D}}(\hat{\theta}+\upsilon+\epsilon) \le h_{\overline{q}}(\hat{\theta})])\\ \ge&\Phi^{-1}(\Pr[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \le h_{\overline{q}}(\hat{\theta})]) - \frac{\|\upsilon\|}{\sigma}.\end{aligned} \tag{15}$$

If and only if when $\|\upsilon\| \le \eta$, there is

$$\begin{aligned}&\Phi^{-1}(\Pr[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \le h_{\overline{q}}(\hat{\theta})]) - \frac{\|\upsilon\|}{\sigma}\\ \ge&\Phi^{-1}(\Pr[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \le h_{\overline{q}}(\hat{\theta})]) - \frac{\eta}{\sigma}\\ =&\Phi^{-1}(\overline{q}) - \frac{\eta}{\sigma}\\ =&\Phi^{-1}(q).\end{aligned} \tag{16}$$

Since the inverse CDF is monotonically increasing, combining Inequality 15 and Inequality 16 yields

$$\Pr[A_{\mathbb{D}}(\hat{\theta}+\upsilon+\epsilon) \le h_{\overline{q}}(\hat{\theta})] \ge q.$$

Note that, by definition,

$$h_q(\hat{\theta}+\upsilon) = \inf\ \{t \mid \Pr[A_{\mathbb{D}}(\hat{\theta}+\upsilon+\epsilon) \le t] \ge q\},$$

Then

$$\begin{aligned}&h_q(\hat{\theta}+\upsilon) \le h_{\overline{q}}(\hat{\theta})\\ \iff &h_q(\hat{\theta}+\upsilon) \le \inf\ \{t \mid \Pr[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \le t] \ge \overline{q}\}.\end{aligned}$$

The proof is concluded. ■

Corollary 1 (*Expected generalization accuracy under parametric smoothing noise*)**.** *Let $A_{\hat{\mathbb{D}}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be an accuracy function of a hypothesis parameterized by $\hat{\theta} \in \Theta$ evaluated on a dataset $\hat{\mathbb{D}} \sim \mathcal{D}$. When $\hat{\theta}$ is under a random perturbation υ, with probability at least $1-\alpha$, we have:*

$$\mathbb{E}_\epsilon[A_{\mathcal{D}}(\hat{\theta}+\epsilon)] \ge \mathbb{E}_\epsilon[A_{\hat{\mathbb{D}}}(\hat{\theta}+\epsilon)] - \sqrt{\frac{\frac{\|\hat{\theta}\|^2}{\sigma^2} + \ln\frac{N}{\alpha}}{2(N-1)}}, \tag{12}$$

where N is the size of $\hat{\mathbb{D}}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Proof: Given the input space $\mathcal{X} := \{x \in \mathbb{R}^d \mid \sum_{i=1}^d x_i^2 \le \tau\}$, a classifier $f_{\hat{\theta}} : \mathbb{R}^d \to \mathbb{R}^k$, an accuracy $A_{\mathcal{D}}(\cdot)$ over data distribution $\mathcal{D}$, and an empirical accuracy $A_{\hat{\mathbb{D}}}(\cdot)$ calculated from a test dataset $\hat{\mathbb{D}}$ of size N. Notice that the generalization error of $f_{\hat{\theta}}$ is

$$err_{\mathcal{D}}(\hat{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{1}[f_{\hat{\theta}}^{(y)}(x) \le \max_{i:i\neq y} f_{\hat{\theta}}^{(i)}(x)], \tag{17}$$

where $f_{\hat{\theta}}^{(i)}(x)$ is the i-th classification score of $f_{\hat{\theta}}(x)$. Similarly, the empirical error on $\hat{\mathbb{D}}$ is

$$err_{\hat{\mathbb{D}}}(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}[f_{\hat{\theta}}^{(y_j)}(x_j) \le \max_{i:i\neq y_j} f_{\hat{\theta}}^{(i)}(x_j)]. \tag{18}$$

When the classifier's weights are randomly perturbed by $\epsilon \sim \pi(0)$, according to Langford & Shawe-Taylor, and McAllester [34], for any $\alpha > 0$:

$$\begin{aligned}&\Pr\left(\mathbb{E}_\epsilon[err_{\mathcal{D}}(\hat{\theta}+\epsilon)] \le \mathbb{E}_\epsilon[err_{\hat{\mathbb{D}}}(\hat{\theta}+\epsilon)] + \sqrt{\frac{D_{KL}(\pi(\hat{\theta}+\epsilon)\|P) + \log\frac{N}{\alpha}}{2(N-1)}}\right)\\ &\ge 1-\alpha,\end{aligned} \tag{19}$$

where P is a prior distribution independent from training data, $\pi(\hat{\theta}+\upsilon)$ is the distribution of $\hat{\theta}+\upsilon$, and D_{KL} measures the Kullback–Leibler (KL) Divergence. Therefore, with probability

TABLE VIII: Trade-off between Train-Time/Certification Noise Level and Learnability (%) on CIFAR100

Train Noise	Certification Noise (σ)	$\eta \times 100$																						
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0	35.0	40.0	45.0	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0	95.0	100.0
0.10	0.10	**1.81**	**1.82**	**1.84**	**1.97**	**2.27**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	1.17	1.17	1.17	1.21	1.32	**1.40**	**1.49**	**1.60**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	1.13	1.13	1.13	1.17	1.22	1.25	1.33	1.40	**1.44**	**1.49**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	1.13	1.13	1.13	1.15	1.17	1.19	1.23	1.28	1.35	1.41	**1.50**	**1.52**	**1.60**	-	-	-	-	-	-	-	-	-	-
	0.50	1.12	1.13	1.13	1.14	1.16	1.18	1.20	1.23	1.25	1.33	1.34	1.35	1.37	1.37	**1.57**	1.57	-	-	-	-	-	-	-
	0.60	1.12	1.12	1.12	1.14	1.16	1.19	1.21	1.24	1.26	1.30	1.32	1.38	1.40	**1.46**	1.48	**1.61**	**1.70**	**1.74**	-	-	-	-	-
	0.70	1.12	1.12	1.12	1.13	1.15	1.16	1.18	1.19	1.21	1.24	1.25	1.28	1.30	1.32	1.34	1.35	1.42	1.43	**1.58**	**1.66**	**1.66**	-	-
	0.80	1.12	1.12	1.12	1.13	1.16	1.16	1.18	1.19	1.20	1.22	1.25	1.27	1.28	1.33	1.34	1.35	1.38	1.38	1.42	1.42	1.48	**1.48**	**1.60**
	0.90	1.11	1.11	1.11	1.12	1.12	1.13	1.14	1.15	1.17	1.17	1.19	1.20	1.21	1.22	1.22	1.25	1.25	1.30	1.33	1.39	1.41	1.44	1.56
	1.00	1.11	1.11	1.11	1.11	1.12	1.13	1.14	1.15	1.16	1.17	1.17	1.19	1.19	1.21	1.23	1.23	1.26	1.28	1.31	1.32	1.34	1.36	1.36
0.15	0.10	**1.85**	**1.86**	**1.88**	**2.01**	**2.33**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.15	1.61	1.18	1.19	1.27	1.39	**1.51**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	1.13	1.13	1.14	1.19	1.25	1.37	**1.45**	**1.66**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	1.12	1.13	1.13	1.15	1.21	1.25	1.33	1.40	**1.48**	**1.50**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	1.13	1.13	1.13	1.16	1.19	1.24	1.30	1.32	1.40	1.43	**1.46**	**1.49**	**1.52**	-	-	-	-	-	-	-	-	-	-
	0.50	1.13	1.13	1.13	1.14	1.16	1.18	1.19	1.21	1.24	1.26	1.28	1.30	1.37	**1.42**	**1.59**	**1.59**	-	-	-	-	-	-	-
	0.60	1.12	1.12	1.12	1.14	1.15	1.15	1.19	1.20	1.22	1.25	1.27	1.30	1.33	1.35	1.38	1.43	1.44	**1.52**	-	-	-	-	-
	0.70	1.12	1.12	1.12	1.13	1.14	1.15	1.17	1.18	1.21	1.25	1.26	1.28	1.35	1.40	1.41	1.43	**1.45**	1.49	**1.54**	**1.74**	**1.74**	-	-
	0.80	1.12	1.12	1.12	1.13	1.14	1.15	1.16	1.17	1.19	1.20	1.22	1.25	1.26	1.27	1.30	1.31	1.31	1.32	1.35	1.39	1.41	**1.41**	**1.48**
	0.90	1.11	1.11	1.11	1.11	1.12	1.13	1.15	1.16	1.18	1.19	1.21	1.22	1.24	1.25	1.27	1.31	1.32	1.34	1.34	1.37	1.39	1.39	1.46
	1.00	1.11	1.11	1.11	1.11	1.12	1.13	1.14	1.14	1.15	1.16	1.18	1.20	1.21	1.22	1.24	1.26	1.28	1.32	1.32	1.33	1.34	1.35	1.41
0.20	0.10	1.24	1.24	1.25	1.33	1.41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	**1.51**	**1.52**	**1.53**	**1.59**	**1.70**	**1.75**	**1.81**	**1.86**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	1.14	1.14	1.15	1.18	1.21	1.24	1.28	1.38	**1.45**	**1.51**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	1.14	1.14	1.15	1.18	1.19	1.23	1.25	1.30	1.37	1.43	**1.49**	**1.54**	**1.68**	-	-	-	-	-	-	-	-	-	-
	0.50	1.12	1.12	1.13	1.14	1.16	1.20	1.23	1.27	1.28	1.31	1.32	1.39	1.44	**1.50**	**1.70**	**1.70**	-	-	-	-	-	-	-
	0.60	1.12	1.12	1.12	1.14	1.16	1.17	1.19	1.21	1.24	1.25	1.30	1.30	1.31	1.36	1.39	1.47	**1.55**	**1.56**	-	-	-	-	-
	0.70	1.11	1.11	1.11	1.12	1.14	1.16	1.17	1.18	1.20	1.23	1.24	1.26	1.28	1.32	1.37	1.38	1.40	1.43	**1.46**	**1.49**	**1.49**	-	-
	0.80	1.12	1.12	1.12	1.12	1.14	1.16	1.18	1.20	1.21	1.23	1.24	1.27	1.30	1.31	1.33	1.33	1.34	1.36	1.38	1.39	1.39	1.39	**1.47**
	0.90	1.13	1.13	1.13	1.13	1.15	1.16	1.17	1.18	1.19	1.19	1.20	1.21	1.22	1.25	1.25	1.28	1.28	1.32	1.32	1.35	1.40	**1.43**	1.44
	1.00	1.12	1.12	1.12	1.13	1.14	1.16	1.16	1.17	1.17	1.18	1.19	1.20	1.22	1.23	1.25	1.26	1.28	1.30	1.31	1.35	1.37	1.40	1.40
0.25	0.10	1.12	1.12	1.13	1.19	1.31	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.20	1.27	1.27	1.28	1.31	1.39	1.47	**1.60**	**1.69**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.25	**1.32**	**1.32**	**1.33**	**1.37**	**1.42**	**1.47**	1.53	1.59	**1.68**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.30	1.18	1.18	1.19	1.23	1.28	1.32	1.35	1.52	1.56	**1.61**	-	-	-	-	-	-	-	-	-	-	-	-	-
	0.40	1.12	1.12	1.12	1.14	1.16	1.19	1.22	1.30	1.34	1.40	**1.44**	**1.47**	**1.58**	-	-	-	-	-	-	-	-	-	-
	0.50	1.12	1.12	1.12	1.15	1.17	1.19	1.21	1.23	1.24	1.25	1.28	1.35	1.41	**1.42**	**1.56**	**1.56**	-	-	-	-	-	-	-
	0.60	1.10	1.10	1.10	1.11	1.13	1.16	1.18	1.21	1.22	1.23	1.27	1.33	1.36	1.38	1.44	1.49	**1.51**	**1.52**	-	-	-	-	-
	0.70	1.10	1.11	1.11	1.11	1.13	1.15	1.17	1.19	1.20	1.21	1.23	1.25	1.27	1.28	1.31	1.31	1.38	1.40	**1.50**	**1.52**	**1.52**	-	-
	0.80	1.10	1.10	1.10	1.12	1.13	1.13	1.15	1.15	1.16	1.18	1.20	1.22	1.23	1.25	1.27	1.28	1.31	1.34	1.36	1.38	1.51	**1.51**	**1.56**
	0.90	1.10	1.10	1.10	1.11	1.13	1.14	1.15	1.16	1.17	1.18	1.19	1.20	1.21	1.21	1.23	1.24	1.25	1.26	1.28	1.30	1.30	1.34	1.34
	1.00	1.10	1.10	1.11	1.11	1.13	1.13	1.14	1.16	1.19	1.20	1.21	1.22	1.23	1.25	1.26	1.27	1.30	1.31	1.33	1.33	1.34	1.37	1.38

TABLE IX: Certification Offset

Data	Accuracy (Online)	C (Online)			Accuracy (Offline)	C (Offline)	
	PUE-B	PUE-10	PUE-1	EMN	PUE-10	EMN	OPS
CIFAR10	9.96	-0.68	-0.12	-4.72	10.32	+0.94	-0.28
CIFAR100	1.00	-0.02	-0.10	-0.68	0.97	+0.01	N/A
ImageNet	1.24	+0.15	+0.15	+0.15	N/A	N/A	N/A

TABLE X: Certified (q, η)-Learnability based on Different Surrogate Architectures (%, $\sigma = 0.25$)

Architecture	$\eta \times 100$								
	0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0
ResNet-18	**10.57**	**10.58**	**10.58**	**10.83**	**11.10**	**11.39**	**11.68**	**12.22**	**13.20**
ResNet-50	7.86	7.87	7.87	8.00	8.15	8.48	8.91	9.95	10.97

at least $1 - \alpha$,

$$\begin{aligned} &\mathbb{E}_\epsilon[err_\mathcal{D}(\hat{\theta} + \epsilon)] \\ \leq &\mathbb{E}_\epsilon[err_{\hat{\mathbb{D}}}(\hat{\theta} + \epsilon)] + \sqrt{\frac{D_{KL}(\pi(\hat{\theta} + \epsilon)||P) + \log \frac{N}{\alpha}}{2(N-1)}}. \end{aligned} \tag{20}$$

Let P be $\mathcal{N}(0, \sigma^2 I)$, and by knowing $\pi(\hat{\theta} + \epsilon) = \mathcal{N}(\hat{\theta}, \sigma^2 I)$, the KL Divergence can be computed as

$$\begin{aligned} D_{KL}(\pi(\hat{\theta} + \epsilon)||P) &= \frac{||\hat{\theta}||^2}{2\sigma^2} + \frac{|\hat{\theta}|}{2}\left[\frac{\sigma^2}{\sigma^2} + \log(\frac{\sigma^2}{\sigma^2}) - 1\right] \\ &= \frac{||\hat{\theta}||^2}{2\sigma^2}, \end{aligned} \tag{21}$$

where $|\hat{\theta}|$ is the size of $\hat{\theta}$. Plugging Equation 21 and

$$A_\mathcal{D}(\hat{\theta} + \epsilon) = 1 - err_\mathcal{D}(\hat{\theta} + \epsilon)$$
$$A_{\hat{\mathbb{D}}}(\hat{\theta} + \epsilon) = 1 - err_{\hat{\mathbb{D}}}(\hat{\theta} + \epsilon)$$

into Inequality 20 concludes the proof. ∎

C. More Experimental Settings and Results

Additional details of hyper-parameters. For all experiments, we set the value $\xi = 8/255$ for CIFAR10 and CIFAR100. On ImageNet, we set $\xi = 16/255$. The training batch size is 128 for all classifiers except DenseNet121, where it is set to 96. The step size of gradient descent during optimizing the noise δ is set to $\xi/10$. We set $U_{train} = 100$ for all training with random weight perturbations. In all experiments of crafting PUE/PUE-B/EMN, the value of M is set to 10 for CIFAR10, 20 for CIFAR100, and 100 for ImageNet. The stop criteria for the PUE/PUE-B/EMN optimization is that the validation error drops below 10%. We use 20% of the training set to optimize the PAP noise and validate the error of the online surrogate. The online surrogate is trained on the full training set with a poisoning rate of 1.0. In offline surrogate training, we use the fully poisoned training datasets for training and validation. The offline surrogate training is stopped when the training error is less than 10% and the surrogate has gone through at least one epoch of training with random weight perturbations.

Comparison of (q, η)-Learnability. Given a surrogate with parameters $\hat{\theta}$ and a corresponding $\hat{\Theta}$, we are interested in the accuracy gain $l_{(q,\eta)}(\hat{\Theta}; \mathbb{D}_s \oplus \delta) - A_\mathbb{D}(\hat{\theta})$ resulting from perturbing $\hat{\theta}$. When comparing two surrogates trained by different methods, the one inducing a higher accuracy gain through certification can locate a certified parameter set on which a tighter certification can be made. Regarding surrogates trained on different PAP noises using the same training methods, their accuracy gains reflect the robustness of the noises towards recovery attacks or uncertainties in parameters. Therefore, we adjust the learnability scores in the comparisons to better demonstrate the differences in accuracy gains. In the online setting, let the accuracy of the PUE-B surrogate be $A_\mathbb{D}(\theta_{PUE-B})$, we offset the certified learnability of $X \in$ {PUE-10, PUE-1, EMN} by adding a constant $C = A_\mathbb{D}(\theta_{PUE-B}) - A_\mathbb{D}(\theta_X)$. In the

TABLE XI: Empirical Top-1 Clean Testing Accuracy (%) of Classifiers Trained on UEs

Data	Model	PUE	EMN	OPS	Clean
CIFAR10	ResNet-18	**10.62**	11.89	12.95	94.90
	ResNet-50	**10.00**	12.37	10.96	94.17
	DenseNet-121	**10.31**	11.86	13.80	95.16
CIFAR100	ResNet-18	2.62	2.49	**1.57**	70.78
	ResNet-50	2.03	**1.61**	1.78	70.89
	DenseNet-121	3.01	2.56	**2.32**	95.16
ImageNet	ResNet-18	**4.56**	4.84	22.32	79.86
	ResNet-50	**3.34**	4.00	26.62	82.36
	DenseNet-121	**4.60**	6.23	22.60	84.04

offline setting, we use PUE-10 as the baseline and accordingly offset certification results from EMN and OPS. The offsets are listed in Table IX.

Ablation study of certification. The ablation study on CIFAR10 and CIFAR100 is presented in Table VII and Table VIII, respectively. We can observe similar trends in the results on the two datasets. Generally, when the train-time noise matches the parametric smoothing noise, higher (q, η)-Learnability can be certified. A train-time noise level of 0.25 can obtain higher (q, η)-Learnability certified at farther η.

Cross-architecture certification. We train offline surrogates based on ResNet-18 and ResNet-50 architectures on the PUE version of CIFAR10, respectively. The CIFAR10 PUEs are crafted based on ResNet-18. Therefore, the ResNet-50 surrogate examines a cross-architecture certification performance. We compare the certification results from these two architectures in Table X. We set $\sigma = 0.25$ in the certification. Similar to the previous comparisons, we use ResNet-18 as the baseline and apply offsets (-2.26) to the (q, η)-Learnability scores certified by ResNet-50.

The results show that ResNet-18 certifies tighter (q, η)-Learnability scores. The results suggest that the surrogate used in certification should match the surrogate used for generating UEs. Note that ResNet-50 has more parameters, therefore the certifications are carried out on parameter spaces with different dimensionality. On top of that, we also notice that ResNet-50 is less capable of finding better classifiers through parametric randomization. Better surrogate training methods could be proposed in the future to efficiently augment surrogates with more parameters.

TABLE XII: Impact of Training Strategies

Strategy	Data			
	PUE	EMN	OPS	Clean
MixUp	24.66	17.79	32.99	95.21
CutOut	12.05	14.42	57.69	95.18
Fast Autoaugment	35.33	22.71	65.61	95.14
ℓ_∞ AT	85.03	85.06	10.45	84.97

TABLE XIII: Training Cost (GPU hour) for PUEs and Online Surrogates

Dataset	Noise Level				
	0.05	0.10	0.15	0.20	0.25
CIFAR10	0.24	1.20	1.20	19.68	39.60
CIFAR100	0.24	2.16	2.16	12.96	38.88
ImageNet	0.40	3.84	5.04	23.28	45.90

TABLE XIV: Training Cost (GPU hour) for Offline Surrogates

Dataset	PAP Noise		
	PUE	EMN	OPS
CIFAR10	2.16	2.16	3.60
CIFAR100	2.16	2.16	N/A

Computational overhead. We also tested the effective noise level for training the surrogate and PUEs. We found that the surrogate and PUEs could converge rapidly on a Gaussian noise whose $\sigma' \leq 0.25$. However, when the noise further increases to $\sigma' \geq 0.3$, the training fails to converge in a limited time. We record the GPU hours of training PUEs and surrogates under different noise levels, on Nvidia Tesla P100 GPUs (Table XIII and Table XIV). The training cost increases with the scale of the train-time noise. Each training run of PUEs and an online surrogate takes around 40 GPU hours when selecting 0.25 as the scale of the train-time noise. Training an offline surrogate with a noise scale of 0.25 takes around 2.16 GPU hours. The training cost is within an acceptable range.

D. Empirical Evaluations

In this section, we investigate the empirical utility of UEs. We conduct empirical evaluations of the PUEs by comparing the clean test accuracy of classifiers trained on PUE, EMN [9], and OPS [35]. We also employ train-time strategies in our experiments to show the robustness of PUE against such strategies. However, as discussed in this paper, we want to emphasize that empirical results could be less instructive due to the uncertainty brought by the randomness in training.

Vanilla training. In this part, we measure the empirical Top-1 accuracy of classifiers trained on PUE-10, EMN, and OPS, respectively. There is no particular training strategy incorporated at this stage and we stick to the training details mentioned at the beginning of the Experiments section. We use the default settings from the corresponding papers of EMN and OPS in the comparison, except for ImageNet OPS. We use Random Resized Crop instead of Center Crop to process ImageNet data to 224×224 for PUE, EMN, and OPS. The results are presented in Table XI. PUE outperforms all competitors on CIFAR10 and ImageNet, and it is on par with EMN and OPS on CIFAR100. We observe that OPS performs poorly on ImageNet. The possible reason is that Random Resized Crop diversifies the pixel distribution in each class, rendering the task of finding the most impactful pixel arduous for OPS.

Training with strategies. Training techniques such as data augmentation and adversarial training are prevalently placed into model training in real-world cases. Though the certified learnability considers only the parameters of the trained model and ignores the training techniques, it is also of interest to test how PUEs react to these strategies when the trained classifier is out of the certifiable parameter set. In this part, we apply different techniques in the training and compare the robustness of PUE, EMN and OPS towards these techniques. Specifically, we tested PUEs along with its competitors on adversarial training (AT) [59], Mixup [60], Cutout [61], and Fast Autoaugment [62]. We use an ℓ_∞ norm of $8/255$ for PGD in the adversarial training. The results are in Table XII. PUE outperforms both EMN and OPS when training using CutOut, and better survives from MixUp and Fast Autoaugment than OPS. In contrast, OPS is barely affected by adversarial training while PUE and EMN can be invalidated.