Certifiable Robustness Against Patch Attacks Using an ERM Oracle

Abstract

Consider patch attacks, where at test-time an adversary manipulates a test image with a patch in order to induce a targeted mis-classification. We consider a recent defense to patch attacks, Patch-Cleanser (Xiang et al., 2022). The Patch-Cleanser algorithm requires a prediction model to have a "two-mask correctness" property, meaning that the prediction model should correctly classify any image when any two blank masks replace portions of the image. To this end, Xiang et al. (2022) learn a prediction model to be robust to two-mask operations by augmenting the training set by adding pairs of masks at random locations of training images, and performing *empirical risk minimization* (ERM) on the augmented dataset. However, in the non-realizable setting when no predictor is perfectly correct on *all* two-mask operations on all images, we exhibit an example where ERM fails. To overcome this challenge, we propose a different algorithm that provably learns a predictor robust to *all* two-mask operations using an ERM oracle, based on prior work by Feige et al. (2015a) .

1. Introduction

Patch attacks (Brown et al., 2017; Karmon et al., 2018; Yang et al., 2020) are an important threat model in the general field of test time evasion attacks (Goodfellow et al., 2014). In a patch attack, the adversary replaces a contiguous block of pixels with an adversarially crafted pattern. Patch attacks can realize physical world attacks to computer vision systems by printing and attaching a patch into an object. To secure performance of computer vision systems against patch-attacks, there has been an active line of research for providing certifiable robustness against them (see e.g., McCoyd et al., 2020; Xiang et al., 2020; Xiang and Mittal, 2021; Metzen and Yatsura, 2021; Zhang et al., 2020; Chiang et al., 2020).

Xiang et al. (2022) recently proposed a state-of-the-art algorithm called Patch-Cleanser that can provably defend against patch attacks. The high level idea of the Patch-Cleanser algorithm is to robustly remove all adversarial pixels of an input image in order to obtain accurate predictions. The main difficulty is that the patch location is unknown. One naive solution is to place a mask at all possible locations of an input image. As long as the masks are large enough, at least one of the masks would cover the patch and remove the adversarial effects of the patch so that the prediction model can induce a correct classification on the input image. However, it is challenging to distinguish between this correct prediction and the predictions on the other masked images. To overcome this challenge, they use a second mask. For each of the one-masked images produced in the first step, they add a second mask at all possible locations. For each one-masked image, if for all possible locations of the second mask, the prediction model outputs the same classification, it means that the first mask was removing the patch, and the agreed-upon prediction is correct. Also, any disagreements implies the contrary.

Crucially, the Patch-Cleanser algorithm relies on a *two-mask correctness* assumption of the prediction model that is defined as follows: for a given input (x, y), if for any pair of masks applied to x, a prediction model F outputs the correct prediction y, then F has two-correctness property on (x, y) (see Xiang et al. (2022, Definition 2)). They show as long as the two-mask correctness property holds, their double-masking algorithm guarantees robustness against patch attacks on the input image (x, y).

In order to train a model with the two-mask correctness property, Xiang et al. (2022) use a heuristic data-augmentation approach as follows. They add pairs of masks at random locations to training images and learn a model that predicts correctly on the masked-images using *empirical risk minimization*.

However, we argue that in the non-realizable setting, when no predictor achieves zero robust loss, *this approach can fail*. Intuitively, an ERM oracle does not distinguish between distributing error over a few perturbations (i.e. masked-out variations) of many input images versus concentrating many mistakes on the perturbations of few input images. However, in the latter case, the robust loss can be much higher than the former case. To wit, to obtain high robust loss the adversary only needs one successful perturbation per clean image. If some input images have many perturbations that fool the classifier, but most input images have none, then the adversary cannot obtain high robust loss. We want to be in the second case. We have included a schematic demonstrating this failure mode in A.4.

Our main contribution is an algorithm to learn a predictor that is robust to a set of masking operations (resulting from the two-mask), using an ERM oracle. The algorithm is based on prior work due to Feige et al. (2015b), but the analysis and application are novel in this work. Combining our algorithm with Patch-Cleanser yields a predictor that is provably robust to adversarial patch attacks.

Setup and Notation Let $\mathcal{X}$ denote the instance space and $\mathcal{Y}$ denote the label space. Our main objective is to be robust against adversarial patches $\mathcal{A} : \mathcal{X} \to 2^{\mathcal{X}}$, where $\mathcal{A}(x)$ represents the (potentially infinite) set of adversarially patched images that an adversary might attack with at test-time. Xiang et al. (2022) showed that even though the space of adversarial patches $\mathcal{A}$ can be exponential or infinite, one can consider a "covering" set $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ of masking operations on images where $|\mathcal{U}(x)|$ is polynomially finite.

Thus for the remainder of the paper, we focus on the task of learning a predictor robust to a perturbation set $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, where $\mathcal{U}(x) \subseteq \mathcal{X}$ is the set of allowed masking operations that can be performed on x. We assume that $\mathcal{U}(x)$ is finite where $|\mathcal{U}(x)| \leq k$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and denote by $\mathrm{vc}(\mathcal{H})$ its VC dimension. Let $\mathsf{ERM}_{\mathcal{H}}$ be an ERM oracle for $\mathcal{H}$. For any set arbitrary set W, denote by $\Delta(W)$ the set of distributions over W. $\mathsf{OPT}_{\mathcal{H}}$ is defined as follows:

$$\mathsf{OPT}_{\mathcal{H}} \triangleq \min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \max_{z \in \mathcal{U}(x)} \mathbb{1}\left[h(z) \neq y\right]. \tag{1}$$

2. Main Result: Minimizing Robust Loss Using an ERM Oracle

In this section, we present our main contribution: an algorithm to learn a predictor that is simultaneously robust to a set of (polynomially many) masking operations, using an $\mathsf{ERM}_{\mathcal{H}}$ oracle. The algorithm is based on prior work due to Feige et al. (2015b), but the analysis and application are novel in this work. The main interesting feature of this algorithm is that it achieves stronger robustness guarantees in the non-realizable regime when $\mathsf{OPT}_{\mathcal{H}} \gg 0$, where the approach of Xiang et al. (2022) — as we highlighted in the introduction — of calling $\mathsf{ERM}_{\mathcal{H}}$ on the inflated dataset: original training points plus all possible perturbations resulting from the allowed masking operations, can provably fail (see e.g., A.4).

Algorithm 1: Feige, Mansour, and Schapire (2015b)

Input: weight update parameter $\eta > 0$, number of rounds T, and training dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Set $w_1(z, (x, y)) = 1$, for each $(x, y) \in S, z \in \mathcal{U}(x)$.

Set $P^1(z, (x, y)) = \frac{w_1(z,(x,y))}{\sum_{z' \in \mathcal{U}(x)} w_1(z',(x,y))}$, for each $(x, y) \in S, z \in \mathcal{U}(x)$.

for *each* $t \leftarrow 1$ **to** T **do**

  Call ERM on the empirical weighted distribution:
  $h_t = \mathrm{argmin}_{h\in\mathcal{H}} \sum_{(x,y)\in S} \sum_{z\in\mathcal{U}(x)} \frac{1}{m} P^t(z, (x, y))\mathbb{1}\left[h_t(z) \neq y\right]$.

  for *each* $(x, y) \in S$ *and* $z \in \mathcal{U}(x)$ **do**

    $w_{t+1}(z, (x, y)) = (1 + \eta\mathbb{1}\left[h_t(z) \neq y\right]) \cdot w_t(z, (x, y))$.

    $P^{t+1}(z, (x, y)) = \frac{w_t(z,(x,y))}{\sum_{z' \in \mathcal{U}(x)} w_t(z',(x,y))}$.

Output: The majority-vote predictor $\mathrm{MAJ}(h_1, \ldots, h_T)$.

Theorem 1. *Set* $T(\epsilon) = \frac{32 \ln k}{\epsilon^2}$ *and* $m(\epsilon, \delta) = O\left(\frac{\mathrm{vc}(\mathcal{H})(\ln k)^2}{\epsilon^4} \ln\left(\frac{\ln k}{\epsilon^2}\right) + \frac{\ln(1/\delta)}{\epsilon^2}\right)$*. Then, for any distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$*, with probability at least* $1 - \delta$ *over* $S \sim \mathcal{D}^{m(\epsilon,\delta)}$*, running Algorithm 1 for* $T(\epsilon)$ *rounds produces* $h_1, \ldots, h_{T(\epsilon)}$ *satisfying:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{z\in\mathcal{U}(x)} \mathbb{1}\left[\mathrm{MAJ}(h_1, \ldots, h_{T(\epsilon)})(z) \neq y\right]\right] \leq 2\mathsf{OPT}_{\mathcal{H}} + \epsilon.$$

Comparison with prior related work As presented, Feige et al. (2015b) only considered *finite* hypothesis classes $\mathcal{H}$ and provided generalization guarantees depending on $\log|\mathcal{H}|$. On the other hand, we consider here infinite classes $\mathcal{H}$ with bounded VC dimension and provide tighter robust generalization bounds. The robust learning guarantee (Attias et al., 2022, Theorem 2) assumes access to a *robust* ERM oracle, which minimizes the robust loss on the training dataset. On the other hand, at the expense of higher sample complexity, we provide a robust learning guarantee using only an ERM oracle in the challenging *non-realizable* setting. Prior work due to Montasser et al. (2020) considered using an ERM oracle for robust learning but only in the simpler realizable setting (when $\mathsf{OPT}_{\mathcal{H}} = 0$).

Before proceeding with the proof Theorem 1, we describe now at a high-level the proof strategy. The main insight is to solve a finite zero-sum game. In particular, our goal is to find a mixed-strategy over the hypothesis class that is approximately close to the value of the game:

$$\mathsf{OPT}_{S,\mathcal{H}} \triangleq \min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \max_{z_i\in\mathcal{U}(x_i)} \mathbb{1}\left[h(z_i)\neq y_i\right].$$

We observe that Algorithm 1 due to (Feige et al., 2015b) solves a similar finite zero-sum game (see Lemma 3), and then we relate it to the value of the game we are interested in (see Lemma 2). Combined together, this only establishes that we can minimize the robust loss on the empirical dataset using an ERM oracle. We then appeal to uniform convergence guarantees for the robust loss in Lemma 4 to show that, with large enough training data, our output predictor achieves robust risk that is close to the value of the game.

Lemma 2. *For any data set $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \in (\mathcal{X}\times\mathcal{Y})^m$,*

$$\mathsf{OPT}_{S,\mathcal{H}} = \min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \max_{z_i\in\mathcal{U}(x_i)} \mathbb{1}\left[h(z_i)\neq y_i\right] \geq \min_{Q\in\Delta(\mathcal{H})} \max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i\sim P_i}\mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}\left[h(z_i)\neq y_i\right].$$

Lemma 3 (Feige, Mansour, and Schapire (2015b))**.** *For any data set $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \in (\mathcal{X}\times\mathcal{Y})^m$, running Algorithm 1 for T rounds produces a mixed-strategy $\hat{Q} = \frac{1}{T}\sum_{t=1}^{T} h_t \in \Delta(\mathcal{H})$ satisfying:*

$$\max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i\sim P_i} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[h_t(z_i)\neq y_i\right] \leq \min_{Q\in\Delta(\mathcal{H})} \max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i\sim P_i}\mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}\left[h(z_i)\neq y_i\right] + 2\sqrt{\frac{\ln k}{T}}.$$

Lemma 4 (VC Dimension for the Robust Loss (Attias et al., 2022))**.** *For any class $\mathcal{H}$ and any $\mathcal{U}$ such that $\sup_{x\in\mathcal{X}}|\mathcal{U}(x)|\leq k$, denote the robust loss class of $\mathcal{H}$ with respect to $\mathcal{U}$ by*

$$\mathcal{L}^{\mathcal{U}}_{\mathcal{H}} = \left\{(x,y)\mapsto \max_{z\in\mathcal{U}(x)} \mathbb{1}\left[h(z)\neq y\right] : h\in\mathcal{H}\right\}.$$

Then, it holds that $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \leq O(\mathrm{vc}(\mathcal{H})\log(k))$.

We are now ready to proceed with the proof of Theorem 1.

Proof Let $S\sim\mathcal{D}^m$ be an iid sample from $\mathcal{D}$, where the size of the sample m will be determined later. By invoking Lemma 3 and Lemma 2, we observe that running Algorithm 1 on S for T rounds, produces $h_1, \dots, h_T$ satisfying

$$\max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i\sim P_i} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[h_t(z_i)\neq y_i\right] \leq \mathsf{OPT}_{S,\mathcal{H}} + \frac{\epsilon}{4}.$$

Next, the average robust loss for the majority-vote predictor $\mathrm{MAJ}(h_1, \ldots, h_T)$ can be bounded from above as follows

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} \max_{z_i \in \mathcal{U}(x_i)} \mathbb{1}\left[\mathrm{MAJ}(h_1, \ldots, h_T)(z_i) \neq y_i\right] \\
&\leq \frac{1}{m}\sum_{i=1}^{m} \max_{z_i \in \mathcal{U}(x_i)} 2 \mathop{\mathbb{E}}_{t\sim[T]} \mathbb{1}\left[h_t(z_i) \neq y_i\right] \\
&= 2\frac{1}{m}\sum_{i=1}^{m} \max_{z_i \in \mathcal{U}(x_i)} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[h_t(z_i) \neq y_i\right] \\
&\leq 2 \max_{P_1 \in \Delta(\mathcal{U}(x_1)), \ldots, P_m \in \Delta(\mathcal{U}(x_m))} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i \sim P_i} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[h_t(z_i) \neq y_i\right] \\
&\leq 2\mathsf{OPT}_{S,\mathcal{H}} + \frac{\epsilon}{2}.
\end{aligned}$$

Next, we invoke Lemma 4 to obtain a uniform convergence guarantee on the robust loss. In particular, we apply Lemma 4 on the *convex-hull* of $\mathcal{H}$: $\mathcal{H}^T = \{\mathrm{MAJ}(h_1, \ldots, h_T) : h_1, \ldots, h_T \in \mathcal{H}\}$. By a classic result due to Blumer, Ehrenfeucht, Haussler, and Warmuth (1989), it holds that $\mathrm{vc}(\mathcal{H}^T) = O(\mathrm{vc}(\mathcal{H}) T \ln T)$. Combining this with Lemma 4 and plugging-in the value of $T = \frac{32 \ln k}{\epsilon^2}$, we get that the VC dimension of the robust loss class of $\mathcal{H}^T$ is bounded from above by

$$\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}^T}) \leq O\left(\frac{\mathrm{vc}(\mathcal{H})(\ln k)^2}{\epsilon^2} \ln\left(\frac{\ln k}{\epsilon^2}\right)\right).$$

Finally, using Vapnik's "General Learning" uniform convergence (Vapnik, 1982), with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ where $m = O\left(\frac{\mathrm{vc}(\mathcal{H})(\ln k)^2}{\epsilon^4} \ln\left(\frac{\ln k}{\epsilon^2}\right) + \frac{\ln(1/\delta)}{\epsilon^2}\right)$, it holds that

$$\forall f \in \mathcal{H}^T : \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\max_{z \in \mathcal{U}(x)} \mathbb{1}\left[f(z) \neq y\right]\right] \leq \frac{1}{m}\sum_{i=1}^{m} \max_{z_i \in \mathcal{U}(x_i)} \mathbb{1}\left[f(z_i) \neq y_i\right] + \frac{\epsilon}{4}.$$

This also applies to the particular output $\mathrm{MAJ}(h_1, \ldots, h_T)$ of Algorithm 1, and thus

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\max_{z \in \mathcal{U}(x)} \mathbb{1}\left[\mathrm{MAJ}(h_1, \ldots, h_{T(\epsilon)})(z) \neq y\right]\right] &\leq \frac{1}{m}\sum_{i=1}^{m} \max_{z_i \in \mathcal{U}(x_i)} \mathbb{1}\left[\mathrm{MAJ}(h_1, \ldots, h_T)(z_i) \neq y_i\right] + \frac{\epsilon}{4} \\
&\leq 2\mathsf{OPT}_{S,\mathcal{H}} + \frac{\epsilon}{2} + \frac{\epsilon}{4}.
\end{aligned}$$

Finally, by applying a standard Chernoff-Hoeffding concentration inequality, we get that $\mathsf{OPT}_{S,\mathcal{H}} \leq \mathsf{OPT}_{\mathcal{H}} + \frac{\epsilon}{8}$. Combining this with the above inequality concludes the proof. ■

3. Conclusion

Per the call for papers, we discuss the scalability of our method, which depends on multiple factors. As is argued/demonstrated empirically in the original Patch-Cleanser paper, their original defense scales to high resolution images. An additional strength of this research direction initiated by Patch-Cleanser and maintained by our approach is that it is *agnostic* to the network/structure of the model, and can be applied as a module on top of any state-of-the-art model. The complexity of Algorithm 1 has two components, the complexity of the ERM Oracle and the number of iterations T. As noted in Section 2, our algorithm makes $T = \Omega(\frac{\ln k}{\epsilon^2})$ oracle calls where $\ln k$ is the bit-complexity of the perturbations and thus we are oracle-efficient.

In order to modify Xiang et al. (2022) to handle the case where two-mask correctness *is not realizable*, we exhibit polynomial time algorithms for learning a classifier that satisfies the two-mask property and analyze the provable robustness of this approach, based upon prior work by Feige et al. (2015a). The key future work that we intend for the full version of this work includes an empirical evaluation of this method and extensions to a new multi-group robustness notion.

Appendix A. Missing Proofs

A.1 Proof of Lemma 2

Proof By definition of $\mathsf{OPT}_{S,\mathcal{H}}$, it follows that

$$\begin{aligned}
\mathsf{OPT}_{S,\mathcal{H}} &= \min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \max_{z_i\in\mathcal{U}(x_i)} \mathbb{1}\left[h(z_i)\neq y_i\right] \\
&\geq \min_{h\in\mathcal{H}} \max_{z_1\in\mathcal{U}(x_1),\dots,z_m\in\mathcal{U}(x_m)} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[h(z_i)\neq y_i\right] \\
&\geq \min_{Q\in\Delta(H)} \max_{z_1\in\mathcal{U}(x_1),\dots,z_m\in\mathcal{U}(x_m)} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}\left[h(z_i)\neq y_i\right] \\
&\geq \min_{Q\in\Delta(\mathcal{H})} \max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i\sim P_i} \mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}\left[h(z_i)\neq y_i\right].
\end{aligned}$$

■

A.2 Proof of Lemma 3

Proof By the minimax theorem and (Feige, Mansour, and Schapire, 2015b, Equation 3 and 9 in proof of Theorem 1), we have that

$$\max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{z_i\sim P_i} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[h_t(z_i)\neq y_i\right] \leq$$

$$\min_{Q\in\Delta(\mathcal{H})} \max_{P_1\in\Delta(\mathcal{U}(x_1)),\dots,P_m\in\Delta(\mathcal{U}(x_m))} \mathop{\mathbb{E}}_{z_i\sim P_i} \mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}\left[h(z_i)\neq y_i\right] + 2\frac{\sqrt{\mathcal{L}^* m \ln k}}{T},$$

where $\mathcal{L}^* = \sum_{i=1}^{m} \max_{z\in\mathcal{U}(x_i)} \sum_{t=1}^{T} \mathbb{1}\left[h_t(z)\neq y\right]$. By observing that $\mathcal{L}^* \leq mT$ and dividing both sides of the inequality above by m, we arrive at the inequality stated in the lemma. ■

A.3 Proof of Lemma 4

Proof By finiteness of $\mathcal{U}$, observe that for any dataset $S \in (\mathcal{X}\times\mathcal{Y})^m$, each robust loss vector in the set of robust loss behaviors:

$$\Pi_{\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}}(S) = \left\{(f(x_1,y_1),\dots,f(x_m,y_m)) : f\in\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}\right\}$$

maps to a 0-1 loss vector on the *inflated set* $S_{\mathcal{U}} = \left\{(z_1^1,y_1),\dots,(z_1^k,y_1),(z_2^1,y_2),\dots,(z_2^k,y_2),\dots,(z_m^1,y_m),\dots,(z_m^k,y_m)\right\}$,

$$\Pi_{\mathcal{H}}(S_{\mathcal{U}}) = \left\{(h(z_1^1),\dots,h(z_1^k),h(z_2^1),\dots,h(z_2^k),\dots,h(z_m^1),\dots,h(z_m^k)) : h\in\mathcal{H}\right\}.$$

Therefore, it follows that $\left|\Pi_{\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}}(S)\right| \leq |\Pi_{\mathcal{H}}(S_{\mathcal{U}})|$. Then, by applying the Sauer-Shelah lemma, it follows that $|\Pi_{\mathcal{H}}(S_{\mathcal{U}})| \leq O((mk)^{\mathrm{vc}(\mathcal{H})})$. Then, by solving for m such that $O((mk)^{\mathrm{vc}(\mathcal{H})}) \leq 2^m$, we get that $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \leq O(\mathrm{vc}(\mathcal{H})\log(k))$. ■

A.4 ERM failure Example

Appendix B. Risk Analysis

Per the call for papers for this workshop, in this section we will include our risk analysis, which is a novel contribution for the authors. Some of this analysis is general to theory papers in robustness and

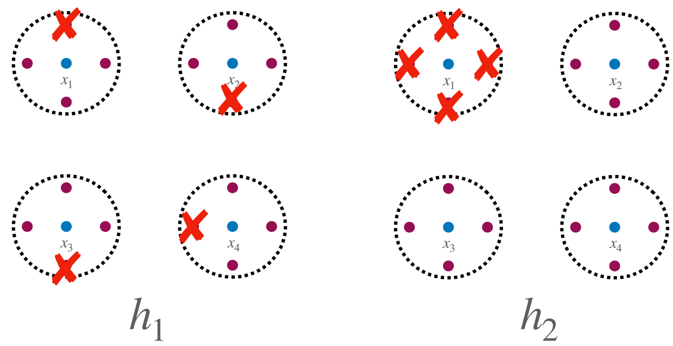

Figure 1: $\{x_i\}$ are natural data points in blue and each is surrounded by their perturbed points in purple. Red means mis-classified points. The approach in this section is to solve the ERM on an inflated data-set consisting of natural points and their perturbations. Observe that both h_1 and h_2 have the same $0-1$ loss on this toy data-set with four points but h_2 has much lower robust loss since it can correctly classify $3/4$ of the original examples no matter what the adversary does, while for h_1 the adversary can perturb any point to induce a mis-classification

some of it is specific to our work. Some of our risk analysis is based on discussion in Hendrycks and Mazeika (2022). This work attempts to mitigate existing risks due to patch attacks.

B.1 Short Term Risk of Patch Attacks

First, some discussion of the short term vulnerability of learned systems. Adversarial attacks of this nature lend themselves immediately to targeted attacks by malicious actors. For instance, if adversarial patch attacks remain a systemic flaw of vision models, and self driving cars with vulnerable vision systems are widely deployed, malefactors could dangerously target specific vehicles. Alternatively, if software to design universal adversarial patches continues to proliferate, then lone wolves could spread patches widely without a specific target and pose an acute and hard to mitigate risk to any driver.

In addition to acute harms caused by attackers using these systems, they could also delay or prevent the beneficial use of AI systems. This type of vulnerability could limit the reliance of automakers on vision systems or delay the implementation of self driving technology. While some of the safety benefits of self driving technology remain conjectural, something on the order of 50,000 Americans die per year in automotive accidents (NHTSA) and on the order of 1 million people annually WHO. There is a plausible argument that self driving technology can mitigate these risks by achieving super-human performance and consistent driving behavior.

Our work mitigates some of this patch risk by exhibiting an algorithm that can learn a classifier that competes with the global optima for this problem.

The most important limitation currently is we have not yet implemented an empirical evaluation. This is intended for a future version of this work.

B.2 Long Term Risk of Patch Attacks

We observe a key long term concern, the risk of catastrophic failures due to AI/ML based controllers subject to patch attacks or other perturbations. For instance while the author was writing this, they observed an advertisement in an undisclosed airport for 'AI for Air Traffic Control'. Safety critical systems increasingly have possibly vulnerable AI sub-systems. There are some efforts to integrate AI/ML techniques into nuclear command and control systems Lowther and McGiffin (2019).

Some of the hypothetical benefits of AI include possibly simplifying decision making for human actors by reducing information overload and giving them the time to make thoughtful choices Lowther and McGiffin (2019). If AI systems are too vulnerable Klare (2020), these benefits would be unrealized and some decision makers may remain stuck with sub-optimal choices. If patch attacks/adversarial attacks remain a credible threat and these control systems are deployed, that could have extreme consequences. For instance, if an early launch warning system had a satellite based vision component focused on missile silos, a patch attack could prevent early detection.

Alternatively, there is also a risk to continuing to use legacy and non-AI systems in that we may be stuck with poor human decision making or static systems.

Moving out in terms of generality, there are also questions raised by adversarial robustness about whether or not models can be relied upon to perform consistently, when subject to natural perturbations of distribution shift, and our work is progress in this direction.

References

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*, 23(175):1–31, 2022.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. *CoRR*, abs/2003.06693, 2020. URL `https://arxiv.org/abs/2003.06693`.

Uriel Feige, Yishay Mansour, and Robert Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657. PMLR, 2015a.

Uriel Feige, Yishay Mansour, and Robert E. Schapire. Learning and inference in the presence of corrupted inputs. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 637–657. JMLR.org, 2015b. URL `http://proceedings.mlr.press/v40/Feige15.html`.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research, 2022. URL `https://arxiv.org/abs/2206.05862`.

Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515. PMLR, 2018.

Michael T Klare. 'skynet' revisited: The dangerous allure of nuclear command automation, 2020. URL `https://www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation`.

Adam Lowther and Curtis McGiffin. MS Windows NT kernel description, 2019. URL `https://warontherocks.com/2019/08/america-needs-a-dead-hand/`.

Michael McCoyd, Won Park, Steven Chen, Neil Shah, Ryan Roggenkemper, Minjune Hwang, Jason Xinyu Liu, and David A. Wagner. Minority reports defense: Defending against adversarial patches. *CoRR*, abs/2004.13799, 2020. URL `https://arxiv.org/abs/2004.13799`.

Jan Hendrik Metzen and Maksym Yatsura. Efficient certified defenses against patch attacks on image classifiers. *CoRR*, abs/2102.04154, 2021. URL `https://arxiv.org/abs/2102.04154`.

Omar Montasser, Steve Hanneke, and Nati Srebro. Reducing adversarially robust learning to non-robust PAC learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien 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. URL https://proceedings.neurips.cc/paper/2020/hash/a822554e5403b1d370db84cfbc530503-Abstract.html.

NHTSA. "newly released estimates show traffic fatalities reached a 16-year high in 2021". URL https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities.

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

WHO. "road traffic injures". URL https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Chong Xiang and Prateek Mittal. Patchguard++: Efficient provable attack detection against adversarial patches. *CoRR*, abs/2104.12609, 2021. URL https://arxiv.org/abs/2104.12609.

Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: Provable defense against adversarial patches using masks on small receptive fields. *CoRR*, abs/2005.10884, 2020. URL https://arxiv.org/abs/2005.10884.

Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. Patchcleanser: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2065–2082, 2022.

Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pages 681–698. Springer, 2020.

Zhanyuan Zhang, Benson Yuan, Michael McCoyd, and David Wagner. Clipped bagnet: Defending against sticker attacks with clipped bag-of-features. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 55–61. IEEE, 2020.