Prompt Certified Machine Unlearning with Randomized Gradient Smoothing and Quantization

Anonymous Author(s)
Affiliation
Address
email

Abstract

The right to be forgotten calls for efficient machine unlearning techniques that make trained machine learning models forget a cohort of data. The combination of training and unlearning operations in traditional machine unlearning methods often leads to the expensive computational cost on large-scale data. This paper presents a Prompt Certified Machine Unlearning algorithm, PCMU, which executes one-time operation of simultaneous training and unlearning in advance for a series of machine unlearning requests, without the knowledge of the removed/forgotten data. First, we establish a connection between randomized smoothing for certified robustness on classification and randomized smoothing for certified machine unlearning on gradient quantization. Second, we propose a prompt certified machine unlearning model based on randomized data smoothing and gradient quantization. We theoretically derive the certified radius R regarding the data change before and after data removals and the certified budget of data removals about R. Last but not least, we present another practical framework of randomized gradient smoothing and quantization, due to the dilemma of producing high confidence certificates in the first framework. We theoretically demonstrate the certified radius R' regarding the gradient change, the correlation between two types of certified radii, and the certified budget of data removals about R'.

1 Introduction

In order to to respect the privacy, machine unlearning techniques aim to enable data owners to proactively remove their data and eliminate its influence from already trained machine learning model upon requests [11, 44, 106, 43, 50, 40, 51, 126, 95]. A straightforward solution is to retrain new models from scratch on the remaining/remembered data, without the knowledge of the removed/forgotten data, as if the retrained model has never seen the forgotten data. However, the naive method is impractical since it often encounters expensive cost over complex models (e.g., DNN) on large data.

This has motivated the recent study of resolving the inefficiency issue of the naive machine unlearning. Existing techniques can be broadly classified into two categories: (1) Exact unlearning algorithms aim to learn an unlearning model with the same performance as the above naive ones retraining from scratch by completely excluding the forgotten data from the training data [102, 13, 61, 18, 77, 83, 9, 103, 43, 11, 4, 8, 117, 16, 15] and (2) Approximate unlearning methods try to bring the parameters of the trained model closer to the naive ones through the relaxation of exact unlearning requirements [4, 49, 46, 115, 78, 85, 127, 57, 92, 63, 43, 50, 92, 105, 44, 45, 50, 94, 47, 38, 78, 79, 81, 118, 125, 46, 16, 134, 80]. certified removal is a certified-removal mechanism that applies a Newton step on the model parameters that largely remove the influence of the deleted data points [50]. GKT is a zero-shot machine unlearning algorithm that imposes the constraint that zero training data is available to the unlearning algorithms [21]. Two recent studies propose online machine unlearning methods for linear regression models [77] and linear support vector machine models [18], for further

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

improving the efficiency of machine unlearning. The former adapts users' requests to delete their data before a specific time bar. The latter conducts only the task of variable support vector machine.

Despite achieving remarkable success, most of the above machine unlearning methods consist of two sequential operations: (1) Training: train a model on the complete training data and (2) Unlearning: generate an unlearning model from the former. The combination of two operations is computationally expensive when training complex models over large datasets. In addition, they often sequentially redo the unlearning operations one by one, when addressing a series of machine unlearning requests.

Randomized smoothing has achieved the state-of-the-art certified robustness guarantees against worst-case attacks by smoothing with isotropic Gaussian distribution [69, 24, 70, 75, 99, 71, 135, 59, 6, 39, 132, 90, 86, 35, 14, 133, 112, 1, 17, 54, 84]. Specially, it takes a base classifier $f(x)$ as an input, and outputs a smooth classifier $g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \underset{\varepsilon \sim \mathcal{D}}{\mathbb{P}}(f(x+\varepsilon) = c)$ by averaging its prediction over isotropic Gaussian noise $\mathcal{D} = \mathcal{N}\left(0, \sigma^2 I\right)$ of the input data x within its neighborhood ε. It provides a tight certified robustness guarantee: $g(x)$ can always return the most probable class c_A, i.e., $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, as long as the perturbation δ is within the certified radius R.

This motivates us to establish a connection between randomized smoothing for certified robustness on classification and randomized smoothing for certified machine unlearning on gradient quantization.. We analogize the data removals on the entire training data (i.e., the perturbations on the entire data) in the machine unlearning to the adversarial attacks (i.e., the perturbations on the data samples) in the certified robustness and liken the output quantized gradients in the former to the output discrete class labels in the latter. Since the output class labels in the latter through randomized smoothing are able to keep unchanged and correct against adversarial attacks within the certified radius, it is highly possible that the output quantized gradients in the former through randomized smoothing can keep unchanged against data removals within the certified budget, which implies that the learnt model shares the same gradients (and parameters) with the naive one retrained on only the remembered data.

Instead of performing two sequential operations of training and unlearning, this work directly trains an unlearning model in advance, without the knowledge of the forgotten data, based on randomized data smoothing and gradient quantization. We propose to execute the randomized smoothing on the average $\overline{x}$ of data samples. When there are data removal requests in the training data, $\overline{x}$ will be updated with a new average $\overline{x}'$, which can be treated as a perturbed version of $\overline{x}$. As the class labels lie in a (countable) small discrete space but the gradients lie in a large continuous space, we propose a gradient quantization technique to produce discrete gradients in a three-class space $\{-1, 0, 1\}$. The randomized smoothing method needs to sample a large number of points surrounding the input data for producing high confidence certificates, e.g., 10^5 samples for 99.9% confidence [24]. In our context, the outputs of randomized smoothing are high-dimensional gradients. We propose to utilize the Taylor expansion to approximate the output gradients of the sampled points for avoiding expensive gradient computation. We theoretically derive the error introduced by the Taylor expansion, the certified radius R regarding the perturbation surrounding $\bar{x}$ and the certified budget B of data removals (i.e., the maximally allowed amount of escaped data samples).

Notice that the response to the data removals is to erase the data samples and their associated labels together. It is necessary to smooth on both of them. However, for a given dataset, the relationship between the samples and labels is a multi-valued non-continuous mapping, which is a non-integrable function. This results in the dilemma of producing high confidence certificates, even with sampling and estimation techniques. Therefore, we propose a feasible solution based on randomized gradient smoothing and gradient quantization. We theoretically demonstrate the certified radius R' regarding the gradient perturbations. Most importantly, we recognize the correlations between two types of radii R and R' and between B and R, which are used to derive the certified budget B about R'. We further integrate the model training, randomized gradient smoothing, and gradient quantization into a unified framework for directly training a machine unlearning model with the data removal certificates as a guidance, for guaranteeing that the model parameters and gradients keep unchanged against the data removals within the certified budget.

In comparison with existing machine unlearning techniques, our randomized gradient smoothing and gradient quantization method exhibits three compelling advantages: (1) It simultaneously executes the training and unlearning operations, which is able to dramatically improve the unlearning efficiency for complex models on large-scale data; (2) The one-time operation of simultaneous training and unlearning can provide the timely response to a series of machine unlearning requests, as long as

the actual data removals are below the certified budget of data removals; and (3) It is agnostic to the removed/forgotten data before performing the unlearning operation.

Empirical evaluation on real datasets demonstrates the superior performance of our PCMU model against several state-of-the-art machine unlearning methods on image classification. More experiments, implementation details, and hyperparameter setting are presented in Appendices A.4-A.6.

2 Background

2.1 Randomized Smoothing for Certified Robustness

Randomized smoothing aims to build a smoothed classifier g from a base classifier f that maps inputs $x \in \mathbb{R}^d$ to classes $c \in C$.

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \underset{\varepsilon \sim \mathcal{D}}{\mathbb{P}}(f(x+\varepsilon) = c) \tag{1}$$

where $\mathcal{D} = \mathcal{N}\left(0, \sigma^2 I\right)$ is a Gaussian probability distribution in $\mathbb{R}^d$ for randomized smoothing. g returns whichever class f is most likely to return when x is perturbed by noise ε.

Let $p_c(x)$ be the output probability of f over class c, i.e., $p_c(x) = \underset{\varepsilon \sim \mathcal{D}}{\mathbb{P}}(f(x+\varepsilon) = c)$. Without loss of generality, we assume that $p_A(x)$ and $p_B(x)$ are the probabilities on the most probable class c_A and the runner-up class c_B respectively. If $\mathbb{P}\left(f(x+\varepsilon) = c_A\right) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c)$, where $\underline{p_A(x)}$ is a lower bound of $p_A(x)$ and $\overline{p_B(x)}$ is an upper bound of $p_B(x)$, then $g(x+\delta) = c_A$ for $\forall \delta \in \mathbb{R}^d, \|\delta\|_2 \leq R$. In this case, the smoothed classifier g can always output the correct prediction as long as the perturbation δ is within a certified l_2-norm radius of R.

Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ satisfy [24]:*

$$\mathbb{P}\left(f(x+\varepsilon) = c_A\right) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \tag{2}$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right) \tag{3}$$

where Φ^{-1} is the inverse of the standard Gaussian CDF.

2.2 Machine Unlearning

Machine unlearning aims to enable the trained models to forget what has been learned from the data to be removed. Specifically, given a training dataset of N samples $D = \{x_i, y_i\}_{i=1}^N$. Each sample x_i is associated with a label $y_i \in \mathcal{Y} = \{1, 2, ..., Y\}$, where Y is the number of classes. A classification model $M(D)$ is trained on the complete training dataset D.

The users can submit a data removal request at any time. Thus, the complete training data D is partitioned into two subsets: $D_f \subseteq D$ denoting the data which we wish the classification model to forget and $D_r \subseteq D$ specifying the data which we want the model to remember ($D = D_f \cup D_r$). The goal of machine unlearning is to unlearn the forgotten data D_f, i.e., eliminate the influence of D_f from $M(D)$. A straightforward solution is to use the remembered data D_r as the training data to retrain a new classification model $M_r(D_r)$ from scratch. However, this naive method is often time-consuming over large-scale datasets. An efficient algorithm is to directly generate a sanitized model $M_u(D, D_f, M)$ from the deployed model $M(D)$ that approximates $M_r(D_r)$, i.e.,

$$M_u(D, D_f, M) \approx M_r(D_r) \tag{4}$$

3 Randomized Data Smoothing and Gradient Quantization

The idea of this work is to establish a connection between randomized smoothing for certified robustness on classification and randomized smoothing for certified machine unlearning on gradient quantization. By leveraging the theory of randomized smoothing and gradient quantization, we theoretically derive the certified radius R regarding the perturbation surrounding the data average $\bar{x}$ and the certified budget B of data removals.

The gradient $G(x, y) \in \mathbb{R}^T$ of a machine learning model is given as follows.

$$G(x, y) = \frac{\partial \mathcal{L}(x, y; w)}{\partial w} \tag{5}$$

where $\mathcal{L}$ is the loss function, e.g., cross-entropy for image classification. w is the model parameter.

We propose to quantize each dimension t ($t = 1, \cdots, T$) of the continuous gradient $G(x, y) \in \mathbb{R}^T$ over a discrete three-class space $\{-1, 0, 1\}$, for mimicking the classification in the randomized smoothing for certified robustness.

$$Q(t) = Softmax([-|t - \sigma^2|, -|t|, -|t + \sigma^2|]) \tag{6}$$

where $Q(t)$ maps a gradient dimension t to a three-dimensional vector $[-|t - \sigma^2|, -|t|, -|t + \sigma^2|]$, where each component denotes the similarity score between t and $-\sigma^2$, 0, or σ^2. σ is the standard deviation of the Gaussian probability distribution in the randomized smoothing and also serves as a quantization threshold in our method. The details of the selection of quantization threshold are presented in Appendix A.2. Therefore, all T gradient dimensions are partitioned into three intervals: $(-\infty, -\sigma^2/2]$ that comes near to $-\sigma^2$, $[-\sigma^2/2, \sigma^2/2]$ that is closer to 0, and $[\sigma^2/2, \infty)$ that approaches σ^2. Each component in $Q(t)$ with the Softmax function also represents the probability of the gradient dimension t belonging to classes -1, 0, or 1. The most probable class $c_A \in \{-1, 0, 1\}$ in $Q(t)$ is assigned to dimension t as a final quantized gradient dimension.

We represent the composition $F(x, y)$ of gradient computation and quantization as follows.

$$F(x, y) = Q(G(x, y)) \tag{7}$$

We use $F^t(x, y)$ to denote the output three-dimensional quantization vector of the tth ($t = 1, \cdots, T$) dimension of the gradient $G(x, y)$ and use $F^t_c(x, y)$ to represent the cth ($c \in \{-1, 0, 1\}$) component of $F^t(x, y)$.

As the data removal is treated as the noise on the entire training data, we use the average $\bar{x}$ of all data samples to represent the entire training data.

$$\bar{x} = \frac{1}{N} \sum_{x_i \in D} x_i, \;\; \bar{y} = \frac{1}{N} \sum_{y_i \in D} y_i \tag{8}$$

where $\bar{y}$ is the average of the class labels of all data samples.

The randomized data smoothing for certified unlearning on gradient quantization is defined below.

$$S^t(\bar{x}, \bar{y}) = \underset{c \in \{-1,0,1\}}{\operatorname{argmax}} \underset{\varepsilon_x, \varepsilon_y \sim \mathcal{D}}{\mathbb{P}} (F^t(\bar{x} + \varepsilon_x, \bar{y} + \varepsilon_y) = c) \tag{9}$$

where $\mathcal{D} = \mathcal{N}\left(0, \sigma^2 I\right)$ is a Gaussian distribution. S^t is a smoothed version of a base gradient quantizatier F^t that maps the tth gradient dimension about inputs $(\bar{x}, \bar{y})$ to gradient classes $c \in \{-1, 0, 1\}$. S^t returns whichever class F^t is most likely to return when $(\bar{x}, \bar{y})$ is perturbed by $(\varepsilon_x, \varepsilon_y)$.

The randomized smoothing method needs to sample a large number of points surrounding the input data for producing high confidence certificates, e.g., 10^5 samples for 99.9% confidence [24]. In our context, computing the gradients for massive samples is extremely inefficient. We utilize the Taylor expansion to approximate the output gradients of sampled points, based on the outputs from the original sample $(\bar{x}, \bar{y})$. For each quantization component $F^t_c(\hat{\bar{x}}, \hat{\bar{y}})$ for a sample in a Gaussian distribution surrounding $(\bar{x}, \bar{y})$, where $\hat{\bar{x}} = \bar{x} + \varepsilon_x, \hat{\bar{y}} = \bar{y} + \varepsilon_y, \varepsilon_x, \varepsilon_y \sim \mathcal{D}$, we take the Taylor expansion of $F^t_c(\hat{\bar{x}}, \hat{\bar{y}})$ at $(\bar{x}, \bar{y})$ as follows.

$$\begin{aligned} F^t_c(\hat{\bar{x}}, \hat{\bar{y}}) =& F^t_c(\bar{x}, \bar{y}) + \frac{\partial F^t_c}{\partial x}(\bar{x}, \bar{y})(\hat{\bar{x}} - \bar{x}) + \frac{\partial F^t_c}{\partial y}(\bar{x}, \bar{y})(\hat{\bar{y}} - \bar{y}) + \\ &+ \frac{1}{2}\{\frac{\partial^2 F^t_c(\bar{x}, \bar{y})}{\partial x^2}(\hat{\bar{x}} - \bar{x})^2 + 2\frac{\partial^2 F^t_c(\bar{x}, \bar{y})}{\partial x \partial y}(\hat{\bar{x}} - \bar{x})(\hat{\bar{y}} - \bar{y}) + \frac{\partial^2 F^t_c(\bar{x}, \bar{y})}{\partial y^2}(\hat{\bar{y}} - \bar{y})^2\} \\ &+ \cdots + O_j(\hat{\bar{x}}, \hat{\bar{y}}) \end{aligned} \tag{10}$$

where $O_j(\hat{\bar{x}}, \hat{\bar{y}}) = \frac{1}{(j+1)!}\{\sum \cdots \sum \frac{\partial^k F^t_c(\xi)}{\partial x^k \partial y^{j+1-k}}(\hat{\bar{x}} - \bar{x})^k (\hat{\bar{y}} - \bar{y})^{j+1-k}\}$, $\xi \in ((\hat{\bar{x}}, \bar{x}), (\hat{\bar{y}}, \bar{y}))$, $j = 3, \cdots, \infty$, and $k = 1, 2, 3$.

The following theorem derives the error introduced by the Taylor expansion.

Theorem 2. *The error introduced by the Taylor expansion of $F^t_c(\hat{\bar{x}}, \hat{\bar{y}})$ at $(\bar{x}, \bar{y})$ is*

$$\epsilon \leq \sum_{j=0}^{\infty} ||\frac{L_{j+1}}{(j+1)!}|| \cdot ||\sigma M||^{j+1} \tag{11}$$

where $L_j = \max_{k=1,\cdots,j} \frac{\partial^j h_i}{\partial x^k \partial y^{j-k}}$ and M is the number of sampled points.

Please refer to Appendix A.3 for detailed proof of Theorem 2.

Notice that

$$\mathbf{P}(F^t(\hat{\bar{x}},\hat{\bar{y}})=c)=\int_{\hat{\bar{x}}}\int_{\hat{\bar{y}}}\mathbf{P}(F_c^t(\hat{\bar{x}},\hat{\bar{y}}))d\hat{\bar{x}}d\hat{\bar{y}}\geq\max_{c\in\{-1,0,1\}}F_c^t(\hat{\bar{x}},\hat{\bar{y}}) \tag{12}$$

Thus, we derive the probabilities $\underline{p_A}$ and $\overline{p_B}$ on the most probable class c_A and the runner-up class c_B in the randomized smoothing for certified machine unlearning ($c_A, c_B \in \{-1,0,1\}$), based on the error introduced by the Taylor expansion.

$$\underline{p_A}=\max_{c\in\{-1,0,1\}}\mathbf{P}(F_c^t(\hat{\bar{x}},\hat{\bar{y}})-\epsilon)=\int_{\hat{\bar{x}}}\int_{\hat{\bar{y}}}\mathbf{P}(F_c^t(\hat{\bar{x}},\hat{\bar{y}})-\epsilon)d\hat{\bar{x}}d\hat{\bar{y}},\ c=c_A \tag{13}$$

$$\overline{p_B}=\max_{c\neq c_A}\mathbf{P}(F_c^t(\hat{\bar{x}},\hat{\bar{y}})+\epsilon)=\int_{\hat{\bar{x}}}\int_{\hat{\bar{y}}}\mathbf{P}(F_c^t(\hat{\bar{x}},\hat{\bar{y}})+\epsilon)d\hat{\bar{x}}d\hat{\bar{y}},\ c\neq c_A \tag{14}$$

Notice that the correlation between the data $\bar{x}$ and its classes $\bar{y}$ is fixed for a given dataset. We denote this correlation as $\bar{y}=H(\bar{x})$ where $H:\mathbb{R}^d\mapsto C$. Thus, the randomized smoothing for certified machine unlearning in Eq.(9) is rewritten as an equivalent one as follows.

$$\tilde{S}^t(\bar{x})=\operatorname*{argmax}_{c\in\{-1,0,1\}}\mathbb{P}_{\varepsilon\sim\mathcal{D}}(\tilde{F}^t(\bar{x}+\varepsilon)=c) \tag{15}$$

where $\tilde{F}^t(\bar{x})=F^t(\bar{x},H(\bar{x}))=F^t(\bar{x},\bar{y})$.

Based on the computed $\underline{p_A}$, $\overline{p_B}$, and $F'(\bar{x})$, we can obtain the certified radius R regarding the perturbation surrounding $\bar{x}$ and the certified budget B of data removal.

Theorem 3. *Let $\varepsilon\sim\mathcal{D}=\mathcal{N}\left(0,\sigma^2 I\right)$ and $\tilde{S}^t(\bar{x})=\operatorname*{argmax}_{c\in\{-1,0,1\}}\mathbb{P}_{\varepsilon\sim\mathcal{D}}(\tilde{F}^t(\bar{x}+\varepsilon)=c)$. Suppose that for a specific $\bar{x}\in\mathbb{R}^d$, there exist $c_A\in\{-1,0,1\}$ and $\underline{p_A},\overline{p_B}\in[0,1]$ such that:*

$$\mathbb{P}\left(\tilde{F}^t(\bar{x}+\varepsilon)=c_A\right)\geq\underline{p_A}\geq\overline{p_B}\geq\max_{c\neq c_A}\mathbb{P}(\tilde{F}^t(\bar{x}+\varepsilon)=c) \tag{16}$$

Then $\tilde{S}^t(\bar{x}+\delta)=c_A$ for all $\|\delta\|_2<R$, where

$$R=\frac{\sigma}{2}\left(\Phi^{-1}\left(\underline{p_A}\right)-\Phi^{-1}\left(\overline{p_B}\right)\right) \tag{17}$$

where Φ^{-1} is the inverse of the standard Gaussian CDF.

Theorem 4. *Let R be the certified radius of $\bar{x}\in\mathbb{R}^d$ based on $\tilde{S}^t(\bar{x})=\operatorname*{argmax}_{c\in\{-1,0,1\}}\mathbb{P}_{\varepsilon\sim\mathcal{D}}(\tilde{F}^t(\bar{x}+\varepsilon)=c)$, then the certified budget of data removal is*

$$B\leq N-\frac{9d\sigma^2}{R^2} \tag{18}$$

Please refer to Appendix A.3 for detailed proof of Theorems 3 and 4.

The above method is effective for certified machine unlearning, but it is computationally expensive to calculate high-dimensional double integrals in $\underline{p_A}$ and $\overline{p_B}$. We can reduce the double integrals to the single integrals through $\bar{y}=H(\bar{x})$. However, H is essentially a multi-valued non-continuous mapping, which is not an integrable function and makes the above method impractical.

$$\underline{p_A}=\int_{\hat{\bar{x}}}\int_{\hat{\bar{y}}}\mathbf{P}(F_c^t(\hat{\bar{x}},\hat{\bar{y}})-\epsilon)d\hat{\bar{x}}d\hat{\bar{y}}=\int_{\hat{\bar{x}}}\mathbf{P}(F_c^t(\hat{\bar{x}},H(\hat{\bar{x}}))-\epsilon)d\hat{\bar{x}},\ c=c_A \tag{19}$$

4 Randomized Gradient Smoothing and Quantization

In order to avoid the dilemma of computing practical $\underline{p_A}$, $\overline{p_B}$ and R, we propose a feasible solution based on randomized gradient smoothing and gradient quantization. We theoretically demonstrate the certified radius R' regarding the gradient perturbations. We recognize the correlations between R and R' and between B and R, which are used to derive the certified budget B about R'.

We first calculate the gradient $G(x_i,y_i)$ in terms of each sample (x_i,y_i) and the gradient average $\bar{G}$.

$$\bar{G}=\frac{1}{N}\sum_{(x_i,y_i)\in D}G(x_i,y_i) \tag{20}$$

The randomized gradient smoothing for certified unlearning on gradient quantization is defined below.

$$S^{t\prime}(\bar{G})=\operatorname*{argmax}_{c\in\{-1,0,1\}}\mathbb{P}_{\varepsilon\sim\mathcal{D}}(Q^t(\bar{G}+\varepsilon)=c) \tag{21}$$

where $S^{t\prime}$ is a smoothed version of a base gradient quantizatier Q^t that maps each dimension t $(t = 1, \cdots, T)$ of the gradient $\bar{G}$ to gradient classes $c \in \{-1, 0, 1\}$. $S^{t\prime}$ returns whichever gradient class Q^t is most likely to return when $\bar{G}$ is perturbed by noise ε.

We compute the probabilities over three intervals of $(-\infty, -\sigma^2/2]$, $[-\sigma^2/2, \sigma^2/2]$, and $[\sigma^2/2, \infty)$.

$$P = \Big\{ \int_{\frac{\sigma_2}{2} - \hat{\bar{G}}^t}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}}\, dz, \int_{-\frac{\sigma_2}{2} - \hat{\bar{G}}^t}^{\frac{\sigma_2}{2} - \hat{\bar{G}}^t} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}}\, dz, \int_{-\infty}^{-\frac{\sigma_2}{2} - \hat{\bar{G}}^t} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}}\, dz \Big\} \tag{22}$$

Now, we can directly generate the corresponding probabilities $\underline{p'_A}$ and $\overline{p'_B}$.

$$\underline{p'_A} = \max P,\ \overline{p'_B} = \max\left\{P - \{\underline{p'_A}\}\right\} \tag{23}$$

By following the similar strategy in Theorem 3, we can derive the corresponding certified radius R'. Namely, $S^{t\prime}(\bar{G} + \delta) = c_A$ for all $\|\delta\|_2 < R'$, where

$$R' = \frac{\sigma}{2}\left(\Phi^{-1}\left(\underline{p'_A}\right) - \Phi^{-1}\left(\overline{p'_B}\right)\right) \tag{24}$$

However, it is difficult to obtain the corresponding certified budget of data removal from R', since R' is related to the perturbations over the gradient $\bar{G}$, instead of the data $\bar{x}$. The following theorem demonstrates the correlation between two types of radii R and R'.

Theorem 5. *Let R and R' be the certified radii of the above two algorithms respectively and L be the Lipschitz constant of gradient $G(x, y)$, then*

$$R \geq \frac{\sqrt{d}}{L} R' \tag{25}$$

By combining Theorems 4 and 5 together, we derive the certified budget B' of data removal from R'.

$$B' \leq N - \frac{36L^2}{(\Phi^{-1}(\underline{p_A}') - \Phi^{-1}(\overline{p_B}'))^2} \tag{26}$$

In addition, we conduct the convergence analysis of our prompt certified machine unlearning algorithm based on randomized gradient smoothing and quantization.

Theorem 6. *Let $S^{t\prime}(\bar{G})$ be the randomized gradient smoothing for certified machine unlearning on gradient quantization, L, L_1, and L_2 be the Lipschitz constants of G, Q^t, and $S^{t\prime}$ respectively, i.e.,*

$$||\nabla S^{t\prime}(a) - \nabla S^{t\prime}(b)||_2 \leq L_2 L_1 L ||a - b||_2 \ for\ any\ a, b \tag{27}$$

If we run gradient descent for k iterations with a fixed step size $s \leq \frac{1}{L_2 L_1 L}$, it will yield a solution $S^{t\prime(k)}$ which satisfies

$$S^{t\prime}(q^{(k)}) - S^{t\prime}(q^*) \leq \frac{||q^{(0)} - q^*||_2^2}{2sk} \tag{28}$$

where $S^{t\prime}(q^{(0)})$ is the initial solution and $S^{t\prime}(q^*)$ is the local optimal solution.

This means that gradient descent is guaranteed to converge and that it converges with rate $\mathcal{O}(1/k)$. Please refer to Appendix A.3 for detailed proof of Theorems 5 and 6.

Finally, we integrate the model training for a specific learning task (e.g., image classification) randomized gradient smoothing, and gradient quantization into a unified framework for directly training a machine unlearning model with the data removal certificates as a guidance, for guaranteeing that the model parameters and gradients keep unchanged against the data removals within the certified budget. The corresponding parameter update is given below.

$$w = w - \eta[S^{1\prime}(\bar{G}), \cdots, S^{T\prime}(\bar{G})] \tag{29}$$

where w is the model parameter and η is a learning rate.

5 Experimental Evaluation

In this section, we have evaluated the effectiveness of our PCMU model and other comparison methods for machine unlearning over three popular image classification datasets: Fashion-MNIST [129, 51, 38], CIFAR-10 [65, 44, 45, 114, 51, 38], and SVHN [93, 50, 8]. We train the classifiers on the training set and test them on the test set for three datasets. We train a convolutional neural network (CNN) on

Table 1: Performance with 10% data removal and CNN on Fashion-MNIST

	Performance				Runtime (s)		
Metric	$Accuracy$	$Error_t$	$Error_r$	$Error_f$	Training	Unlearning	Total
Retrain	88.50	11.50	8.92	11.41	687	629	1,316
Fisher	86.23	13.77	12.61	13.33	671	2,015	2,686
certified removal	77.70	22.30	90.11	89.87	719	181	900
DeltaGrad	84.22	15.78	14.36	15.27	553	141	694
NTK	86.02	13.98	12.81	13.25	671	1,879	2,550
Unrolling SGD	83.44	16.56	41.13	41.00	**356**	63	**419**
SISA	84.46	15.54	14.59	14.52	1,419	1,387	2,806
Adaptive Unlearning	86.02	13.98	12.80	13.25	1,537	1,481	3,018
FedEraser	69.86	30.14	29.88	28.14	677	608	1,285
MCMC unlearning	85.29	14.71	6.62	58.73	621	803	1,424
PCMU	**88.34**	**11.66**	**10.68**	**10.71**	802	**0**	802

Table 2: Performance with 20% data removal and CNN on Fashion-MNIST

	Performance				Runtime (s)		
Metric	$Accuracy$	$Error_t$	$Error_r$	$Error_f$	Training	Unlearning	Total
Retrain	88.21	11.79	9.75	11.76	687	561	1,248
Fisher	86.02	13.98	12.80	13.25	827	1,939	2,766
certified removal	76.69	23.31	90.29	89.82	711	347	1,058
DeltaGrad	84.11	15.89	14.73	13.21	570	140	710
NTK	86.03	13.97	12.81	13.25	827	1,807	2,634
Unrolling SGD	85.56	14.44	38.27	37.50	**371**	65	**436**
SISA	83.90	16.10	14.64	14.71	1,419	1,351	2,770
Adaptive Unlearning	75.13	24.87	25.37	25.19	1,537	1,419	2,956
FedEraser	71.93	28.07	27.07	27.13	654	586	1,240
MCMC unlearning	84.52	15.48	6.95	63.02	995	778	1,773
PCMU	**88.34**	**11.66**	**10.25**	**11.47**	802	**0**	802

Fashion-MNIST for clothing classification. We train LeNet over CIFAR-10 for image classification. We apply the ResNet-18 architecture on SVHN for street view house number identification. We evaluate the performance of various machine unlearning methods on three datasets with the ratio of data removal between 5% and 20%.

Baselines. We compare the PCMU model with nine state-of-the-art machine unlearning models. **Fisher** is a scrubbing procedure that removes information from the trained weights, without the need to access the original training data, nor to retrain the entire network [44]. **certified removal** provides a strong theoretical guarantee that a model from which data is removed cannot be distinguished from a model that never observed the data to begin with [50]. **DeltaGrad** is a rapid retraining machine learning model based on information cached during the training phase [127]. **NTK** removes dependency on a cohort of training data from a trained deep network that improves upon and generalizes previous methods to different readout functions [45]. **Unrolling SGD** is a taxonomy of approximate unlearning which concludes with verification error as a metric to study as it subsumes a large class of unlearning criteria [114]. **SISA** is a practical approach for unlearning that relies on data sharding and slicing to reduce the computational overhead of unlearning [8]. **Adaptive Unlearning** gives a general reduction from deletion guarantees against adaptive sequences to deletion guarantees against non-adaptive sequences. [51]. **FedEraser** is a federated unlearning methodology that can eliminate the influences of a federated client's data on the global model while significantly reducing the time consumption [78]. **MCMC unlearning** designs an MCMC influence function to characterize the knowledge learned from data, which then delivers the MCMC unlearning algorithm [38]. To our best knowledge, this work is the first to execute one-time operation of simultaneous training and unlearning in advance for a series of machine unlearning requests.

Variants of PCMU model. We evaluate two versions of PCMU to show the strengths of different techniques. PCMU-N uses the basic model with only the gradient quantization. PCMU operates with the full support of both the randomized gradient smoothing and the gradient quantization techniques. Notice that the gradient quantization is a necessary operation to convert continuous gradients to discrete gradient classes for the randomized gradient smoothing in our PCMU model. Thus, we cannot validate the version with only the randomized gradient smoothing.

Evaluation metrics. By following the same settings in several representative machine unlearning models [44, 45, 114, 38], we use four popular measures in machine unlearning to verify the performance of different methods: $Accuracy$, $Error_f$ (classification errors on the forgortten data D_f), $Error_r$ (errors on the remembered data D_r), and $Error_t$ (errors on the test data). Since the model $M_r(D_r)$ (**Retrain**) that uses only the remembered data D_r as the training data retrained from

Table 3: Performance with 10% data removal and LeNet on CIFAR-10

	Performance				Runtime (s)		
Metric	$Accuracy$	$Error_t$	$Error_r$	$Error_f$	Training	Unlearning	Total
Retrain	64.31	35.69	27.44	36.16	846	745	1,591
Fisher	62.39	37.61	33.76	33.78	693	2189	2,882
certified removal	36.91	63.09	90.02	89.66	749	174	923
DeltaGrad	61.46	38.54	23.12	23.92	859	493	1,352
NTK	62.36	37.64	33.76	35.44	693	2,047	2,740
Unrolling SGD	59.69	40.31	41.69	49.80	**511**	168	**679**
SISA	58.01	41.99	34.01	34.40	1,594	1,533	3,127
Adaptive Unlearning	43.35	56.65	55.80	55.21	1,176	293	1,469
FedEraser	51.63	48.37	43.57	48.62	1,190	984	2,174
MCMC unlearning	60.70	39.30	4.53	26.98	1,322	718	2,040
PCMU	**64.33**	**35.67**	**24.21**	**37.68**	903	**0**	903

Table 4: Performance with 20% data removal and LeNet on CIFAR-10

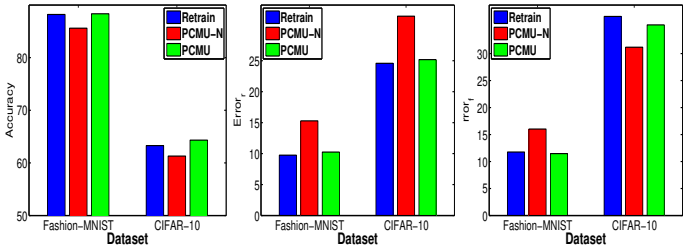
	Performance				Runtime (s)		
Metric	$Accuracy$	$Error_t$	$Error_r$	$Error_f$	Training	Unlearning	Total
Retrain	63.29	36.71	24.59	36.89	846	673	1,519
Fisher	Failed due to out of memory				Failed due to out of memory		
certified removal	36.03	63.97	90.14	89.84	749	430	1,179
DeltaGrad	61.55	38.45	22.61	22.20	864	499	1,363
NTK	Failed due to out of memory				Failed due to out of memory		
Unrolling SGD	60.39	39.61	40.17	47.00	**511**	327	**838**
SISA	57.17	42.83	35.75	35.74	1,594	1,465	3,059
Adaptive Unlearning	41.23	58.77	57.52	59.22	1,176	321	1,497
FedEraser	51.66	48.34	48.53	50.14	1,190	980	2,170
MCMC unlearning	61.33	38.67	5.29	30.42	1,322	714	2,766
PCMU	**64.33**	**35.67**	**25.18**	**35.32**	903	**0**	903

scratch has never seen the forgotten data D_f, it is usually used as the gold standard for evaluating the performance of machine unlearning algorithms [44, 38]. Ideally, the accuracy and three errors of the unlearning models should match that of the retrained model $M_r(D_r)$.

Machine unlearning accuracy with varying ratios of data removal. Tables 1-4 exhibit the accuracy obtained by eleven machine unlearning approaches by varying the ratio of unlearning request / data removal between 10% and 20%. Retrain represents the model retrained on only the remembered data D_r from scratch, without the knowledge of the forgotten data D_f. A machine unlearning algorithm with more similar performance to the Retrain model achieves a better unlearning result. It is observed that among ten approaches except the Retrain model, no matter how large the ratios of data removal are, the PCMU method achieves the closest accuracy to the Retrain model in all tests, showing the effectiveness of PCMU to the machine unlearning. Compared to the absolute performance difference between other baselines and the Retrain model, PCMU, on average, achieves at least 5.56% and 15.17% improvement of absolute accuracy difference on Fashion-MNIST and CIFAR-10 respectively. Notice that the accuracy and error on test data by our PCMU keeps unchanged since it performs one-time operation of simultaneous training and unlearning for addressing multiple unlearning requests. In addition, the promising performance of PCMU over Fashion-MNIST and CIFAR-10 implies that PCMU has great potential as a general machine unlearning solution to other image datasets, which is desirable in practice.

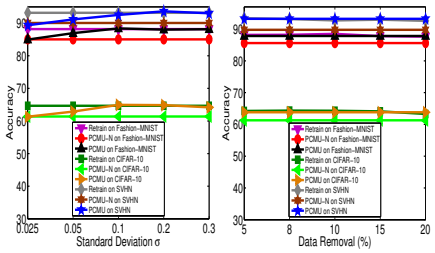
Machine unlearning error with varying ratios of data removal. Tables 1-4 also show the classification errors on the deleted data D_f ($Error_f$), errors on the remembered data D_r ($Error_r$), and errors on the test data ($Error_t$) by eleven machine unlearning methods respectively. We have observed that the performance of our PCMU method behaves similarly and achieves at least 13.73% and 16.01% boost of absolute error difference on two datasets respectively. PCMU substantially outperforms the performance of other baselines in most experiments, especially on the CIFAR-10 dataset. In addition, the errors by our PCMU are not sensitive to the ratio of data removals. This is because that our PCMU method performs one-time operation of simultaneous training and unlearning when addressing a series of machine unlearning requests, as long as the ratio of actual data removals is below the certified budget of data removals in our PCMU. However, other baselines need to sequentially handle these machine unlearning requests one by one.

Ablation study. Figure 1 exhibits the unlearning performance with the Retrain model and two variants of the PCMU model over two datasets of Fashion-MNIST and CIFAR-10 respectively.



(a) $Accuracy$ (b) $Error_r$ (c) $Error_f$

Figure 1: Performance of PCMU variants with 20% data removal on two datasets



(a) Standard Deviation (b) Data Removal (%)

Figure 2: Performance with varying parameters

We have observed the exact PCMU achieves the closest accuracy and errors to the Retrain model over two datasets, which are obviously better than PCMU-N. A reasonable explanation is that our PCMU method utilizes the randomized gradient smoothing and gradient quantization techniques for supporting certified machine unlearning. It further uses the certificates as a guidance to train the machine unlearning model, for guaranteeing that the model parameters and gradients keep unchanged against the data removals within the certified budget.

Running time. Tables 1-4 report the running time achieved by all comparison methods over two dataset to produce machine unlearning results respectively. We observe that PCMU scales well with deep neural network architectures over different image datasets and shows good efficiency for machine unlearning. Our PCMU method achieves better efficiency than most baseline methods, except DeltaGrad and Unrolling SGD. As discussed above, our PCMU method performs one-time operation of simultaneous training and unlearning when addressing a series of machine unlearning requests. However, DeltaGrad and Unrolling SGD need to sequentially handle these machine unlearning requests one by one. This is clearly a computationally expensive process when the number of machine unlearning requests is huge.

Impact of standard deviation. Figure 2 (a) measures the performance effect of standard deviation of the Gaussian distribution in the randomized smoothing for machine unlearning by varying σ from 0.025 to 0.3. Notice that the Retrain and PCMU-N models do not contain the module of randomized smoothing. Thus, their accuracy scores keep unchanged with varying σ. We have witnessed the performance curves by PCMU initially increase quickly and then become stable or even slight drop when σ continuously increases. Initially, a large σ can help utilize the strength of randomized gradient smoothing and quantization for directly training a machine unlearning model in advance. Later on, when σ continues to increase and goes beyond some thresholds, the performance curves become stable. A rational guess is that after the data removals have been already certified at a certain threshold and considered in the training of machine unlearning models, our PCMU model is able to generate a good machine unlearning result. When σ continuously increases, this does not affect the performance of machine unlearning any more.

Impact of data removal ratio. Figure 2 (b) evaluates the accuracy impact of data removal ratios varying from 5% to 20% on three datasets of Fishion-MNIST, CIFAR-10, and SVHN. It is observed that when changing data removal ratios, the accuracy by our PCMU model matchs well with that of the retrained model from scratch. The performance by our PCMU model keeps relatively stable, since our method directly trains a unlearning model based on the certified budget of data removals in advance and performs one-time operation of simultaneous training and unlearning, as long as the ratio of actual data removals is below the certified budget of data removals. Thus, we do not need to re-unlearn the model when an unlearning request is coming. This demonstrates the potential of our PCMU model on addressing a series of machine unlearning requests.

6 Conclusions

In this work, we have proposed a prompt certified machine unlearning algorithm that executes one-time operation of simultaneous training and unlearning in advance. First, we establish a connection between randomized smoothing for certified robustness on classification and randomized smoothing for certified machine unlearning on gradient quantization. Second, we propose a certified machine unlearning model based on randomized data smoothing and gradient quantization. Finally, we present another practical framework of randomized gradient smoothing and quantization, due to the dilemma of producing high confidence certificates in the first framework.

References

[1] M. Alfarra, A. Bibi, N. Khan, P. H. S. Torr, and B. Ghanem. Deformrs: Certifying input deformations with randomized smoothing. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, (AAAI'22), February 22-March 1, 2022, Vancouver, Canada*, 2022.

[2] C. Anil, J. Lucas, and R. B. Grosse. Sorting out lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 291–301, 2019.

[3] S. Axler, P. Bourdon, and W. Ramey. *Harmonic Function Theory*. Wiley, 2001.

[4] T. Baumhauer, P. Schöttle, and M. Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *CoRR*, abs/2002.02730, 2020.

[5] A. Blum, T. Dick, N. Manoj, and H. Zhang. Random smoothing might be unable to certify $l\infty$ robustness for high-dimensional images. *J. Mach. Learn. Res.*, 21:211:1–211:21, 2020.

[6] A. Bojchevski, J. Klicpera, and S. Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 12-18 July 2020*, 2020.

[7] R. Bonic and J. Frampton. Smooth functions on banach manifolds. *Journal of Mathematics and Mechanics*, 15(5):877–898, 1966.

[8] L. Bourtoule, V. Chandrasekaran, C. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *Proceedings of the 42nd IEEE Symposium on Security and Privacy*, 2021.

[9] J. Brophy and D. Lowd. Machine unlearning for random forests. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1092–1104. PMLR, 2021.

[10] R. Bunel, I. Turkaslan, P. H. S. Torr, P. Kohli, and P. K. Mudigonda. A unified view of piecewise linear neural network verification. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4795–4804, 2018.

[11] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pages 463–480. IEEE Computer Society, 2015.

[12] N. Carlini and D. A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pages 3–14, 2017.

[13] G. Cauwenberghs and T. A. Poggio. Incremental and decremental support vector machine learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 409–415. MIT Press, 2000.

[14] C. Chen, B. Kailkhura, R. A. Goldhahn, and Y. Zhou. Certifiably-robust federated adversarial learning via randomized smoothing. In *IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems, MASS 2021, Denver, CO, USA, October 4-7, 2021*, pages 173–179. IEEE, 2021.

[15] C. Chen, F. Sun, M. Zhang, and B. Ding. Recommendation unlearning. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Médini, editors, *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2768–2777. ACM, 2022.

[16] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang. When machine unlearning jeopardizes privacy. In Y. Kim, J. Kim, G. Vigna, and E. Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 896–911. ACM, 2021.

[17] R. Chen, J. Li, J. Yan, P. Li, and B. Sheng. Input-specific robustness certification for randomized smoothing. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, (AAAI'22), February 22-March 1, 2022, Vancouver, Canada*, 2022.

[18] Y. Chen, J. Xiong, W. Xu, and J. Zuo. A novel online incremental and decremental learning algorithm based on variable support vector machine. *Clust. Comput.*, 22(Supplement):7435–7445, 2019.

[19] C. Cheng, G. Nührenberg, and H. Ruess. Maximum resilience of artificial neural networks. In *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*, pages 251–268, 2017.

[20] P. Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studer, and T. Goldstein. Certified defenses for adversarial patches. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[21] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. S. Kankanhalli. Zero-shot machine unlearning. *CoRR*, abs/2201.05629, 2022.

[22] M. Cissé, P. Bojanowski, E. Grave, Y. N. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 854–863, 2017.

[23] J. E. J. Cohen, T. Huster, and R. Cohen. Universal lipschitz approximation in bounded depth neural networks. *CoRR*, abs/1904.04861, 2019.

[24] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 1310–1320, 2019.

[25] F. Croce, M. Andriushchenko, and M. Hein. Provable robustness of relu networks via maximization of linear regions. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2057–2066, 2019.

[26] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[27] S. Dutta, S. Jha, S. Sankaranarayanan, and A. Tiwari. Output range analysis for deep feedforward neural networks. In *NASA Formal Methods - 10th International Symposium, NFM 2018, Newport News, VA, USA, April 17-19, 2018, Proceedings*, pages 121–138, 2018.

[28] K. Dvijotham, S. Gowal, R. Stanforth, R. Arandjelovic, B. O'Donoghue, J. Uesato, and P. Kohli. Training verified learners with learned verifiers. *CoRR*, abs/1805.10265, 2018.

[29] K. Dvijotham, R. Stanforth, S. Gowal, T. A. Mann, and P. Kohli. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 550–559, 2018.

[30] K. D. Dvijotham, J. Hayes, B. Balle, J. Z. Kolter, C. Qin, A. György, K. Xiao, S. Gowal, and P. Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[31] R. Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*, pages 269–286, 2017.

[32] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11423–11434, 2019.

[33] H. Feng, C. Wu, G. Chen, W. Zhang, and Y. Ning. Regularized training and tight certification for randomized smoothed classifier with provable robustness. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3858–3865, 2020.

[34] M. Fischer, M. Baader, and M. T. Vechev. Certified defense to image transformations via randomized smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[35] M. Fischer, M. Baader, and M. Vechev. Scaleable certified segmentation via randomized smoothing. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[36] M. Fischetti and J. Jo. Deep neural networks and mixed integer linear optimization. *Constraints An Int. J.*, 23(3):296–309, 2018.

[37] A. Fromherz, K. Leino, M. Fredrikson, B. Parno, and C. Pasareanu. Fast geometric projections for local robustness certification. In *9th International Conference on Learning Representations, ICLR 2021, Online, May 4-7, 2021, Conference Track Proceedings*, 2021.

[38] S. Fu, F. He, and D. Tao. Knowledge removal in sampling-based bayesian inference. In *10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022.

[39] Z. Gao, R. Hu, and Y. Gong. Certified robustness of graph classification against topology attack with randomized smoothing. In *2020 IEEE Global Communications Conference, GLOBECOM 2020, Taipei, Taiwan, December 8-10, 2020*, 2020.

[40] S. Garg, S. Goldwasser, and P. N. Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In A. Canteaut and Y. Ishai, editors, *Advances in Cryptology - EUROCRYPT 2020 - 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, May 10-14, 2020, Proceedings, Part II*, volume 12106 of *Lecture Notes in Computer Science*, pages 373–402. Springer, 2020.

[41] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. T. Vechev. AI2: safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 3–18, 2018.

[42] A. Ghiasi, A. Shafahi, and T. Goldstein. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[43] A. Ginart, M. Y. Guan, G. Valiant, and J. Zou. Making AI forget you: Data deletion in machine learning. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3513–3526, 2019.

[44] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9301–9309. Computer Vision Foundation / IEEE, 2020.

[45] A. Golatkar, A. Achille, and S. Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 383–398. Springer, 2020.

[46] A. Golatkar, A. Achille, A. Ravichandran, M. Polito, and S. Soatto. Mixed-privacy forgetting in deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 792–801. Computer Vision Foundation / IEEE, 2021.

[47] J. Gong, O. Simeone, and J. Kang. Bayesian variational federated learning and unlearning in decentralized networks. In *22nd IEEE International Workshop on Signal Processing Advances in Wireless Communications, SPAWC 2021, Lucca, Italy, September 27-30, 2021*, pages 216–220. IEEE, 2021.

[48] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. A. Mann, and P. Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *CoRR*, abs/1810.12715, 2018.

[49] L. Graves, V. Nagisetty, and V. Ganesh. Amnesiac machine learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11516–11524. AAAI Press, 2021.

[50] C. Guo, T. Goldstein, A. Y. Hannun, and L. van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3832–3842. PMLR, 2020.

[51] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites. Adaptive machine unlearning. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 16319–16330, 2021.

[52] J. Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 3413–3421, 2020.

[53] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034, 2015.

[54] M. Z. Horváth, M. N. Müller, M. Fischer, and M. T. Vechev. Robust and accurate - compositional architectures for randomized smoothing. In *10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022.

[55] P. Huang, R. Stanforth, J. Welbl, C. Dyer, D. Yogatama, S. Gowal, K. Dvijotham, and P. Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4081–4091, 2019.

[56] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I*, pages 3–29, 2017.

[57] Z. Izzo, M. A. Smart, K. Chaudhuri, and J. Zou. Approximate data deletion from machine learning models. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 2021.

[58] J. Jeong and J. Shin. Consistency regularization for certified robustness of smoothed classifiers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[59] J. Jia, X. Cao, B. Wang, and N. Z. Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[60] J. Jia, X. Cao, and N. Z. Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021.

[61] M. Karasuyama and I. Takeuchi. Multiple incremental decremental learning of support vector machines. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 907–915. Curran Associates, Inc., 2009.

[62] G. Katz, C. W. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I*, pages 97–117, 2017.

[63] M. E. Khan and S. Swaroop. Knowledge-adaptation priors. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19757–19770, 2021.

[64] V. Krishnan, A. A. A. Makdah, and F. Pasqualetti. Lipschitz bounds and provably robust training by laplacian smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[65] A. Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[66] A. Kumar, A. Levine, S. Feizi, and T. Goldstein. Certifying confidence via randomized smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[67] A. Kumar, A. Levine, T. Goldstein, and S. Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 5458–5467, 2020.

[68] S. Lang. *Differential and Riemannian Manifolds*. Springer, 1995.

[69] M. Lécuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672, 2019.

[70] G. Lee, Y. Yuan, S. Chang, and T. S. Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 4911–4922, 2019.

[71] A. Levine and S. Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 4585–4593, 2020.

[72] A. Levine and S. Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *9th International Conference on Learning Representations, ICLR 2021, Online, May 4-7, 2021, Conference Track Proceedings*, 2021.

[73] A. Levine, A. Kumar, T. Goldstein, and S. Feizi. Tight second-order certificates for randomized smoothing. *CoRR*, abs/2010.10549, 2020.

[74] B. Li, C. Chen, W. Wang, and L. Carin. Second-order adversarial attack and certifiable robustness. *CoRR*, abs/1809.03113, 2018.

[75] B. Li, C. Chen, W. Wang, and L. Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9459–9469, 2019.

[76] Q. Li, S. Haque, C. Anil, J. Lucas, R. B. Grosse, and J. Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 15364–15376, 2019.

[77] Y. Li, C. Wang, and G. Cheng. Online forgetting process for linear regression models. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 217–225. PMLR, 2021.

[78] G. Liu, X. Ma, Y. Yang, C. Wang, and J. Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *29th IEEE/ACM International Symposium on Quality of Service, IWQOS 2021, Tokyo, Japan, June 25-28, 2021*, pages 1–10. IEEE, 2021.

[79] Y. Liu, Z. Ma, Y. Yang, X. Liu, J. Ma, and K. Ren. Revfrf: Enabling cross-domain random forest training with revocable federated learning. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[80] Y. Liu, M. Fan, C. Chen, X. Liu, Z. Ma, L. Wang, and J. Ma. Backdoor defense with machine unlearning. In *41st IEEE Conference on Computer Communications, INFOCOM 2022, Virtual, May 2-5, 2022*. IEEE, 2022.

[81] Y. Liu, L. Xu, X. Yuan, C. Wang, and B. Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *41st IEEE Conference on Computer Communications, INFOCOM 2022, Virtual, May 2-5, 2022*. IEEE, 2022.

[82] A. Lomuscio and L. Maganti. An approach to reachability analysis for feed-forward relu neural networks. *CoRR*, abs/1706.07351, 2017.

[83] A. Mahadevan and M. Mathioudakis. Certifiable machine unlearning for linear models. *CoRR*, abs/2106.15093, 2021.

[84] T. Maho, T. Furon, and E. L. Merrer. Randomized smoothing under attack: How good is it in pratice? In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Singapore, May 22-27, 2022*, 2022.

[85] N. G. Marchant, B. I. P. Rubinstein, and S. Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, (AAAI'22), February 22-March 1, 2022, Vancouver, Canada*, 2022.

[86] A. Mehra, B. Kailkhura, P. Chen, and J. Hamm. How robust are randomized smoothing based defenses to data poisoning? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13244–13253. Computer Vision Foundation / IEEE, 2021.

[87] J. H. Metzen and M. Yatsura. Efficient certified defenses against patch attacks on image classifiers. In *9th International Conference on Learning Representations, ICLR 2021, Online, May 4-7, 2021, Conference Track Proceedings*, 2021.

[88] M. Mirman, T. Gehr, and M. T. Vechev. Differentiable abstract interpretation for provably robust neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3575–3583, 2018.

[89] J. Mohapatra, C. Ko, T. Weng, P. Chen, S. Liu, and L. Daniel. Higher-order certification for randomized smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[90] J. Mohapatra, C. Ko, L. Weng, P. Chen, S. Liu, and L. Daniel. Hidden cost of randomized smoothing. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 4033–4041. PMLR, 2021.

[91] M. N. Mueller, M. Balunovic, and M. Vechev. Certify or predict: Boosting certified robustness with compositional architectures. In *9th International Conference on Learning Representations, ICLR 2021, Online, May 4-7, 2021, Conference Track Proceedings*, 2021.

[92] S. Neel, A. Roth, and S. Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In V. Feldman, K. Ligett, and S. Sabato, editors, *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR, 2021.

[93] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011*, 2011.

[94] Q. P. Nguyen, B. K. H. Low, and P. Jaillet. Variational bayesian unlearning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[95] Q. P. Nguyen, R. Oikawa, D. M. Divakaran, M. C. Chan, and B. K. H. Low. Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten. In *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22), May 30-June 3, 2022, Nagasaki, Japan*, 2022.

[96] A. Raghunathan, J. Steinhardt, and P. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10900–10910, 2018.

[97] E. Rosenfeld, E. Winston, P. Ravikumar, and J. Z. Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 8230–8241, 2020.

[98] A. Ruoss, M. Baader, M. Balunovic, and M. Vechev. Efficient certification of spatial robustness. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021.

[99] H. Salman, J. Li, I. P. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11289–11300, 2019.

[100] H. Salman, G. Yang, H. Zhang, C. Hsieh, and P. Zhang. A convex relaxation barrier to tight robustness verification of neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9832–9842, 2019.

[101] S. Saralajew, L. Holdijk, and T. Villmann. Fast adversarial robustness certification of nearest prototype classifiers for arbitrary seminorms. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[102] S. Schelter. "amnesia" - machine learning models that can forget user data very fast. In *10th Conference on Innovative Data Systems Research, CIDR 2020, Amsterdam, The Netherlands, January 12-15, 2020, Online Proceedings*. www.cidrdb.org, 2020.

[103] S. Schelter, S. Grafberger, and T. Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In G. Li, Z. Li, S. Idreos, and D. Srivastava, editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 1545–1557. ACM, 2021.

[104] J. Schuchardt, A. Bojchevski, J. Klicpera, and S. Günnemann. Collective robustness certificates. In *9th International Conference on Learning Representations, ICLR 2021, Online, May 4-7, 2021, Conference Track Proceedings*, 2021.

[105] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh. Remember what you want to forget: Algorithms for machine unlearning. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18075–18086, 2021.

[106] T. Shibata, G. Irie, D. Ikami, and Y. Mitsuzumi. Learning with selective forgetting. In Z. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 989–996. ijcai.org, 2021.

[107] G. Singh, T. Gehr, M. Mirman, M. Püschel, and M. T. Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 10825–10836, 2018.

[108] G. Singh, T. Gehr, M. Püschel, and M. T. Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL):41:1–41:30, 2019.

[109] S. Singla and S. Feizi. Robustness certificates against adversarial examples for relu networks. *CoRR*, abs/1902.01235, 2019.

[110] S. Singla and S. Feizi. Second-order provable defenses against adversarial attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 8981–8991, 2020.

[111] W. A. Strauss. *Partial Differential Equations: An Introduction, 2nd Edition*. Wiley, 2007.

[112] P. Súkeník, A. Kuvshinov, and S. Günnemann. Intriguing properties of input-dependent randomized smoothing. *CoRR*, abs/2110.05365, 2021.

[113] J. Teng, G.-H. Lee, and Y. Yuan. l_1 adversarial robustness certificates: a randomized smoothing approach. *OpenReview*, 2019.

[114] A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot. Unrolling SGD: understanding factors influencing machine unlearning. *CoRR*, abs/2109.13398, 2021.

[115] A. Thudi, H. Jia, I. Shumailov, and N. Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. *CoRR*, abs/2110.11891, 2021.

[116] Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6542–6551, 2018.

[117] E. Ullah, T. Mai, A. Rao, R. A. Rossi, and R. Arora. Machine unlearning via algorithmic stability. In M. Belkin and S. Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 4126–4142. PMLR, 2021.

[118] J. Wang, S. Guo, X. Xie, and H. Qi. Federated unlearning via class-discriminative pruning. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Médini, editors, *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 622–632. ACM, 2022.

[119] S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6369–6379, 2018.

[120] S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6369–6379, 2018.

[121] T. Weng, H. Zhang, H. Chen, Z. Song, C. Hsieh, L. Daniel, D. S. Boning, and I. S. Dhillon. Towards fast computation of certified robustness for relu networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5273–5282, 2018.

[122] T. Weng, P. Zhao, S. Liu, P. Chen, X. Lin, and L. Daniel. Towards certificated model robustness against weight perturbations. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6356–6363, 2020.

[123] E. Wong and J. Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5283–5292, 2018.

[124] E. Wong, F. R. Schmidt, J. H. Metzen, and J. Z. Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8410–8419, 2018.

[125] C. Wu, S. Zhu, and P. Mitra. Federated unlearning with knowledge distillation. *CoRR*, abs/2201.09441, 2022.

[126] G. Wu, M. Hashemi, and C. Srinivasa. PUMA: performance unchanged model augmentation for training data removal. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, (AAAI'22), February 22-March 1, 2022, Vancouver, Canada*, 2022.

[127] Y. Wu, E. Dobriban, and S. B. Davidson. Deltagrad: Rapid retraining of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10355–10366. PMLR, 2020.

[128] Y. Wu, A. Bojchevski, A. Kuvshinov, and S. Günnemann. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 3763–3771. PMLR, 2021.

[129] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[130] K. Xu, Z. Shi, H. Zhang, Y. Wang, K. Chang, M. Huang, B. Kailkhura, X. Lin, and C. Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[131] G. Yang, T. Duan, J. E. Hu, H. Salman, I. P. Razenshteyn, and J. Li. Randomized smoothing of all shapes and sizes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 10693–10705, 2020.

[132] M. Ye, C. Gong, and Q. Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3465–3475, 2020.

[133] H. Zeng, J. Su, and F. Huang. Certified defense via latent space randomized smoothing with orthogonal encoders. *CoRR*, abs/2108.00491, 2021.

[134] Y. Zeng, S. Chen, W. Park, Z. Mao, M. Jin, and R. Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022.

[135] R. Zhai, C. Dan, D. He, H. Zhang, B. Gong, P. Ravikumar, C. Hsieh, and L. Wang. MACER: attack-free and scalable robust training via maximizing certified radius. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[136] D. Zhang, M. Ye, C. Gong, Z. Zhu, and Q. Liu. Black-box certification with randomized smoothing: A functional optimization based framework. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[137] H. Zhang, T. Weng, P. Chen, C. Hsieh, and L. Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4944–4953, 2018.

[138] R. Y. Zhang. On the tightness of semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 1, 3, and 4.
   (b) Did you describe the limitations of your work? [Yes] See Section A.7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section A.7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section A.3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section A.3.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We include the citations and URLs of all datasets used in this work and all codes of third-party baselines in Sections 5 and A.6. Since the datasets used are all public datasets and our methodologies and the hyperparameter settings are explicitly described in Section 3, 4, 5, and A.6, our codes and experiments can be easily reproduced on top of a GPU server. We promise to release our open-source codes on GitHub and maintain a project website with detailed documentation for long-term access by other researchers and end-users after the paper is accepted.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section A.6.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section A.6.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Sections 5 and A.6.

(b) Did you mention the license of the assets? [Yes] See Section A.6.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] Since the datasets used are all public datasets and our methodologies and the hyperparameter settings are explicitly described in Section 3, 4, 5, and A.6, our codes and experiments can be easily reproduced on top of a GPU server. We promise to release our open-source codes on GitHub and maintain a project website with detailed documentation for long-term access by other researchers and end-users after the paper is accepted.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]