

---

# Provably Robust Streaming Models with a Sliding Window

---

Anonymous Authors<sup>1</sup>

## Abstract

The literature on provable robustness in machine learning has primarily focused on static prediction problems, such as image classification, in which input samples are assumed to be independent and model performance is measured as an expectation over the input distribution. Robustness certificates are derived for individual input instances with the assumption that the model is evaluated on each instance separately. However, in many deep learning applications such as online content recommendation and stock market analysis, models use historical data to make predictions. Robustness certificates based on the assumption of independent input samples are not directly applicable in such scenarios. In this work, we focus on the provable robustness of machine learning models in the context of data streams, where inputs are presented as a sequence of potentially correlated items. We derive robustness certificates for models that use a fixed-size sliding window over the input stream. Our guarantees hold for the average model performance across the entire stream and are independent of stream size, making them suitable for large data streams. We perform experiments on speech detection and human activity recognition tasks and show that our certificates can produce meaningful performance guarantees against adversarial perturbations.

## 1. Introduction

Deep neural network (DNN) models are increasingly being adopted for real-time decision-making and prediction tasks. Once a neural network is trained, it is often required to make fast predictions on an evolving stream of inputs, as in algorithmic trading (Zhang et al., 2017; Krauss et al., 2017; Korczak & Hemes, 2017; Fischer & Krauss, 2018;

Ozbayoglu et al., 2020), human action recognition (Yang et al., 2015; Ordonez & Roggen, 2016; Ronao & Cho, 2016) and speech detection (Graves & Schmidhuber, 2005; Dennis et al., 2019; Hsiao et al., 2020). However, despite their impressive performance, DNNs are known to malfunction under tiny perturbations of the input, designed to fool them into making incorrect predictions (Szegedy et al., 2014; Biggio et al., 2013; Goodfellow et al., 2015; Madry et al., 2018; Carlini & Wagner, 2017). This vulnerability is not limited to static models like image classifiers and has been demonstrated for streaming models as well (Braverman et al., 2021; Mladenovic et al., 2022; Ben-Eliezer et al., 2020; Ben-Eliezer & Yoge, 2020). Such input corruptions, commonly known as adversarial attacks, make DNNs especially risky for safety-critical applications of streaming models such as health monitoring (Ignatov, 2018; Stamate et al., 2017; Lee et al., 2019; Cai et al., 2020) and autonomous driving (Bojarski et al., 2016; Xu et al., 2017; Janai et al., 2020). What makes the adversarial streaming setting more challenging than the static one is that the adversary can exploit historical data to strengthen its attack. For instance, it could wait for a critical decision-making point, such as a trading algorithm making a buy/sell recommendation or an autonomous vehicle approaching a stop sign, before generating an adversarial perturbation.

Over the years, a long line of research has been dedicated to mitigating the adversarial vulnerabilities of DNNs. These methods seek to improve the empirical robustness of a model by introducing input corruptions during training (Kurakin et al., 2017; Buckman et al., 2018; Guo et al., 2018; Dhillon et al., 2018; Li & Li, 2017; Grosse et al., 2017; Gong et al., 2017). However, such empirical defenses have been shown to break down under stronger adversarial attacks (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018; Tramer et al., 2020). This motivated the study of provable robustness in machine learning which seeks to obtain verifiable guarantees on the predictive performance of a DNN. Several certified defense techniques have been developed over the years, most notable of which are based on convex relaxation (Wong & Kolter, 2018; Raghunathan et al., 2018; Singla & Feizi, 2019; Chiang et al., 2020; Singla & Feizi, 2020), interval-bound propagation (Gowal et al., 2018; Huang et al., 2019; Dvijotham et al., 2018; Mirman et al., 2018) and randomized smoothing (Cohen et al., 2019;

<sup>1</sup>Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

055 Lécuyer et al., 2019; Li et al., 2019; Salman et al., 2019;  
 056 Levine & Feizi, 2021). Most research in provable robustness  
 057 has focused on static prediction tasks like image classifica-  
 058 tion and the streaming machine learning (ML) setting has  
 059 not yet been considered.

060 In this work, we derive provable robustness guarantees for  
 061 the streaming setting where inputs are presented as a se-  
 062 quence of potentially correlated items. Our objective is  
 063 to design robustness certificates that produce guarantees  
 064 on the average model performance over long, potentially  
 065 infinite, data streams. Our threat model is defined as a man-  
 066 in-the-middle adversary present between the DNN and the  
 067 data stream that can perturb the input items before they  
 068 are passed to the DNN. The adversary is constrained by a  
 069 limit on the average perturbation added to the inputs. We  
 070 show that a DNN that randomizes the inputs before making  
 071 predictions is guaranteed to achieve a certain performance  
 072 level for any adversary within the threat model. Unlike  
 073 many randomized smoothing-based approaches that aggre-  
 074 gate predictions over several noised samples ( $\sim 10^6$ ) of the  
 075 input instance, our procedure only requires one sample of  
 076 the randomized input, keeping the computational complex-  
 077 ity of the DNN unchanged. Our certificates are independent  
 078 of the stream length, making them suitable for large streams.  
 079

080 **Technical Challenges:** Provably robustness procedures de-  
 081 veloped for static tasks like image classification assume that  
 082 the inputs are sampled independently from the data distri-  
 083 bution. Robustness certificates are derived for individual  
 084 input instances with the assumption that the DNN is applied  
 085 to each instance separately and the adversarial perturbation  
 086 added to one instance does not affect the DNN’s predic-  
 087 tions on another. However, in the streaming ML setting, the  
 088 prediction at a given time-step is dependent on past input  
 089 items in the data stream and a worst-case adversary can  
 090 exploit this dependence between inputs to adapt its strategy  
 091 and strengthen its attack. A robustness certificate that is  
 092 derived based on the assumption of independence of input  
 093 samples may not hold for such correlated inputs. Thus, there  
 094 is a need to design provable robustness techniques tailored  
 095 specifically for the streaming ML setting.

096 Out of the existing certified robustness techniques, random-  
 097 ized smoothing has become prominent due to its model-  
 098 agnostic nature, scalability for high-dimensional problems  
 099 (Lécuyer et al., 2019), and flexibility to adapt to different  
 100 machine learning paradigms like reinforcement learning  
 101 and structured outputs (Kumar et al., 2021; Wu et al., 2021;  
 102 Kumar & Goldstein, 2021). This makes randomized smoothing  
 103 a suitable candidate for provable robustness in stream-  
 104 ing ML. However, conventional randomized smoothing ap-  
 105 proaches require several evaluations ( $\sim 10^6$ ) of the predic-  
 106 tion model on different noise vectors in order to produce a  
 107 robust output. This significantly increases the computational  
 108 cost.

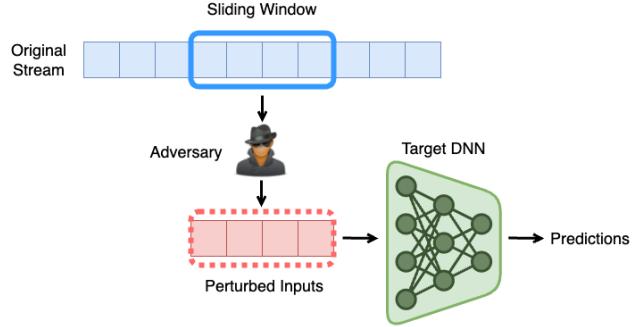


Figure 1. Adversarial Streaming Threat Model.

requirements of the model making them infeasible for real-world streaming applications which require decisions to be made in a short time frame such as high-frequency trading and autonomous driving. Our goal is to obtain robustness guarantees for a simple technique that only adds a single noise vector to the DNN’s input.

Existing works on provable robustness in reinforcement learning (Kumar et al., 2021; Wu et al., 2021) indicate that if the prediction at a given time-step is a function of the entire stream till that step, the robustness guarantees worsen with the length of the stream and become vacuous for large stream sizes. The tightness analysis of these certificates suggests that it might be difficult to achieve robustness guarantees that are independent of the stream size. However, many practical streaming models use only a bounded number of past input items in order to make predictions at a given time step. Recent work by Efroni et al. (2022) has also shown that near-optimal performance can be achieved by only observing a small number of past inputs for several real-world sequential decision-making problems. This raises the natural question:

Can we obtain better certificates if the DNN only used a fixed number of inputs from the stream?

**Our Contributions:** We design a robustness certificate for streaming models that use a fixed-sized sliding window over the data stream to make predictions (see Figure 1). In our setting, the DNN only uses the part of the data stream inside the window at any given time step. We certify the average performance  $Z$  of the model over a stream of size  $t$ :

$$Z = \frac{\sum_{i=1}^t f_i}{t},$$

where each  $f_i$  measures the predictive performance of the DNN at time-step  $i$  as a value in the range  $[0, 1]$ .

The adversary is allowed to perturb the input items inside the window at every time step separately. The strength of

110 the adversary is limited by a bound  $\epsilon$  on the average size of  
 111 the perturbation added:

$$\frac{\sum_{i=1}^t \sum_{k=1}^w d(x_i, x_i^k)}{wt} \leq \epsilon,$$

117 where  $x_i$  and  $x_i^k$  are the input item at time-step  $i$  and its  
 118  $k$ th adversarial perturbation respectively,  $w$  is the window  
 119 size and  $d$  is a distance function to measure the size of  
 120 the adversarial perturbations, e.g.,  $d(x_i, x_i^k) = \|x_i - x_i^k\|_2$ .  
 121 Our adversarial threat model is general enough to subsume  
 122 the scenario where the attacker only perturbs each stream  
 123 element only once as a special case where all  $x_i^k$ 's are set to  
 124 some  $x'_i$ .

125 Our main theoretical result shows that the difference be-  
 126 tween the clean performance of a robust streaming model  
 127  $\tilde{Z}$  and that in the presence of an adversarial attack  $\tilde{Z}_\epsilon$  is  
 128 bounded as follows:

$$|\tilde{Z} - \tilde{Z}_\epsilon| \leq w\psi(\epsilon), \quad (1)$$

133 where  $\psi(\cdot)$  is a concave function that bounds the total varia-  
 134 tion between the smoothing distributions at two input points  
 135 as a function of the distance between them (condition (4)  
 136 in Section 3). Such an upper bound always exists for any  
 137 smoothing distribution. For example, when the distance  
 138 between the points is measured using the  $\ell_2$ -norm and the  
 139 smoothing distribution is a Gaussian  $\mathcal{N}(0, \sigma^2 I)$  with  
 140 variance  $\sigma^2$ , then the concave upper bound is given by  
 141  $\psi(\cdot) = \text{erf}(\cdot/2\sqrt{2}\sigma)$ . Our robustness certificate is inde-  
 142 pendent of the length of the stream and depends only on  
 143 the window size  $w$  and average perturbation size  $\epsilon$ . This  
 144 suggests that streaming ML models with smaller window  
 145 sizes are provably more robust to adversarial attacks.

146 We perform experiments on two real-world applications –  
 147 human activity recognition and speech keyword detection.  
 148 We use the UCI HAR dataset (Reyes-Ortiz et al., 2012) for hu-  
 149 man activity recognition and the Speech commands dataset  
 150 (Warden, 2018) for speech keyword detection. We train  
 151 convolutional networks that take sliding windows as inputs  
 152 and provide robustness guarantees for their performance. In  
 153 our experiments, we consider two different scenarios for  
 154 the adversary. In the first case, the adversary can perturb an  
 155 input only once. In the more general second scenario, the  
 156 adversary can perturb each sliding window separately, making  
 157 it a powerful attacker. We develop strong adversaries  
 158 for both of these scenarios and show their effectiveness in  
 159 our experiments. We then show that our certificates provide  
 160 meaningful robustness guarantees in the presence of such  
 161 strong adversaries. Consistent with our theory, our experi-  
 162 ments also demonstrate that a smaller window size  $w$  gives  
 163 a stronger certificate.

## 2. Related Work

The adversarial streaming setup has been studied extensively in recent years. Mladenovic et al. (2022) designed an attack for transient data streams that do not allow the adversary to re-attack past input items. In their setting, the adversary only has partial knowledge of the target DNN and the perturbations applied in previous time steps are irrevocable. Their objective is to produce an adversarial attack with minimal access to the data stream and the target model. Our goal, on the other hand, is to design a provably robust method that can defend against as general and as strong an adversary as possible. We assume that the adversary has full knowledge of the parameters of the target DNN and can change the adversarial perturbations added in previous time steps. Our threat model includes transient data streams as a special case and applies even to adversaries that only have partial access to the DNN. Streaming adversarial attacks have also been studied for sampling algorithms such as Bernoulli sampling and reservoir sampling (Ben-Eliezer & Yogev, 2020). Here, the goal of the adversary is to create a stream that is unrepresentative of the actual data distribution. Other works have studied the adversarial streaming setup for specific data analysis problems like frequency moment estimation (Ben-Eliezer et al., 2020), submodular maximization (Mitrovic et al., 2017), coresets construction and row sampling (Braverman et al., 2021). In this work, we focus on a robustness certificate for general DNN models in the streaming setting under the conventional notion of adversarial attacks in machine learning literature. We use a sliding-window computational model which has been extensively studied over several years for many streaming applications (Ganardi et al., 2019; Feigenbaum et al., 2005; Datar & Motwani, 2007). Recently Efroni et al. (2022) also showed that a short-term memory is sufficient for several real-world reinforcement learning tasks.

A closely related setting is that of adversarial reinforcement learning. Adversarial attacks have been designed that either directly corrupt the observations of the agent (Huang et al., 2017; Behzadan & Munir, 2017; Pattanaik et al., 2018) or introduce adversarial behavior in a competing agent (Gleave et al., 2020). Robust training methods, such as adding adversarial noise (Kamalaruban et al., 2020; Vinitsky et al., 2020) and training with a learned adversary in an online alternating fashion (Zhang et al., 2021), have been proposed to improve the robustness of RL agents. Several certified defenses have also been developed over the years. For instance, Zhang et al. (2020) developed a method that can certify the actions of an RL agent at each time step under a fixed adversarial perturbation budget. It can certify the total reward obtained at the end of an episode if each of the intermediate actions is certifiably robust. Our streaming formulation allows the adversary to choose the budget at each time step as long as the average perturbation size

165 remains below  $\epsilon$  over time. Our framework also does not  
 166 require each prediction to be robust in order to certify the  
 167 average performance of the DNN. More recent works in  
 168 certified RL can produce robustness guarantees on the total  
 169 reward without requiring every intermediate action to be  
 170 robust or the adversarial budget to be fixed (Kumar et al.,  
 171 2021; Wu et al., 2021). However, these certificates degrade  
 172 for longer streams and the tightness analysis of these certifi-  
 173 cates indicates that this dependence on stream size may not  
 174 be improved. Our goal is to keep the robustness guarantees  
 175 independent of stream size so that they are suitable even for  
 176 large streams.

177 The literature on provable robustness has primarily focused  
 178 on static prediction problems like image classification. One  
 179 of the most prominent techniques in this line of research is  
 180 randomized smoothing. For a given input image, this tech-  
 181 nique aggregates the output of a DNN on several noisy ver-  
 182 sions of the image to produce a robust class label (Lécuyer  
 183 et al., 2019; Cohen et al., 2019). This is the first approach  
 184 that scaled up to high-dimensional image datasets like Im-  
 185 ageNet for  $\ell_2$ -norm bounded adversaries.. It does not make  
 186 any assumptions on the underlying neural network such as  
 187 Lipschitz continuity or a specific architecture, making it  
 188 suitable for conventional DNNs that are several layers deep.  
 189 However, randomized smoothing also suffers some funda-  
 190 mental limitations for higher norms such as the  $\ell_\infty$ -norm  
 191 (Kumar et al., 2020). Due to its flexible nature, randomized  
 192 smoothing has also been adapted for tasks beyond classifi-  
 193 cation, such as segmentation and deep generative modeling,  
 194 with multi-dimensional and structured outputs like images,  
 195 segmentation masks, and language (Kumar & Goldstein,  
 196 2021). For such outputs, robustness certificates are designed  
 197 in terms of a distance metric in the output space such as  
 198 LPIPS distance, intersection-over-union and total variation  
 199 distance. However, provable robustness in the static setting  
 200 assumes a fixed budget on the size of the adversarial  
 201 perturbation for each input instance and does not allow the  
 202 adversary to choose a different budget for each instance. In  
 203 our streaming threat model, we allow the adversary the flex-  
 204 ibility of allocating the adversarial budget to different time  
 205 steps in an effective way, attacking more critical input items  
 206 with a higher budget and conserving its budget at other time  
 207 steps. Recent work on provable robustness against Wasser-  
 208 stein shifts of the data distribution allows the adversary to  
 209 choose the attack budget for each instance differently (Ku-  
 210 mar et al., 2022). However, unlike our streaming setting,  
 211 the input instances are drawn independently from the data  
 212 distribution and the adversarial perturbation applied to one  
 213 instance does not impact the performance of the DNN on  
 214 another.

### 3. Preliminaries and Notation

**Streaming ML Setting:** We define a data stream of size  $t$  as a sequence of input items  $x_1, x_2, \dots, x_i, \dots, x_t$  generated one-by-one from an input space  $\mathcal{X}$  over discrete time steps. At each time step  $i$ , a DNN model  $\mu$  makes a prediction that may depend on no more than  $w$  of the previous inputs. We refer to the contiguous block of past input items as a window  $W_i \in \mathcal{X}^{\min(i,w)}$  of size  $w$  defined as follows:

$$W_i = \begin{cases} (x_1, x_2, \dots, x_i) & \text{for } i \leq w \\ (x_{i-w+1}, x_{i-w+2}, \dots, x_i) & \text{otherwise.} \end{cases}$$

The performance of the model  $\mu$  at time step  $i$  is given by a function  $f_i : \mathcal{X}^{\min(i,w)} \rightarrow [0, 1]$  that passes the window  $W_i$  through the model  $\mu$ , compares the prediction with the ground truth and outputs a value in the range  $[0, 1]$ . For instance, in speech recognition, the window  $W_i$  would represent the audio from the past few seconds which gets fed to the model  $\mu$ . The function  $f_i = \mathbf{1}\{\mu(W_i) = y_i\}$  could indicate whether the prediction of  $\mu$  matches the ground truth  $y_i$ . Similarly, in autonomous driving, we can define a performance function  $f_i = \text{IoU}(\mu(W_i), y_i)$  that measures the average intersection-over-union of the segmentation mask of the surrounding environment. We define the overall performance  $Z$  of the model  $\mu$  as an average over the  $t$  time-steps:

$$Z = \frac{\sum_{i=1}^t f_i}{t}.$$

**Threat Model:** An adversary  $A$  is present between the DNN and the data stream which can perturb the inputs with the objective of minimizing the average performance  $Z$  of the DNN (see Figure 1). Let  $x'_i$  be the perturbed input at step  $i$ . We define a constraint on the amount by which the adversary can perturb the inputs as a bound on the average distance between the original input items  $x_i$  and their perturbed versions  $x'_i$ :

$$\frac{\sum_{i=1}^t d(x_i, x'_i)}{t} \leq \epsilon, \quad (2)$$

where  $d$  is a function that measures the distance between a pair of input items from  $\mathcal{X}$ , e.g.,  $d(x_i, x'_i) = \|x_i - x'_i\|_2$ . The adversary seeks to minimize the overall performance  $Z$  of the model without violating the above constraint, i.e.,

$$\min_{A \in \mathcal{A}_\epsilon} \sum_{i=1}^t f_i(A(x_i), A(x_{i-1}), \dots, A(x_{i-w+1}))/t,$$

where  $\mathcal{A}_\epsilon$  is the set of all adversaries satisfying constraint (2). We also study another threat model where the adversary is allowed to attack an input item  $x_i$  in every window that it appears in. We denote the  $k$ -th attack of  $x_i$  as  $x_i^k$  and redefine the above constraint as follows:

$$\frac{\sum_{i=1}^t \sum_{k=1}^w d(x_i, x_i^k)}{wt} \leq \epsilon \quad (3)$$

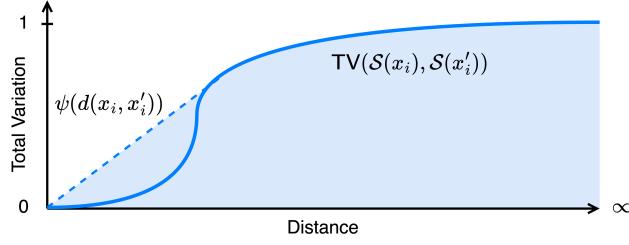


Figure 2. Constructing a concave upper bound  $\psi(\cdot)$  for any smoothing distribution  $\mathcal{S}$ .

This threat model is more general than the one defined by constraint (2) because it subsumes this constraint as a special case when all  $x_i^k$  are equal to  $x'_i$ . Thus, any robustness guarantee that holds for this stronger threat model must also hold for the previous one.

**Robustness Procedure:** Our goal is to design a procedure that has provable robustness guarantees against the above threat models. We define a robust prediction model  $\tilde{\mu}$ : Given an input  $x_i \in \mathcal{X}$ , we sample a point  $\tilde{x}_i$  from a probability distribution  $\mathcal{S}(x_i)$  around  $x_i$  (e.g.,  $\mathcal{N}(x_i, \sigma^2 I)$ ) and evaluate the model  $\mu$  on  $\tilde{x}_i$ . Define the performance of  $\tilde{\mu}$  at time-step  $i$  to be the expected value of  $f_i$  under the randomized inputs, i.e.,

$$\tilde{f}_i = \mathbb{E}_{\tilde{x}_i \sim \mathcal{S}(x_i)} [f_i(\tilde{x}_i, \tilde{x}_{i-1}, \dots, \tilde{x}_{i-w+1})]$$

and the overall performance as  $\tilde{Z} = \sum_{i=1}^t \tilde{f}_i / t$ .

Let  $\psi(\cdot)$  be a concave function bounding the total variation between the distributions  $\mathcal{S}(x_i)$  and  $\mathcal{S}(x'_i)$  as a function of the distance between them, i.e.,

$$TV(\mathcal{S}(x_i), \mathcal{S}(x'_i)) \leq \psi(d(x_i, x'_i)). \quad (4)$$

Such a bound always exists regardless of the shape of the smoothing distribution because as the distance between the points  $x_i$  and  $x'_i$  goes from 0 to  $\infty$ , the total variation goes from 0 to 1. A trivial concave bound could be obtained by simply taking the convex hull of the region under the total variation curve (see Figure 2). However, to find a closed-form expression for  $\psi$ , we need to analyze different smoothing distributions and distance functions separately. If the smoothing distribution is a Gaussian  $\mathcal{N}(0, \sigma^2 I)$  with variance  $\sigma^2$  and the distance is measured using the  $\ell_2$ -norm, as in all of our experiments, then  $\psi(\|x_i - x'_i\|_2) = \text{erf}(\|x_i - x'_i\|_2 / 2\sqrt{2}\sigma)$ , where erf is the Gauss error function. For a uniform smoothing distribution within an interval of size  $b$  in each dimension of  $x_i$  and the  $\ell_1$ -distance metric,  $\psi(\|x_i - x'_i\|_1) = \|x_i - x'_i\|_1 / b$ . See Appendix C for proof.

## 4. Robustness Certificate

In this section, we prove robustness guarantees for the simpler threat model defined by constraint (2) where each input item is allowed to be attacked only once. We include complete proofs of our theorems for this threat model in this section for clarity. The proofs for the more general case in the next section use similar techniques and have been included in the appendix. In the following lemma, we bound the change in the performance function  $\tilde{f}_i$  at each time-step  $i$  using the function  $\psi$  and the size of the adversarial perturbation added at each step. For the proof, we first decompose the change in the value of this function into components for each input item. Since each of these components can be expressed as the difference of the expected value of a function in the range  $[0, 1]$  under two probability distributions, they can be bounded by the total variation of these distributions.

**Lemma 4.1.** *The change in each  $\tilde{f}_i$  under an adversary in  $\mathcal{A}_\epsilon$  is bounded as*

$$|\tilde{f}_i(x_i, x_{i-1}, \dots, x_{i-s+1}) - \tilde{f}_i(x'_i, x'_{i-1}, \dots, x'_{i-s+1})| \leq \sum_{j=i}^{i-s+1} \psi(d(x_j, x'_j)),$$

where  $s = \min(i, w)$ .

*Proof.* The left-hand side of the above inequality can be re-written as:

$$\begin{aligned} & |\tilde{f}_i(x_i, x_{i-1}, \dots, x_{i-s+1}) - \tilde{f}_i(x'_i, x'_{i-1}, \dots, x'_{i-s+1})| \\ &= |\tilde{f}_i(x_i, x_{i-1}, \dots, x_{i-s+1}) - \tilde{f}_i(x'_i, x_{i-1}, \dots, x_{i-s+1}) \\ &\quad + \tilde{f}_i(x'_i, x_{i-1}, \dots, x_{i-s+1}) - \tilde{f}_i(x'_i, x'_{i-1}, \dots, x'_{i-s+1})| \\ &= \left| \sum_{j=i}^{i-s+1} \tilde{f}_i(x'_i, \dots, x_j, \dots, x_{i-s+1}) \right. \\ &\quad \left. - \tilde{f}_i(x'_i, \dots, x'_j, \dots, x_{i-s+1}) \right| \\ &\leq \sum_{j=i}^{i-s+1} |\tilde{f}_i(x'_i, \dots, x_j, \dots, x_{i-s+1}) \\ &\quad - \tilde{f}_i(x'_i, \dots, x'_j, \dots, x_{i-s+1})| \end{aligned}$$

The two terms in each summand differ only in the  $j$ th input. Thus, the  $j$ th term in the above summation can be written as the difference of the expected value of some  $[0, 1]$ -function  $q_j$  under the distributions  $\mathcal{S}(x_j)$  and  $\mathcal{S}(x'_j)$ , i.e.,  $|\mathbb{E}_{\tilde{x} \sim \mathcal{S}(x_j)}[q_j(\tilde{x})] - \mathbb{E}_{\tilde{x} \sim \mathcal{S}(x'_j)}[q_j(\tilde{x})]|$ , which can be upper bounded by the total variation between  $\mathcal{S}(x_j)$  and  $\mathcal{S}(x'_j)$ . Here,  $q_j$  is given by:

$$q_j(\chi) = \mathbb{E}[f_i(\tilde{x}'_i, \dots, \tilde{x}'_{j-1}, \chi, \tilde{x}_{j+1}, \dots, \tilde{x}_{i-s+1})],$$

where  $\chi \in \mathcal{X}$  is the  $j$ th input item, the inputs before  $\chi$  are drawn from the respective adversarially shifted smoothing distributions and the inputs after  $\chi$  are drawn from the original distributions, i.e.,  $\tilde{x}'_i \sim \mathcal{S}(x'_i), \dots, \tilde{x}'_{j-1} \sim \mathcal{S}(x'_{j-1})$  and  $\tilde{x}_{j+1} \sim \mathcal{S}(x_{j+1}), \dots, \tilde{x}_{i-s+1} \sim \mathcal{S}(x_{i-s+1})$ .

Without loss of generality, assume  $\mathbb{E}_{\tilde{\chi} \sim \mathcal{S}(x_j)}[q_j(\tilde{\chi})] \geq \mathbb{E}_{\tilde{\chi} \sim \mathcal{S}(x'_j)}[q_j(\tilde{\chi})]$ . Then,

$$\begin{aligned}
 & |\mathbb{E}_{\tilde{\chi} \sim \mathcal{S}(x_j)}[q_j(\tilde{\chi})] - \mathbb{E}_{\tilde{\chi} \sim \mathcal{S}(x'_j)}[q_j(\tilde{\chi})]| \\
 &= \int_{\mathcal{X}} q_j(x) \mu_1(x) dx - \int_{\mathcal{X}} q_j(x) \mu_2(x) dx \\
 &\quad (\mu_1 \text{ and } \mu_2 \text{ are the PDFs of } \mathcal{S}(x_j) \text{ and } \mathcal{S}(x'_j)) \\
 &= \int_{\mathcal{X}} q_j(x) (\mu_1(x) - \mu_2(x)) dx \\
 &= \int_{\mu_1 > \mu_2} q_j(x) (\mu_1(x) - \mu_2(x)) dx \\
 &\quad - \int_{\mu_2 > \mu_1} q_j(x) (\mu_2(x) - \mu_1(x)) dx \\
 &\leq \int_{\mu_1 > \mu_2} \max_{x' \in \mathcal{X}} q_j(x') (\mu_1(x) - \mu_2(x)) dx \\
 &\quad - \int_{\mu_2 > \mu_1} \min_{x' \in \mathcal{X}} q_j(x') (\mu_2(x) - \mu_1(x)) dx \\
 &\leq \int_{\mu_1 > \mu_2} (\mu_1(x) - \mu_2(x)) dz \\
 &\quad (\text{since } \max_{x' \in \mathcal{X}} q_j(x') \leq 1 \text{ and } \min_{x' \in \mathcal{X}} q_j(x') \geq 0) \\
 &= \frac{1}{2} \int_{\mathcal{X}} |\mu_1(x) - \mu_2(x)| dx = \text{TV}(\mathcal{S}(x_1), \mathcal{S}(x_2)).
 \end{aligned}$$

The equality in the last line follows from the fact that  $\int_{\mu_1 > \mu_2} (\mu_1(x) - \mu_2(x)) dx = \int_{\mu_2 > \mu_1} (\mu_2(x) - \mu_1(x)) dx = \frac{1}{2} \int_{\mathcal{X}} |\mu_1(x) - \mu_2(x)| dx$ .

Therefore, from condition (4), we have:

$$\begin{aligned}
 & |\tilde{f}_i(x'_i, \dots, x_j, \dots, x_{i-w+1}) - \tilde{f}_i(x'_i, \dots, x'_j, \dots, x_{i-w+1})| \\
 &\leq \text{TV}(\mathcal{S}(x_j), \mathcal{S}(x'_j)) \leq \psi(d(x_j, x'_j)).
 \end{aligned}$$

This proves the statement of the lemma.  $\square$

Now we use the above lemma to prove the main robustness guarantee. We first decompose the change in the average performance into the average of the differences at each time step. Then we apply lemma 4.1 to bound each difference with the function  $\psi$  of the per-step perturbation size. We then utilize the convex nature of  $\psi$  to convert this average over the performance differences to an average of perturbation sizes, which completes the proof.

**Theorem 4.2.** Let  $\tilde{Z}_\epsilon$  to be the minimum  $\tilde{Z}$  for an adversary in  $\mathcal{A}_\epsilon$ . Then,

$$|\tilde{Z} - \tilde{Z}_\epsilon| \leq w\psi(\epsilon).$$

*Proof.* Let  $\tilde{Z}'$  be the overall performance of  $\tilde{M}$  under an adversary. Then,

$$\begin{aligned}
 |\tilde{Z} - \tilde{Z}'| &= \left| \frac{\sum_{i=1}^t \tilde{f}_i(x_i, x_{i-1}, \dots, x_{i-s+1})}{t} \right. \\
 &\quad \left. - \frac{\sum_{i=1}^t \tilde{f}_i(x'_i, x'_{i-1}, \dots, x'_{i-s+1})}{t} \right| \\
 &\quad (\text{where } s = \min(i, w)) \\
 &\leq \frac{1}{t} \sum_{i=1}^t \left| \tilde{f}_i(x_i, x_{i-1}, \dots, x_{i-s+1}) \right. \\
 &\quad \left. - \tilde{f}_i(x'_i, x'_{i-1}, \dots, x'_{i-s+1}) \right| \\
 &\leq \sum_{i=1}^t \sum_{j=i}^{i-s+1} \psi(d(x_j, x'_j))/t \quad (\text{from lemma 4.1}) \\
 &\leq w \sum_{i=1}^t \psi(d(x_i, x'_i))/t \\
 &\quad (\text{since each term appears at most } w \text{ times}) \\
 &\leq w\psi \left( \sum_{i=1}^t d(x_i, x'_i)/t \right) \\
 &\quad (\psi \text{ is concave and Jensen's inequality})
 \end{aligned}$$

Therefore, for the worst-case adversary in  $\mathcal{A}_\epsilon$ , we have

$$|\tilde{Z} - \tilde{Z}_\epsilon| \leq w\psi(\epsilon)$$

from constraint (2) on the average distance between the original and perturbed inputs.  $\square$

Although the above certificate is designed for the sliding-window computational model for streaming applications, it may also be applied to the static tasks like classification with a fixed adversarial budget for all inputs by setting  $w = 1$ . In Appendix D, we compare the our bound with that obtained by Cohen et al. (2019) for an  $\ell_2$ -norm bounded adversary and a Gaussian smoothing distribution. While the above bound is not tight, our analysis shows that the gap with static  $\ell_2$ -certificate is small for meaningful robustness guarantees.

## 5. Attacking Each Window

Now we consider the case where the adversary is allowed to attack each window seen by the target DNN separately. The threat model in this section is defined using constraint (3). It is able to re-attack an input item  $x_i$  in each new window. Similar to the definition of a window in Section 3, define an adversarially corrupted window  $W'_i$  as:

$$W'_i = \begin{cases} (x_1^i, x_2^{i-1}, \dots, x_i^1) & \text{for } i \leq w \\ (x_{i-w+1}^w, x_{i-w+2}^{w-1}, \dots, x_i^1) & \text{otherwise,} \end{cases}$$

330 where  $x_i^k$  is the  $k^{\text{th}}$  perturbed instance of  $x_i$ .

331 Similar to the certificate derived in Section 4, we first bound  
 332 the change in the per-step performance function and then  
 333 use that result to prove the final robustness guarantee. We  
 334 formulate the following lemma similar to Lemma 4.1 but  
 335 accounting for the fact that each input item can be perturbed  
 336 multiple times.  
 337

**Lemma 5.1.** *The change in each  $\tilde{f}_i$  under an adversary in  $\mathcal{A}_\epsilon$  is bounded as*

$$341 \quad |\tilde{f}_i(W_i) - \tilde{f}_i(W'_i)| \leq \sum_{j=i-s+1}^i \psi(d(x_j, x_j^{i+1-j})),$$

344 where  $s = \min(i, w)$ .

345 The proof is available in Appendix A.

346 We prove the same certified robustness bound as in Section 4  
 347 but the  $\epsilon$  here is defined according to constraint (3).

349 **Theorem 5.2.** *Let  $\tilde{Z}_\epsilon$  to be the minimum  $\tilde{Z}$  for an adversary  
 350 in  $\mathcal{A}_\epsilon$ . Then,*

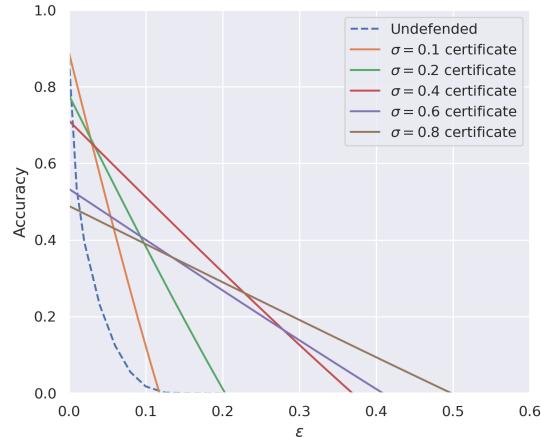
$$352 \quad |\tilde{Z} - \tilde{Z}_\epsilon| \leq w\psi(\epsilon).$$

353 The proof is available in Appendix B.

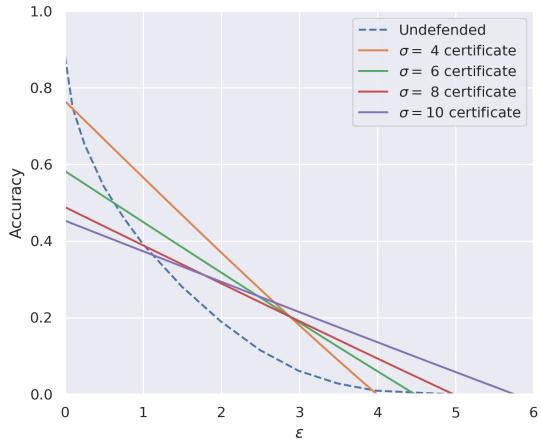
## 356 6. Experiments

357 We test our certificates for two streaming tasks – speech  
 358 keyword detection and human activity recognition. We use  
 359 a subset of the Speech commands dataset (Warden, 2018)  
 360 for our speech keyword detection task. The subset we use  
 361 contains ten keyword classes, corresponding to utterances  
 362 of numbers from zero to nine recorded at a sample rate of  
 363 16 kHz. This dataset also contains noise clips such as audio  
 364 of running tap water and exercise bike. We add these noise  
 365 clips to the speech audio to simulate real-world scenarios  
 366 and stitch them together to generate longer audio clips. We  
 367 use the UCI HAR dataset (Reyes-Ortiz et al., 2012) for  
 368 human activity recognition. This contains a 6-D triaxial  
 369 accelerometer and gyroscope readings measured with  
 370 human subjects. The objective in HAR is to recognize various  
 371 human activities based on sensor readings. The UCI HAR  
 372 dataset contains signals recorded at 50 Hz that correspond  
 373 to six human activities such as standing, sitting, laying,  
 374 walking, walking up, and walking down.

376 We use the M5 network described in (Dai et al., 2017) with  
 377 an SGD optimizer and an initial learning rate of 0.1, which  
 378 we anneal using a cosine scheduler. For the speech detection  
 379 task, we train a M5 network with 128 channels for 30 epochs  
 380 with a batch size of 128. For the human activity recognition  
 381 task, we use a M5 network with 32 channels for 30 epochs  
 382 with a batch size of 256. We apply isotropic Gaussian noise  
 383 for smoothing and use the  $\ell_2$ -norm to define the average  
 384



(a) Speech keyword detection



(b) Human activity recognition

Figure 3. Certificates against online adversarial attacks for varying smoothing noises. Here we can perturb each input only once. The average size of perturbation is computed as per equation 2.

distance measure  $d$ . For the speech keyword detection task, we use smoothing noises with standard deviations of 0.1, 0.2, 0.4, 0.6, and 0.8. For the human activity recognition task, we use smoothing noises with standard deviations of 4, 6, 8, and 10. See Appendix E for more details on the experiments. We compute certificates for both scenarios, where the input is attacked only once and where each window can be attacked with the ability to re-attack inputs. These experiments show that our certificates provide meaningful guarantees against adversarial perturbations.

### 6.1. Attacking an input only once

We evaluate the robustness of undefended models using a custom-made attack that is constrained by the  $\ell_2$ -norm budget, as described in equation 2. To adhere to this constraint at each time-step  $j$ , the attacker must only perturb

---

385   **Algorithm 1** Our streaming attack  
 386   **Input:** time-step  $j$ , clean inputs  $x_j, x_{j-1}, \dots, x_{j-w+1}$ ,  
 387   perturbed inputs  $x'_{j-1}, \dots, x'_{j-w+1}$ , attack budget  $\epsilon$ ,  
 388   search parameter  $\alpha \in \mathbb{N}$ .  
 389    $d_{j-1} = \sum_{i=1}^{j-1} d(x_i, x'_i)$   
 390    $\text{budget}_j = j\epsilon - d_{j-1}$   
 391   **for**  $i = 0$  **to**  $\alpha$  **do**  
 392      $\epsilon' = \frac{i}{\alpha} \cdot \text{budget}_j$   
 393      $x = \arg \min_x f_j(x, \dots, x'_{j-w+1})$  s.t.  $d(x, x_j) \leq \epsilon'$   
 394     **if**  $f_j(x'_j, \dots, x'_{j-w+1}) = 0$  **then**  
 395        $x'_j = x$   
 396       **break**  
 397     **else**  
 398        $x'_j = x_j$   
 399     **end if**  
 400   **end for**

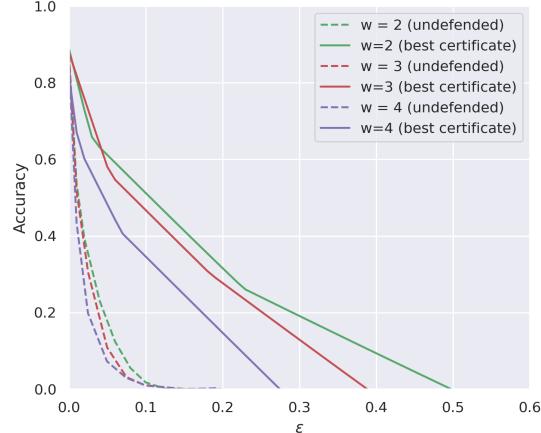
---

403   the input  $x_j$ , since the previous inputs  $(x_{j-w+1}, \dots, x_{j-1})$   
 404   have already been perturbed. This creates a significant challenge  
 405   in creating a strong adversary. We design an adversary that only perturbs  
 406   the last input  $x_j$  at every time-step  $j$  using projected gradient descent to minimize  $f_j$ . In our  
 407   experiments, we set  $f_j = 1$  if the model outputs the correct class and  $f_j = 0$  when the model misclassifies. We  
 408   linearly search using grid search parameter  $\alpha$  for the smallest distance  $d(x_j, x'_j)$  such that the input  $(x'_{j-w+1}, \dots, x'_j)$   
 409   leads to a misclassification at time-step  $j$ . We perturb  $x_j$  if  
 410    $(x'_{j-w+1}, \dots, x'_j)$  leads to misclassification and the average  
 411   distance budget at time-step  $j$  is less than  $\epsilon$ . Else, we do not  
 412   perturb  $x_j$ . In this manner, our attack perturbs the streaming  
 413   input in a greedy fashion. See Algorithm 1 for details.

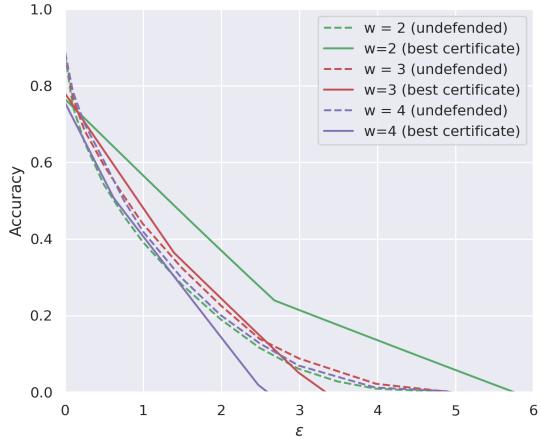
417   We conduct our streaming attack on the keyword recognition  
 418   task with a window size of  $w = 2$ , where each input  $x_j$   
 419   is a 4000-dimensional vector in the range  $[0, 1]$ . We also  
 420   perform the attack on the human activity recognition task  
 421   with  $w = 2$ , where each input  $x_j$  is a 250x6-dimensional  
 422   matrix. We use search parameter  $\alpha = 15$ . We plot the  
 423   results of our certificates for various smoothing noises (see  
 424   Figure 3). Note that the attack budget  $\epsilon$  is calculated as  
 425   per the definition in equation 2. In Figure 4, we also plot  
 426   our best certificates across various smoothing noises for  
 427   different window sizes  $w$ . This plot supports our theory  
 428   that streaming models with smaller window sizes are more  
 429   robust to adversarial perturbations. Figure 7 in Appendix  
 430   F shows that the empirical performance of smooth models  
 431   after the online adversarial attack is lower bound by our  
 432   certificates. These plots validate our certificates.  
 433

## 434   6.2. Attacking each window

435   Now, we perform experiments for the attack setting de-  
 436   scribed in Section 5. Note that here we need to calculate the  
 437   attack budget  $\epsilon$  based on equation 3. In this setting, we can  
 438



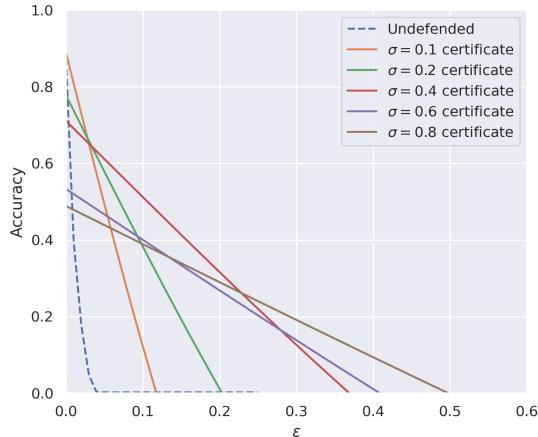
(a) Speech keyword detection



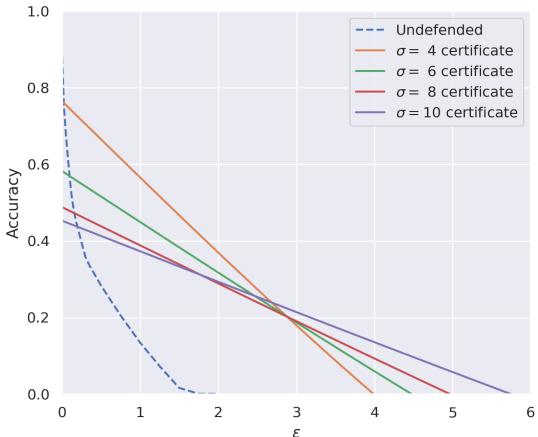
(b) Human activity recognition

Figure 4. Best certificates across varying smoothing noises for different window sizes. Streaming models with smaller window sizes are more robust to adversarial perturbations.

re-attack an input for every window, making it a stronger attack. To attack the undefended models, we search for window perturbations that lead to misclassification using a minimum distance budget. Similar to our previous attack in Section 6.1, we only perturb a window at time-step  $j$  if the average window distance at time-step  $j$  is less than  $\epsilon$ . Also, we do not perturb a window if the window can not be perturbed to reduce the performance  $f_j$ . In Figure 5, we plot our certificates for this attack setting along with the accuracy of the undefended model for different attack budgets. These experiments show that our certificates produce meaningful performance guarantees against adversarial perturbations even if an attacker has the ability to re-attack the inputs.



(a) Speech keyword detection



(b) Human activity recognition

Figure 5. Certificates against online adversarial attacks for varying smoothing noises. Here we attack each window with the ability to re-attack inputs. The average size of perturbation is computed as per equation 3.

## 7. Conclusion

In this work, we design provable robustness guarantees for streaming machine learning models with a sliding window. Our certificates provide a lower bound on the average performance of a streaming DNN model in the presence of an adversary. The adversarial budget in our threat model is defined in terms of the average size of the perturbations added to the input items across the entire stream. This allows the adversary to allocate a different budget to each input item and leads to a more general threat model than the static setting. Our certificates are independent of the stream length and can handle long, potentially infinite, streams. They are also applicable for adversaries that are allowed to re-attack past inputs leading to strong robustness guarantees covering a wide range of attack strategies.

Our robustness procedure simply augments the inputs with random noise. Unlike conventional randomized smoothing techniques, our method only requires one noised sample per prediction keeping the computational requirements of the DNN model unchanged. It does not make any assumptions about the DNN model such as Lipschitz continuity or a specific architecture and is applicable for conventional DNNs that are several layers deep. Our experimental results show that our certificates can obtain meaningful robustness guarantees for real-world streaming applications. Our results show that the certified performance of a robust model depends only on the window size and smaller windows lead to models that are provably more robust than larger windows.

To the best of our knowledge, this is the first attempt at designing adversarial robustness certificates for the streaming setting. We note that our robustness guarantees are not proven to be tight and could be improved upon by future work. We hope our work inspires further investigation into provable robustness methods for streaming ML models.

## References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Behzadan, V. and Munir, A. Vulnerability of deep reinforcement learning to policy induction attacks. In Perner, P. (ed.), *Machine Learning and Data Mining in Pattern Recognition - 13th International Conference, MLD-M 2017, New York, NY, USA, July 15-20, 2017, Proceedings*, volume 10358 of *Lecture Notes in Computer Science*, pp. 262–275. Springer, 2017. doi: 10.1007/978-3-319-62416-7\_19. URL [https://doi.org/10.1007/978-3-319-62416-7\\_19](https://doi.org/10.1007/978-3-319-62416-7_19).

Ben-Eliezer, O. and Yoge, E. The adversarial robustness of sampling. In Suciu, D., Tao, Y., and Wei, Z. (eds.), *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020*, pp. 49–62. ACM, 2020. doi: 10.1145/3375395.3387643. URL <https://doi.org/10.1145/3375395.3387643>.

Ben-Eliezer, O., Jayaram, R., Woodruff, D. P., and Yoge, E. A framework for adversarially robust streaming algorithms. In Suciu, D., Tao, Y., and Wei, Z. (eds.), *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020*, pp. 63–80. ACM,

- 495      2020. doi: 10.1145/3375395.3387658. URL <https://doi.org/10.1145/3375395.3387658>.
- 496
- 497
- 498 Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic,  
499 N., Laskov, P., Giacinto, G., and Roli, F. Evasion at-  
500 tacks against machine learning at test time. In Blockeel,  
501 H., Kersting, K., Nijssen, S., and Zelezný, F. (eds.), *Ma-  
502 chine Learning and Knowledge Discovery in Databases  
503 - European Conference, ECML PKDD 2013, Prague,  
504 Czech Republic, September 23-27, 2013, Proceedings,  
505 Part III*, volume 8190 of *Lecture Notes in Computer  
506 Science*, pp. 387–402. Springer, 2013. doi: 10.1007/  
507 978-3-642-40994-3\\_25. URL [https://doi.org/10.1007/978-3-642-40994-3\\_25](https://doi.org/10.1007/978-3-642-40994-3_25).
- 508
- 509 Bojarski, M., Testa, D. D., Dworakowski, D., Firner,  
510 B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M.,  
511 Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba,  
512 K. End to end learning for self-driving cars. *CoRR*,  
513 abs/1604.07316, 2016. URL <http://arxiv.org/abs/1604.07316>.
- 514
- 515 Braverman, V., Hassidim, A., Matias, Y., Schain, M., Silwal,  
516 S., and Zhou, S. Adversarial robustness of streaming al-  
517 gorithms through importance sampling. In Beygelzimer,  
518 A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.),  
519 *Advances in Neural Information Processing Systems*,  
520 2021. URL [https://openreview.net/forum?id=83A-0x6Pfi\\_](https://openreview.net/forum?id=83A-0x6Pfi_).
- 521
- 522 Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. J.  
523 Thermometer encoding: One hot way to resist adversarial  
524 examples. In *6th International Conference on Learning  
525 Representations, ICLR 2018, Vancouver, BC, Canada,  
526 April 30 - May 3, 2018, Conference Track Proceedings*,  
527 2018.
- 528
- 529 Cai, L., Gao, J., and Zhao, D. A review of the applica-  
530 tion of deep learning in medical image classification and  
531 segmentation. *Annals of translational medicine*, 8(11),  
532 2020.
- 533
- 534 Carlini, N. and Wagner, D. A. Adversarial examples are  
535 not easily detected: Bypassing ten detection methods.  
536 In *Proceedings of the 10th ACM Workshop on Artificial  
537 Intelligence and Security, AISec@CCS 2017, Dallas, TX,  
538 USA, November 3, 2017*, pp. 3–14, 2017.
- 539
- 540 Chiang, P.-y., Ni, R., Abdelkader, A., Zhu, C., Studer,  
541 C., and Goldstein, T. Certified defenses for adver-  
542 sarial patches. In *8th International Conference on Learning  
543 Representations*, 2020.
- 544
- 545 Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial  
546 robustness via randomized smoothing. In Chaudhuri, K.  
547 and Salakhutdinov, R. (eds.), *Proceedings of the 36th In-  
548 ternational Conference on Machine Learning*, volume 97  
549 of *Proceedings of Machine Learning Research*, pp. 1310–  
1320, Long Beach, California, USA, 09–15 Jun 2019.  
PMLR.
- 550
- Dai, W., Dai, C., Qu, S., Li, J., and Das, S. Very deep con-  
551 volutional neural networks for raw waveforms. In *2017  
552 IEEE international conference on acoustics, speech and  
553 signal processing (ICASSP)*, pp. 421–425. IEEE, 2017.
- Datar, M. and Motwani, R. *The Sliding-Window Compu-  
554 tation Model and Results*, pp. 149–167. Springer US,  
555 Boston, MA, 2007. ISBN 978-0-387-47534-9. doi:  
556 10.1007/978-0-387-47534-9\_8. URL [https://doi.org/10.1007/978-0-387-47534-9\\_8](https://doi.org/10.1007/978-0-387-47534-9_8).
- Dennis, D., Acar, D. A. E., Mandikal, V., Sadasivan,  
557 V. S., Saligrama, V., Simhadri, H. V., and Jain, P.  
558 Shallow rnn: Accurate time-series classification  
559 on resource constrained devices. In Wallach, H.,  
560 Larochelle, H., Beygelzimer, A., d’Alché-Buc, F.,  
561 Fox, E., and Garnett, R. (eds.), *Advances in Neural  
562 Information Processing Systems*, volume 32. Curran As-  
563 sociates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/76d7c0780ceb8fbf964c102ebc16d75f-Paper.pdf>.
- Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein,  
564 J., Kossaifi, J., Khanna, A., and Anandkumar, A. Stochas-  
565 tic activation pruning for robust adversarial defense. In  
566 *6th International Conference on Learning Representa-  
567 tions, ICLR 2018, Vancouver, BC, Canada, April 30 -  
568 May 3, 2018, Conference Track Proceedings*, 2018.
- Dvijotham, K., Gowal, S., Stanforth, R., Arandjelovic, R.,  
569 O’Donoghue, B., Uesato, J., and Kohli, P. Training veri-  
570 fied learners with learned verifiers, 2018.
- Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S.  
571 Provably reinforcement learning with a short-term mem-  
572 ory. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári,  
573 C., Niu, G., and Sabato, S. (eds.), *International Confer-  
574 ence on Machine Learning, ICML 2022, 17-23 July 2022,  
575 Baltimore, Maryland, USA*, volume 162 of *Proceedings  
576 of Machine Learning Research*, pp. 5832–5850. PMLR,  
577 2022. URL <https://proceedings.mlr.press/v162/efroni22a.html>.
- Feigenbaum, J., Kannan, S., and Zhang, J. Computing diameter in the streaming and sliding-window mod-  
578 els. *Algorithmica*, 41(1):25–41, 2005. doi: 10.1007/s00453-004-1105-2. URL <https://doi.org/10.1007/s00453-004-1105-2>.
- Fischer, T. and Krauss, C. Deep learning with  
579 long short-term memory networks for financial mar-  
580 ket predictions. *European Journal of Operational  
581 Research*, 2022.

- 550      Research, 270(2):654–669, 2018. ISSN 0377-  
 551      2217. doi: <https://doi.org/10.1016/j.ejor.2017.11.054>. URL <https://www.sciencedirect.com/science/article/pii/S0377221717310652>.
- 552      Ganardi, M., Hucke, D., and Lohrey, M. Derandomization for sliding window algorithms with strict correctness. In van Bevern, R. and Kucherov, G. (eds.), *Computer Science - Theory and Applications - 14th International Computer Science Symposium in Russia, CSR 2019, Novosibirsk, Russia, July 1-5, 2019, Proceedings*, volume 11532 of *Lecture Notes in Computer Science*, pp. 237–249. Springer, 2019. doi: 10.1007/978-3-030-19955-5\_21. URL [https://doi.org/10.1007/978-3-030-19955-5\\_21](https://doi.org/10.1007/978-3-030-19955-5_21).
- 553      Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=HJgEMpVFwB>.
- 554      Gong, Z., Wang, W., and Ku, W. Adversarial and clean data are not twins. *CoRR*, abs/1704.04960, 2017.
- 555      Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- 556      Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models, 2018.
- 557      Graves, A. and Schmidhuber, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2005.06.042>. URL <https://www.sciencedirect.com/science/article/pii/S0893608005001206>. IJCNN 2005.
- 558      Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. D. On the (statistical) detection of adversarial examples. *CoRR*, abs/1702.06280, 2017.
- 559      Guo, C., Rana, M., Cissé, M., and van der Maaten, L. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.
- 560      Hsiao, R., Can, D., Ng, T., Travadi, R., and Ghoshal, A. Online automatic speech recognition with listen, attend and spell model. *IEEE Signal Processing Letters*, 27: 1889–1893, 2020.
- 561      Huang, P., Stanforth, R., Welbl, J., Dyer, C., Yogatama, D., Gowal, S., Dvijotham, K., and Kohli, P. 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*, pp. 4081–4091, 2019. doi: 10.18653/v1/D19-1419. URL <https://doi.org/10.18653/v1/D19-1419>.
- 562      Huang, S. H., Papernot, N., Goodfellow, I. J., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=ryvlRyBK1>.
- 563      Ignatov, A. Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, 62:915–922, 2018. ISSN 1568-4946. doi: <https://doi.org/10.1016/j.asoc.2017.09.027>. URL <https://www.sciencedirect.com/science/article/pii/S1568494617305665>.
- 564      Janai, J., Güney, F., Behl, A., Geiger, A., et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1-3):1–308, 2020.
- 565      Kamalaruban, P., Huang, Y.-T., Hsieh, Y.-P., Rolland, P., Shi, C., and Cevher, V. Robust reinforcement learning via adversarial training with langevin dynamics. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8127–8138. Curran Associates, Inc., 2020.
- 566      Korczak, J. and Hemes, M. Deep learning for financial time series forecasting in a-trader system. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 905–912, 2017. doi: 10.15439/2017F449.
- 567      Krauss, C., Do, X. A., and Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702, 2017. ISSN 0377-2217. doi: <https://doi.org/10.1016/j.ejor.2016.10.031>. URL <https://www.sciencedirect.com/science/article/pii/S0377221716308657>.

- 605 Kumar, A. and Goldstein, T. Center smoothing: Certified ro-  
 606 bustness for networks with structured outputs. *Advances*  
 607 in *Neural Information Processing Systems*, 34, 2021.
- 608
- 609 Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Curse of  
 610 dimensionality on randomized smoothing for certifiable  
 611 robustness. In *Proceedings of the 37th International*  
 612 *Conference on Machine Learning, ICML 2020, 13-18*  
 613 *July 2020, Virtual Event*, volume 119 of *Proceedings*  
 614 of *Machine Learning Research*, pp. 5458–5467. PMLR,  
 615 2020. URL <http://proceedings.mlr.press/v119/kumar20b.html>.
- 616
- 617 Kumar, A., Levine, A., and Feizi, S. Policy smoothing  
 618 for provably robust reinforcement learning. *CoRR*,  
 619 abs/2106.11420, 2021. URL <https://arxiv.org/abs/2106.11420>.
- 620
- 621 Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Certi-  
 622 fying model accuracy under distribution shifts. *CoRR*,  
 623 abs/2201.12440, 2022. URL <https://arxiv.org/abs/2201.12440>.
- 624
- 625 Kurakin, A., Goodfellow, I. J., and Bengio, S. Adver-  
 626 sarial machine learning at scale. In *5th International*  
 627 *Conference on Learning Representations, ICLR 2017,*  
 628 *Toulon, France, April 24-26, 2017, Conference Track*  
 629 *Proceedings*, 2017. URL <https://openreview.net/forum?id=BJm4T4Kgx>.
- 630
- 631 Lécyuer, M., Atlidakis, V., Geambasu, R., Hsu, D., and  
 632 Jana, S. Certified robustness to adversarial examples  
 633 with differential privacy. In *2019 IEEE Symposium on*  
 634 *Security and Privacy, SP 2019, San Francisco, CA, USA,*  
 635 *May 19-23, 2019*, pp. 656–672, 2019.
- 636
- 637 Lee, G., Nho, K., Kang, B., Sohn, K.-A., and Kim, D.  
 638 Predicting alzheimer’s disease progression using multi-  
 639 modal deep learning approach. *Scientific reports*, 9(1):  
 640 1–12, 2019.
- 641
- 642 Levine, A. and Feizi, S. Improved, deterministic smoothing  
 643 for L1 certified robustness. *CoRR*, abs/2103.10834, 2021.  
 644 URL <https://arxiv.org/abs/2103.10834>.
- 645
- 646 Li, B., Chen, C., Wang, W., and Carin, L. Certified ad-  
 647 versarial robustness with additive noise. In *Advances*  
 648 in *Neural Information Processing Systems 32: Annual*  
 649 *Conference on Neural Information Processing Systems*  
 650 *2019, NeurIPS 2019, 8-14 December 2019, Vancouver,*  
 651 *BC, Canada*, pp. 9459–9469, 2019.
- 652
- 653 Li, D., Langlois, T. R., and Zheng, C. Scene-aware audio  
 654 for 360 videos. *ACM Transactions on Graphics (TOG)*,  
 655 37(4):1–12, 2018.
- 656
- 657 Li, X. and Li, F. Adversarial examples detection in deep  
 658 networks with convolutional filter statistics. In *IEEE In-*  
 659 *ternational Conference on Computer Vision, ICCV 2017,*  
 660 *Venice, Italy, October 22-29, 2017*, pp. 5775–5783, 2017.
- 661
- 662 Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and  
 663 Vladu, A. Towards deep learning models resistant to  
 664 adversarial attacks. In *6th International Conference on*  
 665 *Learning Representations, ICLR 2018, Vancouver, BC,*  
 666 *Canada, April 30 - May 3, 2018, Conference Track*  
 667 *Proceedings*, 2018.
- 668
- 669 Mirman, M., Gehr, T., and Vechev, M. Differentiable  
 670 abstract interpretation for provably robust neural  
 671 networks. In Dy, J. and Krause, A. (eds.), *Pro-*  
 672 *ceedings of the 35th International Conference on Ma-*  
 673 *chine Learning*, volume 80 of *Proceedings of Machine*  
 674 *Learning Research*, pp. 3578–3586. PMLR, 10–15 Jul  
 675 2018. URL <http://proceedings.mlr.press/v80/mirman18b.html>.
- 676
- 677 Mitrovic, S., Bogunovic, I., Norouzi-Fard, A., Tarnawski, J.,  
 678 and Cevher, V. Streaming robust submodular maximiza-  
 679 tion: A partitioned thresholding approach. In Guyon, I.,  
 680 von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R.,  
 681 Vishwanathan, S. V. N., and Garnett, R. (eds.), *Ad-*  
 682 *vances in Neural Information Processing Systems 30: Annu-*  
 683 *al Conference on Neural Information Processing Systems*  
 684 *2017, December 4-9, 2017, Long Beach, CA, USA*, pp.  
 685 4557–4566, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3baa271bc35fe054c86928f7016e8ae6-Abstract.html>.
- 686
- 687 Mladenovic, A., Bose, J., berard, H., Hamilton, W. L.,  
 688 Lacoste-Julien, S., Vincent, P., and Gidel, G. On-  
 689 line adversarial attacks. In *International Conference*  
 690 *on Learning Representations*, 2022. URL [https://openreview.net/forum?id=bYGSzbCM\\_i](https://openreview.net/forum?id=bYGSzbCM_i).
- 691
- 692 Ordóñez, F. J. and Roggen, D. Deep convolutional and  
 693 lstm recurrent neural networks for multimodal wearable  
 694 activity recognition. *Sensors*, 16(1), 2016. ISSN 1424-  
 695 8220. doi: 10.3390/s16010115. URL <https://www.mdpi.com/1424-8220/16/1/115>.
- 696
- 697 Ozbayoglu, A. M., Gudelek, M. U., and Sezer, O. B. Deep  
 698 learning for financial applications: A survey. *Applied Soft*  
 699 *Computing*, 93:106384, 2020.
- 700
- 701 Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and  
 702 Chowdhary, G. Robust deep reinforcement learning  
 703 with adversarial attacks. In André, E., Koenig, S., Das-  
 704 tani, M., and Sukthankar, G. (eds.), *Proceedings of the*  
 705 *17th International Conference on Autonomous Agents*  
 706 *and MultiAgent Systems, AAMAS 2018, Stockholm, Swe-*  
 707 *den, July 10-15, 2018*, pp. 2040–2042. International

- 660 Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018. URL <http://dl.acm.org/citation.cfm?id=3238064>.  
 661  
 662  
 663
- 664 Raghunathan, A., Steinhardt, J., and Liang, P. Semidefinite relaxations for certifying robustness to adversarial examples. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 10900–10910, Red Hook, NY, USA, 2018. Curran Associates Inc.  
 665  
 666  
 667  
 668  
 669
- 670 Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., and Parra, X. Uci machine learning repository: Human activity recognition using smartphones data set, 2012.  
 671  
 672  
 673
- 674 Ronao, C. A. and Cho, S.-B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235–244, 2016. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2016.04.032>. URL <https://www.sciencedirect.com/science/article/pii/S0957417416302056>.  
 675  
 676  
 677  
 678  
 679
- 680 Salman, H., Li, J., Razenshteyn, I. P., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. 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*, pp. 11289–11300, 2019.  
 681  
 682  
 683  
 684  
 685  
 686  
 687  
 688  
 689
- 690 Singla, S. and Feizi, S. Robustness certificates against adversarial examples for relu networks. *CoRR*, abs/1902.01235, 2019.  
 691  
 692  
 693  
 694  
 695  
 696  
 697  
 698  
 699
- 700 Singla, S. and Feizi, S. 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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8981–8991. PMLR, 2020. URL <http://proceedings.mlr.press/v119/singla20a.html>.  
 701  
 702  
 703  
 704  
 705  
 706  
 707
- 708 Stamate, C., Magoulas, G., Kueppers, S., Nomikou, E., Daskalopoulos, I., Luchini, M., Moussouri, T., and Rousos, G. Deep learning parkinson's from smartphone data. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 31–40, 2017. doi: 10.1109/PERCOM.2017.7917848.  
 709  
 710  
 711  
 712  
 713  
 714
- 715 Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- 716 Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses, 2020.  
 717  
 718  
 719 Uesato, J., O'Donoghue, B., Kohli, P., and van den Oord, A. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, pp. 5032–5041, 2018.  
 720  
 721  
 722 Vinitsky, E., Du, Y., Parvate, K., Jang, K., Abbeel, P., and Bayen, A. M. Robust reinforcement learning using adversarial populations. *CoRR*, abs/2008.01825, 2020. URL <https://arxiv.org/abs/2008.01825>.  
 723  
 724 Warden, P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*, April 2018. URL <https://arxiv.org/abs/1804.03209>.  
 725  
 726  
 727 Wong, E. and Kolter, J. Z. Provably defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, pp. 5283–5292, 2018.  
 728  
 729  
 730 Wu, F., Li, L., Huang, Z., Vorobeychik, Y., Zhao, D., and Li, B. CROP: certifying robust policies for reinforcement learning through functional smoothing. *CoRR*, abs/2106.09292, 2021. URL <https://arxiv.org/abs/2106.09292>.  
 731  
 732  
 733 Xu, H., Gao, Y., Yu, F., and Darrell, T. End-to-end learning of driving models from large-scale video datasets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3530–3538. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.376. URL <https://doi.org/10.1109/CVPR.2017.376>.  
 734  
 735  
 736 Yang, J. B., Nguyen, M. N., San, P. P., Li, X. L., and Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15*, pp. 3995–4001. AAAI Press, 2015. ISBN 9781577357384.  
 737  
 738  
 739  
 740  
 741  
 742  
 743  
 744  
 745  
 746  
 747  
 748  
 749  
 750  
 751  
 752  
 753  
 754  
 755  
 756  
 757  
 758  
 759  
 760  
 761  
 762  
 763  
 764  
 765  
 766  
 767  
 768  
 769  
 770  
 771  
 772  
 773  
 774  
 775  
 776  
 777  
 778  
 779  
 780  
 781  
 782  
 783  
 784  
 785  
 786  
 787  
 788  
 789  
 790  
 791  
 792  
 793  
 794  
 795  
 796  
 797  
 798  
 799  
 800  
 801  
 802  
 803  
 804  
 805  
 806  
 807  
 808  
 809  
 810  
 811  
 812  
 813  
 814  
 815  
 816  
 817  
 818  
 819  
 820  
 821  
 822  
 823  
 824  
 825  
 826  
 827  
 828  
 829  
 830  
 831  
 832  
 833  
 834  
 835  
 836  
 837  
 838  
 839  
 840  
 841  
 842  
 843  
 844  
 845  
 846  
 847  
 848  
 849  
 850  
 851  
 852  
 853  
 854  
 855  
 856  
 857  
 858  
 859  
 860  
 861  
 862  
 863  
 864  
 865  
 866  
 867  
 868  
 869  
 870  
 871  
 872  
 873  
 874  
 875  
 876  
 877  
 878  
 879  
 880  
 881  
 882  
 883  
 884  
 885  
 886  
 887  
 888  
 889  
 890  
 891  
 892  
 893  
 894  
 895  
 896  
 897  
 898  
 899  
 900  
 901  
 902  
 903  
 904  
 905  
 906  
 907  
 908  
 909  
 910  
 911  
 912  
 913  
 914  
 915  
 916  
 917  
 918  
 919  
 920  
 921  
 922  
 923  
 924  
 925  
 926  
 927  
 928  
 929  
 930  
 931  
 932  
 933  
 934  
 935  
 936  
 937  
 938  
 939  
 940  
 941  
 942  
 943  
 944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 966  
 967  
 968  
 969  
 970  
 971  
 972  
 973  
 974  
 975  
 976  
 977  
 978  
 979  
 980  
 981  
 982  
 983  
 984  
 985  
 986  
 987  
 988  
 989  
 990  
 991  
 992  
 993  
 994  
 995  
 996  
 997  
 998  
 999  
 1000  
 1001  
 1002  
 1003  
 1004  
 1005  
 1006  
 1007  
 1008  
 1009  
 1010  
 1011  
 1012  
 1013  
 1014  
 1015  
 1016  
 1017  
 1018  
 1019  
 1020  
 1021  
 1022  
 1023  
 1024  
 1025  
 1026  
 1027  
 1028  
 1029  
 1030  
 1031  
 1032  
 1033  
 1034  
 1035  
 1036  
 1037  
 1038  
 1039  
 1040  
 1041  
 1042  
 1043  
 1044  
 1045  
 1046  
 1047  
 1048  
 1049  
 1050  
 1051  
 1052  
 1053  
 1054  
 1055  
 1056  
 1057  
 1058  
 1059  
 1060  
 1061  
 1062  
 1063  
 1064  
 1065  
 1066  
 1067  
 1068  
 1069  
 1070  
 1071  
 1072  
 1073  
 1074  
 1075  
 1076  
 1077  
 1078  
 1079  
 1080  
 1081  
 1082  
 1083  
 1084  
 1085  
 1086  
 1087  
 1088  
 1089  
 1090  
 1091  
 1092  
 1093  
 1094  
 1095  
 1096  
 1097  
 1098  
 1099  
 1100  
 1101  
 1102  
 1103  
 1104  
 1105  
 1106  
 1107  
 1108  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115  
 1116  
 1117  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133  
 1134  
 1135  
 1136  
 1137  
 1138  
 1139  
 1140  
 1141  
 1142  
 1143  
 1144  
 1145  
 1146  
 1147  
 1148  
 1149  
 1150  
 1151  
 1152  
 1153  
 1154  
 1155  
 1156  
 1157  
 1158  
 1159  
 1160  
 1161  
 1162  
 1163  
 1164  
 1165  
 1166  
 1167  
 1168  
 1169  
 1170  
 1171  
 1172  
 1173  
 1174  
 1175  
 1176  
 1177  
 1178  
 1179  
 1180  
 1181  
 1182  
 1183  
 1184  
 1185  
 1186  
 1187  
 1188  
 1189  
 1190  
 1191  
 1192  
 1193  
 1194  
 1195  
 1196  
 1197  
 1198  
 1199  
 1200  
 1201  
 1202  
 1203  
 1204  
 1205  
 1206  
 1207  
 1208  
 1209  
 1210  
 1211  
 1212  
 1213  
 1214  
 1215  
 1216  
 1217  
 1218  
 1219  
 1220  
 1221  
 1222  
 1223  
 1224  
 1225  
 1226  
 1227  
 1228  
 1229  
 1230  
 1231  
 1232  
 1233  
 1234  
 1235  
 1236  
 1237  
 1238  
 1239  
 1240  
 1241  
 1242  
 1243  
 1244  
 1245  
 1246  
 1247  
 1248  
 1249  
 1250  
 1251  
 1252  
 1253  
 1254  
 1255  
 1256  
 1257  
 1258  
 1259  
 1260  
 1261  
 1262  
 1263  
 1264  
 1265  
 1266  
 1267  
 1268  
 1269  
 1270  
 1271  
 1272  
 1273  
 1274  
 1275  
 1276  
 1277  
 1278  
 1279  
 1280  
 1281  
 1282  
 1283  
 1284  
 1285  
 1286  
 1287  
 1288  
 1289  
 1290  
 1291  
 1292  
 1293  
 1294  
 1295  
 1296  
 1297  
 1298  
 1299  
 1300  
 1301  
 1302  
 1303  
 1304  
 1305  
 1306  
 1307  
 1308  
 1309  
 1310  
 1311  
 1312  
 1313  
 1314  
 1315  
 1316  
 1317  
 1318  
 1319  
 1320  
 1321  
 1322  
 1323  
 1324  
 1325  
 1326  
 1327  
 1328  
 1329  
 1330  
 1331  
 1332  
 1333  
 1334  
 1335  
 1336  
 1337  
 1338  
 1339  
 1340  
 1341  
 1342  
 1343  
 1344  
 1345  
 1346  
 1347  
 1348  
 1349  
 1350  
 1351  
 1352  
 1353  
 1354  
 1355  
 1356  
 1357  
 1358  
 1359  
 1360  
 1361  
 1362  
 1363  
 1364  
 1365  
 1366  
 1367  
 1368  
 1369  
 1370  
 1371  
 1372  
 1373  
 1374  
 1375  
 1376  
 1377  
 1378  
 1379  
 1380  
 1381  
 1382  
 1383  
 1384  
 1385  
 1386  
 1387  
 1388  
 1389  
 1390  
 1391  
 1392  
 1393  
 1394  
 1395  
 1396  
 1397  
 1398  
 1399  
 1400  
 1401  
 1402  
 1403  
 1404  
 1405  
 1406  
 1407  
 1408  
 1409  
 1410  
 1411  
 1412  
 1413  
 1414  
 1415  
 1416  
 1417  
 1418  
 1419  
 1420  
 1421  
 1422  
 1423  
 1424  
 1425  
 1426  
 1427  
 1428  
 1429  
 1430  
 1431  
 1432  
 1433  
 1434  
 1435  
 1436  
 1437  
 1438  
 1439  
 1440  
 1441  
 1442  
 1443  
 1444  
 1445  
 1446  
 1447  
 1448  
 1449  
 1450  
 1451  
 1452  
 1453  
 1454  
 1455  
 1456  
 1457  
 1458  
 1459  
 1460  
 1461  
 1462  
 1463  
 1464  
 1465  
 1466  
 1467  
 1468  
 1469  
 1470  
 1471  
 1472  
 1473  
 1474  
 1475  
 1476  
 1477  
 1478  
 1479  
 1480  
 1481  
 1482  
 1483  
 1484  
 1485  
 1486  
 1487  
 1488  
 1489  
 1490  
 1491  
 1492  
 1493  
 1494  
 1495  
 1496  
 1497  
 1498  
 1499  
 1500  
 1501  
 1502  
 1503  
 1504  
 1505  
 1506  
 1507  
 1508  
 1509  
 1510  
 1511  
 1512  
 1513  
 1514  
 1515  
 1516  
 1517  
 1518  
 1519  
 1520  
 1521  
 1522  
 1523  
 1524  
 1525  
 1526  
 1527  
 1528  
 1529  
 1530  
 1531  
 1532  
 1533  
 1534  
 1535  
 1536  
 1537  
 1538  
 1539  
 1540  
 1541  
 1542  
 1543  
 1544  
 1545  
 1546  
 1547  
 1548  
 1549  
 1550  
 1551  
 1552  
 1553  
 1554  
 1555  
 1556  
 1557  
 1558  
 1559  
 1560  
 1561  
 1562  
 1563  
 1564  
 1565  
 1566  
 1567  
 1568  
 1569  
 1570  
 1571  
 1572  
 1573  
 1574  
 1575  
 1576  
 1577  
 1578  
 1579  
 1580  
 1581  
 1582  
 1583  
 1584  
 1585  
 1586  
 1587  
 1588  
 1589  
 1590  
 1591  
 1592  
 1593  
 1594  
 1595  
 1596  
 1597  
 1598  
 1599  
 1600  
 1601  
 1602  
 1603  
 1604  
 1605  
 1606  
 1607  
 1608  
 1609  
 1610  
 1611  
 1612  
 1613  
 1614  
 1615  
 1616  
 1617  
 1618  
 1619  
 1620  
 1621  
 1622  
 1623  
 1624  
 1625  
 1626  
 1627  
 1628  
 1629  
 1630  
 1631  
 1632  
 1633  
 1634  
 1635  
 1636  
 1637  
 1638  
 1639  
 1640  
 1641  
 1642  
 1643  
 1644  
 1645  
 1646  
 1647  
 1648  
 1649  
 1650  
 1651  
 1652  
 1653  
 1654  
 1655  
 1656  
 1657  
 1658  
 1659  
 1660  
 1661  
 1662  
 1663  
 1664  
 1665  
 1666  
 1667  
 1668  
 1669  
 1670  
 1671  
 1672  
 1673  
 1674  
 1675  
 1676  
 1677  
 1678  
 1679  
 1680  
 1681  
 1682  
 1683  
 1684  
 1685  
 1686  
 1687  
 1688  
 1689  
 1690  
 1691  
 1692  
 1693  
 1694  
 1695  
 1696  
 1697  
 1698  
 1699  
 1700  
 1701  
 1702  
 1703  
 1704  
 1705  
 1706  
 1707  
 1708  
 1709  
 1710  
 1711  
 1712  
 1713  
 1714  
 1715  
 1716  
 1717  
 1718  
 1719  
 1720  
 1721  
 1722  
 1723  
 1724  
 1725  
 1726  
 1727  
 1728  
 1729  
 1730  
 1731  
 1732  
 1733  
 1734  
 1735  
 1736  
 1737  
 1738  
 1739  
 1740  
 1741  
 1742  
 1743  
 1744  
 1745  
 1746  
 1747  
 1748  
 1749  
 1750  
 1751  
 1752  
 1753  
 1754  
 1755  
 1756  
 1757  
 1758  
 1759  
 1760  
 1761  
 1762  
 1763  
 1764  
 1765  
 1766  
 1767  
 1768  
 1769  
 1770  
 1771  
 1772  
 1773  
 1774  
 1775  
 1776  
 1777  
 1778  
 1779  
 1780  
 1781  
 1782  
 1783  
 1784  
 1785  
 1786  
 1787  
 1788  
 1789  
 1790  
 1791  
 1792  
 1793  
 1794  
 1795  
 1796  
 1797  
 1798  
 1799  
 1800  
 1801  
 1802  
 1803  
 1804  
 1805  
 1806  
 1807  
 1808  
 1809  
 1810  
 1811  
 1812  
 1813  
 1814  
 1815  
 1816  
 1817  
 1818  
 1819  
 1820  
 1821  
 1822  
 1823  
 1824  
 1825  
 1826  
 1827  
 1828  
 1829  
 1830  
 1831  
 1832  
 1833  
 1834  
 1835  
 1836  
 1837  
 1838  
 1839  
 1840  
 1841  
 1842  
 1843  
 1844  
 1845  
 1846  
 1847  
 1848  
 1849  
 1850  
 1851  
 1852  
 1853  
 1854  
 1855  
 1856  
 1857  
 1858  
 1859  
 1860  
 1861  
 1862  
 1863  
 1864  
 1865  
 1866  
 1867  
 1868  
 1869  
 1870  
 1871  
 1872  
 1873  
 1874  
 1875  
 1876  
 1877  
 1878  
 1879  
 1880  
 1881  
 1882  
 1883  
 1884  
 1885  
 1886  
 1887  
 1888

715 learned optimal adversary. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=sCZbhBvqQaU>.

716  
717  
718  
719 Zhang, L., Aggarwal, C., and Qi, G.-J. Stock price  
720 prediction via discovering multi-frequency trading pat-  
721 terns. In *Proceedings of the 23rd ACM SIGKDD Inter-  
722 national Conference on Knowledge Discovery and  
723 Data Mining*, KDD '17, pp. 2141–2149, New York,  
724 NY, USA, 2017. Association for Computing Machin-  
725 ery. ISBN 9781450348874. doi: 10.1145/3097983.  
726 3098117. URL <https://doi.org/10.1145/3097983.3098117>.

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

## A. Proof of Lemma 5.1

**Statement.** The change in each  $\tilde{f}_j$  under an adversary in  $\mathcal{A}_\epsilon$  is bounded as

$$|\tilde{f}_j(W_j) - \tilde{f}_j(W'_j)| \leq \sum_{i=j-w+1}^j \psi(d(x_i, x_i^{j+1-i})).$$

*Proof.* The left-hand side of the above inequality can be re-written as:

$$\begin{aligned} |\tilde{f}_j(W_j) - \tilde{f}_j(W'_j)| &= |\tilde{f}_j(x_{j-w+1}, \dots, x_j) - \tilde{f}_j(x_{j-w+1}^w, \dots, x_j^1)| \\ &= |\tilde{f}_j(x_{j-w+1}, \dots, x_{j-1}, x_j) - \tilde{f}_j(x_{j-w+1}, \dots, x_{j-1}, x_j^1) \\ &\quad + \tilde{f}_j(x_{j-w+1}, \dots, x_{j-1}, x_j^1) - \tilde{f}_j(x_{j-w+1}^w, \dots, x_{j-1}^2, x_j^1)| \\ &= \left| \sum_{k=1}^w \tilde{f}_j(x_{j-w+1}, \dots, x_{j-k+1}, x_{j-k+2}^{k-1}, \dots, x_j^1) - \tilde{f}_j(x_{j-w+1}, \dots, x_{j-k+1}^k, x_{j-k+2}^{k-1}, \dots, x_j^1) \right| \\ &\leq \sum_{k=1}^w \left| \tilde{f}_j(x_{j-w+1}, \dots, x_{j-k+1}, x_{j-k+2}^{k-1}, \dots, x_j^1) - \tilde{f}_j(x_{j-w+1}, \dots, x_{j-k+1}^k, x_{j-k+2}^{k-1}, \dots, x_j^1) \right| \end{aligned}$$

The two terms in each summand differ only in the  $(j-k+1)$ -th input. Thus, it can be written as the difference of the expected value of some  $[0, 1]$ -function  $q$  under the distributions  $\mathcal{S}(x_{j-k+1})$  and  $\mathcal{S}(x_{j-k+1}^k)$ , i.e.,  $|\mathbb{E}_{\tilde{x}_{j-k+1} \sim \mathcal{S}(x_{j-k+1})}[q(\tilde{x}_{j-k+1})] - \mathbb{E}_{\tilde{x}_{j-k+1}^k \sim \mathcal{S}(x_{j-k+1}^k)}[q(\tilde{x}_{j-k+1}^k)]|$  which can be upper bounded by the total variation between  $\mathcal{S}(x_{j-k+1})$  and  $\mathcal{S}(x_{j-k+1}^k)$ . Therefore, from condition (4), we have:

$$\begin{aligned} &|\tilde{f}_j(x_{j-w+1}, \dots, x_{j-k+1}, x_{j-k+2}^{k-1}, \dots, x_j^1) - \tilde{f}_j(x_{j-w+1}, \dots, x_{j-k+1}^k, x_{j-k+2}^{k-1}, \dots, x_j^1)| \\ &\leq \text{TV}(\mathcal{S}(x_{j-k+1}), \mathcal{S}(x_{j-k+1}^k)) \leq \psi(d(x_{j-k+1}, x_{j-k+1}^k)). \end{aligned}$$

This proves the statement of the lemma.  $\square$

## B. Proof of Theorem 5.2

**Statement.** Let  $\tilde{Z}_\epsilon$  to be the minimum  $\tilde{Z}$  for an adversary in  $\mathcal{A}_\epsilon$ . Then,

$$|\tilde{Z} - \tilde{Z}_\epsilon| \leq w\psi(\epsilon).$$

*Proof.* Let  $\tilde{Z}'$  be the overall performance of  $\tilde{M}$  under an adversary. Then,

$$\begin{aligned} |\tilde{Z} - \tilde{Z}'| &= \left| \frac{\sum_{j=1}^t \tilde{f}_j(W_j)}{t} - \frac{\sum_{j=1}^t \tilde{f}_j(W'_j)}{t} \right| \\ &\leq \frac{\sum_{j=1}^t |\tilde{f}_j(W_j) - \tilde{f}_j(W'_j)|}{t} \\ &\leq \sum_{j=1}^t \sum_{k=1}^w \psi(d(x_{j-k+1}, x_{j-k+1}^k))/t && \text{(from lemma 5.1)} \\ &\leq \sum_{j=1}^t \sum_{k=1}^w \psi(d(x_j, x_j^k))/t \\ &= w \sum_{j=1}^t \sum_{k=1}^w \psi(d(x_j, x_j^k))/wt \\ &\leq w\psi \left( \sum_{j=1}^t \sum_{k=1}^w d(x_j, x_j^k)/wt \right) && (\psi \text{ is concave and Jensen's inequality}) \end{aligned}$$

825 Therefore, for the worst-case adversary in  $\mathcal{A}_\epsilon$ , we have  
 826  
 827

$$|\tilde{Z} - \tilde{Z}_\epsilon| \leq w\psi(\epsilon)$$

828 from constraint (2) on the average distance between the original and perturbed inputs.  $\square$   
 829

### 830 C. Function $\psi$ for Different Distributions

831 For an isometric Gaussian distribution,  
 832

$$834 \text{TV}(\mathcal{N}(x_i, \sigma^2 I), \mathcal{N}(x'_i, \sigma^2 I)) = \text{erf}(\|x_i - x'_i\|_2 / 2\sqrt{2}\sigma). \\ 835$$

836 *Proof.* Due to the isometric symmetry of the Gaussian distribution and the  $\ell_2$ -norm, the total variation between the two  
 837 distributions is the same as when they are separated by the same  $\ell_2$ -distance but only in the first coordinate. It is equivalent  
 838 to shifting a univariate normal distribution by the same amount. Therefore, the total variation between the two distributions  
 839 is equal to the difference in the probability of a normal random variable with variance  $\sigma^2$  being less than  $\|x_i - x'_i\|_2/2$  and  
 840  $-\|x_i - x'_i\|_2/2$ , i.e.,  $\Phi(\|x_i - x'_i\|_2/2\sigma) - \Phi(-\|x_i - x'_i\|_2/2\sigma)$  where  $\Phi$  is the standard normal CDF.  
 841

$$\begin{aligned} 842 \text{TV}(\mathcal{N}(x_i, \sigma^2 I), \mathcal{N}(x'_i, \sigma^2 I)) &= \Phi(\|x_i - x'_i\|_2/2\sigma) - \Phi(-\|x_i - x'_i\|_2/2\sigma) \\ 843 &= 2\Phi(\|x_i - x'_i\|_2/2\sigma) - 1 \\ 844 &= 2\left(\frac{1 + \text{erf}(\|x_i - x'_i\|_2/2\sqrt{2}\sigma)}{2}\right) - 1 \\ 845 &= \text{erf}(\|x_i - x'_i\|_2/2\sqrt{2}\sigma). \\ 846 \\ 847 \\ 848 \end{aligned}$$

$\square$

851 For a uniform smoothing distribution  $\mathcal{U}(x_i, b)$  between  $x_{ij} - b/2$  and  $x_{ij} + b/2$  in each dimension  $j$  of  $x_i$  for some  $b \geq 0$ ,  
 852  $\text{TV}(\mathcal{U}(x_i, b), \mathcal{U}(x'_i, b)) \leq \|x_i - x'_i\|_1/b$ . When  $\|x_i - x'_i\|_1$  is constrained, the overlap between  $\mathcal{U}(x_i, b)$  and  $\mathcal{U}(x'_i, b)$  is  
 853 minimized when the shift is only along one dimension.  
 854

### 855 D. Comparison with Existing Certificates for Static Tasks

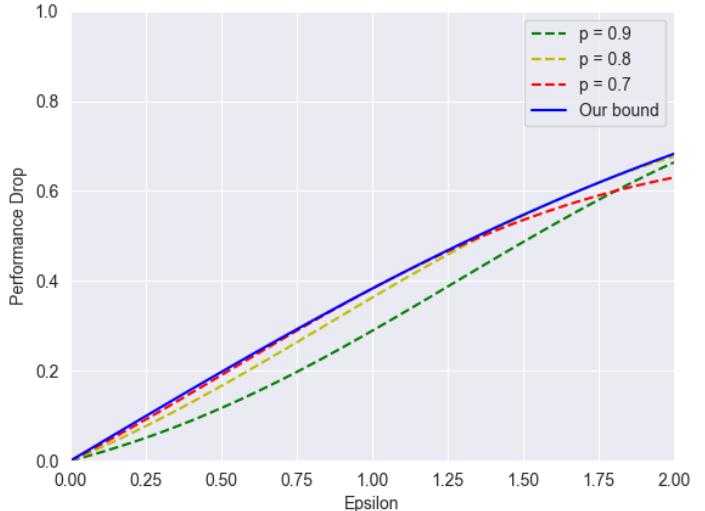
856 In this section, we compare our bound when applied to  
 857 the static setting of classification, i.e., window size  $w = 1$   
 858 in bound (1), to that obtained by Cohen et al. (2019) for  
 859 an  $\ell_2$  adversary and a Gaussian smoothing distribution.  
 860 As discussed in Appendix C, the  $\psi$  function for this case  
 861 takes the form of the Gauss error function erf. Thus our  
 862 bound on the drop in the smoothed model's performance  
 863 against an  $\ell_2$  adversary is given by:  
 864

$$865 |\tilde{Z} - \tilde{Z}_\epsilon| \leq \text{erf}(\epsilon/2\sqrt{2}\sigma). \\ 866$$

867 Cohen et al. (2019)'s certificate bounds the worst-case  
 868 adversarial performance as a function of the clean performance.  
 869 If the probability of predicting the correct class  
 870 is  $p$  on the original input, the probability of that in the  
 871 presence of an adversary is bounded by  $\Phi(\Phi^{-1}(p) - \epsilon/\sigma)$ .  
 872 Therefore, the performance drop  $\Delta p$  is bounded by:  
 873

$$874 \Delta p \leq p - \Phi\left(\Phi^{-1}(p) - \frac{\epsilon}{\sigma}\right). \quad (5) \\ 875$$

876 Figure 6 compares the two bounds for different values of  
 877  $p$ . We keep  $\sigma = 1$  as it only has a scaling effect along the  
 878 879



875 Figure 6. Comparison between our bound and Cohen et al. (2019)'s  
 876 certificate for an  $\ell_2$  adversary and a Gaussian smoothing distribution.  
 877 The solid blue curve corresponds to our bound and the dashed curves  
 878 represent bound (5) for different values of  $p$ . We keep  $\sigma = 1$  as it  
 879 only has a scaling effect along the  $x$ -axis.

880  $x$ -axis. The bound from the  $\ell_2$  certificate by Cohen et al. (2019) is tighter than ours, mainly because it takes the clean  
881 performance  $p$  of the smoothed model into account. However, the gap between the two bounds is small in the range where  $\epsilon$   
882 goes from 0 to 2, by which point the certified performance drops by more than 60%. Thus for most meaningful robustness  
883 guarantees, our certificates are almost at par with the best-known  $\ell_2$  certificates. The key advantage of our certificates over  
884 those for the static setting is that they are applicable for an adaptive adversary that can allocate different attack budgets for  
885 different input items in the stream.  
886

## 887 E. Experimental details

888 We use a single NVIDIA RTX A4000 GPU with four AMD EPYC 7302P Processors. For our main experiments with UCI  
889 HAR and Speech Commands datasets, we use window size  $w = 2$  with inputs belonging to  $\mathbb{R}^{250 \times 6}$  and  $\mathbb{R}^{4000}$ . The UCI  
890 HAR dataset consists of long streaming inputs with sample-level annotations. For a window  $W_j$ , the label is the majority  
891 class that is present in that window. The signals in the HAR dataset are standardized to have mean 0 and variance 1. For  
892 the speech keyword detection task, we use a subset of the Speech commands dataset that consists of long noise clips and  
893 one-second-long speech keyword clips. The labels for each audio clip are available. We utilize all the long noise clips  
894 and clips belonging to the classes belonging speech utterances of numbers from zero to nine to make longer clips for our  
895 streaming case. We add noise clips to the keyword audios to make them more similar to real-world scenarios. Each clip  
896 is stitched together (Li et al., 2018) with arbitrarily long noise between each keyword clip. To make transitions between  
897 the audio smooth, we use exponential decays to overlap keyword audio clips for stitching, with noise in the background.  
898 Hence, for the speech keyword detection, we have 11 classes for labels – zero to nine and a noise class. A window is labeled  
899 to be the majority class in that window.  
900

901 For training, we use M5 networks with 32 channels for HAR. We train for 30 epochs with a batch-size of 256 using SGD  
902 with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 0.0001. We use a cosine annealing learning rate  
903 scheduler. For training the robust models, we use different smoothing noises with standard deviations 4, 6, 8, and 10. For  
904 training on the keyword detection data, we use M5 networks with 128 channels for HAR. We train for 30 epochs with a  
905 batch-size of 128 using SGD with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 0.0001. We use a  
906 cosine annealing learning rate scheduler. For training the robust models, we use different smoothing noises with standard  
907 deviations 0.1, 0.2, 0.4, 0.6, and 0.8. For attacking the trained models, we use PGD  $\ell_2$  attacks for both the datasets. PGD is  
908 run for 100 steps with a step size of  $2\epsilon'/100$  where  $\epsilon'$  is the  $\ell_2$  attack budget.  
909

## 910 F. Attacking the Smooth Models

911 In this section, we empirically demonstrate that the performance of our smooth models in the presence of an adversary is  
912 lower bound by the certificates we obtain. For these experiments, we consider an adversary that can only attack an input  
913 once as in Section 6.1. We show our results on the Human Activity Recognition dataset in Figure 7 for a window size of  
914  $w = 2$ . As seen in the plots, the empirical performance of the smooth models, after the online adversarial attack, is always  
915 better than our theoretical certificates. Hence, this experiment validates our certificates.  
916

917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934

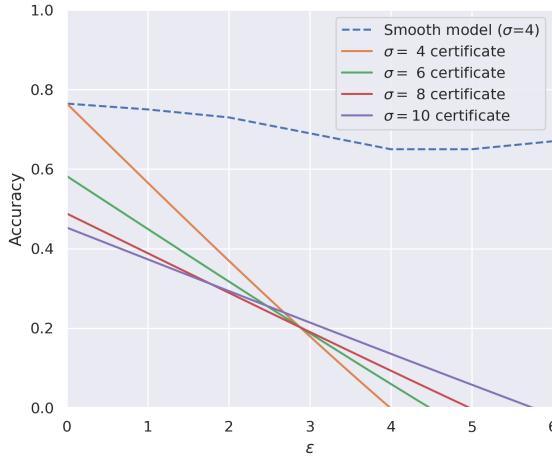
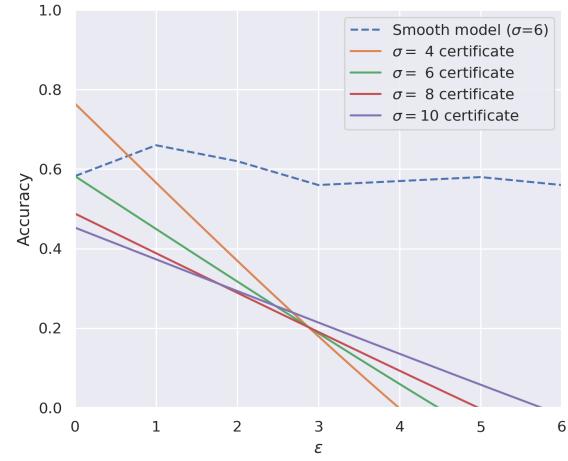
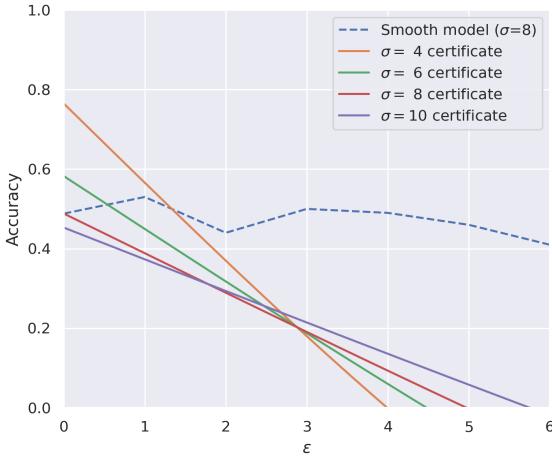
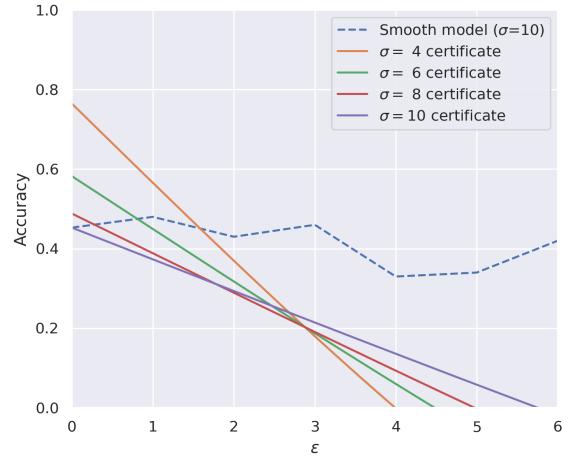

 (a) Attacking model with smoothing noise  $\sigma = 4$ 

 (b) Attacking model with smoothing noise  $\sigma = 6$ 

 (c) Attacking model with smoothing noise  $\sigma = 8$ 

 (d) Attacking model with smoothing noise  $\sigma = 10$ 

Figure 7. Certificates against online adversarial attacks for varying smoothing noises. We attack smooth models trained with different smoothing noises in these plots. Here we can perturb each input only once. The average size of perturbation is computed as per equation 2.