Average Margin Regularization for Classifiers

Matt Olfat and Anil Aswani {molfat,aaswani}@berkeley.edu

Abstract— **Adversarial robustness has become an important research topic given empirical demonstrations on the lack of robustness of deep neural networks. Unfortunately, recent theoretical results suggest that adversarial training induces a strict tradeoff between classification accuracy and adversarial robustness. In this paper, we propose and then study a new regularization for any margin classifier or deep neural network. We motivate this regularization by a novel generalization bound that shows a tradeoff in classifier accuracy between maximizing its margin and average margin. We thus call our approach an average margin (AM) regularization, and it consists of a linear term added to the objective. We theoretically show that for certain distributions AM regularization can both improve classifier accuracy and robustness to adversarial attacks. We conclude by using both synthetic and real data to empirically show that AM regularization can strictly improve both accuracy and robustness for support vector machine's (SVM's), relative to unregularized classifiers and adversarially trained classifiers.**

I. INTRODUCTION

There is renewed interest in robust learning due to the observed fragility of deep classifiers to imperceptible adversarial corruptions [1]–[4]. Such classifiers are often key elements in a variety of cyber-physical systems (CPS) like self-driving cars, smart homes, or smart grids. Adversarial perturbations on sensors within CPS can have disastrous consequences, and this has been exhibited by several major attacks on mission-critical CPS [5]–[8]. In response, numerous adversarial training approaches have been proposed, both in the context of linear margin classifiers [9]–[11] and deep classifiers [2], [12]. These approaches train classifiers so as to minimize loss with respect to adversarially-perturbed data.

Interestingly, such adversarial methods have been shown to be equivalent to a particular type of regularization in the context of linear margin classifiers and regression [11], [13]–[15]. While regularization protects against (though does not always eliminate) overfitting [16], such outcomes critically depend upon having regularization that is congruous to the underlying data distributions [2], [9], [14], [17]–[20].

Here, we argue that adversarial training ignores notable attributes of the data. For instance, image data often has manifold structure [21]–[23]. Similar claims may be made about the dynamics of learned systems in system identification or reinforcment learning contexts [24], [25]. Yet adversarial training regularizes with respect to full-dimensional perturbations and not with respect to any underlying manifold

structure. This is significant because the imperceptibility of the most successful adversarial perturbations suggests that they lie orthogonal to these manifolds [1]. Thus, any robust methodology that does not exploit this kind of structure will likely remain susceptible to adversarial attacks.

Recently, [26] have claimed that there is a strict trade-off between the accuracy of a classifier and its robustness to adversarial perturbations. They augment their argument with demonstrations of this inverse relationship on a specific dataset, and claim that adversarial training best minimizes the cost of robustness. This paper shows that this is trade-off is not general and that, in fact, robustness and accuracy can grow concurrently for broad classes of datasets.

We make three contributions: First, we develop a novel generalization bound that shows classifier accuracy depends on a tradeoff between minimum and average margin. Second, we propose a new regularization that we call average margin (AM) regularization. This regularization consists of a linear term added to the objective, and is hence amenable to efficient numerical computation. We prove that for certain distributions, AM regularization can improve both accuracy and adversarial robustness of a classifier. Third, we use synthetic and real data to empirically show that AM regularization can generate support vector machine (SVM) classifiers that strictly dominate (in terms of accuracy and robustness) classifiers computed with or without adversarial training. Taken together, these results suggest that the phenomenon of adversarial fragility is an issue of overfitting rather than a fundamental issue unique to adversarial attacks.

A. Robust Linear SVM

Linear SVM relies upon on maximizing training margin by minimizing the *hinge-loss*, and several methods have been proposed to improve its robustness. Some approaches truncate the hinge-loss function [27]–[32], but a major issue with this approach is that it forfeits the convexity of the original problem. Other methods [33], [34] penalize outliers uniformly while maintaining the convexity of the hinge-loss, and [17] generates sparse projections to minimize the visibility of outliers. Instead of attempting to devalue outliers, [35] formulate a mixed-integer problem with the hinge-loss that removes them, and [19] redesigns the loss function entirely using bounds on the leave-one-out cross-validation error. Adversarial training has also been considered: [9]–[11] take a minimax approach, solving bilevel programs to design classifiers robust to worst-case perturbations of either a given distribution or a given magnitude. Robustness of classifiers to noise in the label, as opposed to only the features, has also been considered [11], [36], [37].

B. Robust Deep Classifiers

Adversarial fragility is pronounced in deep classifiers. [1] identified the sensitivity of deep learners to adversarial noise, and [2], [12] followed up with empirical and theoretical examinations of this instability. It has been shown that relatively minute and visually unrecognizable perturbations (in the case of images) can significantly impact the accuracy of these learners, and some work has been done on characterizing these minimal deviations [3], [4]. Notably, [38] showed that many existing methods for adversarially-robust deep classification are not wholly effective. However, there has been a spate of promising recent results in this direction, some of which come with theoretical guarantees [39]–[41].

C. Outline

Section II describes our notation and presents the problem setup that will be considered. Next, Section III introduces a novel generalization bound, proposes the average margin (AM) regularization, and then theoretically studies properties of this regularization for a specific distribution. Section III presents empirical results comparing linear classifiers that have been computed using different regularization and adversarial training approaches, using multiple datasets.

II. NOTATION AND SETUP

We use $\mathcal{N}(\mu, \Sigma)$ to refer to a multivariate normal distribution with mean μ and covariance matrix Σ. Also, let $\mathbf{0}_d$ and $\mathbf{1}_d$ refer to vectors of length d with all entries zero and one, respectively, and I_d to the $d \times d$ identity matrix. These subscripts will be dropped when the size is obvious due to context. In contrast, the function notation $\mathbf{1}(\cdot)$ refers to the indicator function. We use X_i to denote the i-th row of a matrix X. For some kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, let the matrix $K(X, X')$ be such that $K(X, X')_{ij} = k(X_i, X'_j)$.

Consider data observations $(x, y) \sim \mathcal{D}$ where $x \in \mathbb{R}^d$ and $y \in \{+1, -1\}$. In binary classification, the goal is compute a classifier $h : \mathbb{R}^d \to \{+1, -1\}$ to predict a label y from from a feature vector x. For a margin classifier from a family $\mathcal{H}$, this is achieved by minimizing $\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i h(x_i))$, which is the sample average of some given loss function $\ell(\cdot)$. The expected classification error rate of a classifier h is defined as $\mathcal{L}(h) = \mathbb{E}(\mathbf{1}(y \neq \text{sign}(h(x))))$.

Adversarial robustness for a classifier h refers to its ability to maintain accuracy in predicting a label y when given a corrupted corresponding feature vector $x + \delta$, where corruption δ is chosen by an adversary and has magnitude bounded by a quantity e. The expected adversarial classification error rate of a classifier h when the adversary can perturb data by e magnitude is $\mathcal{L}(h, e) = \mathbb{E}(\max_{\delta : \|\delta\| \leq e} \mathbf{1}(y \neq h(x + \delta)))$. The inner maximization is interpreted as an adversary choosing an attack. Also note that $\mathcal{L}(h, 0) = \mathcal{L}(h)$.

III. A NEW REGULARIZATION

Margin classifiers primarily use only data near the boundaries of different classes for the purpose of estimating the parameters of the classifier. However, in the low (relative to dimensionality) data regime this can be problematic.
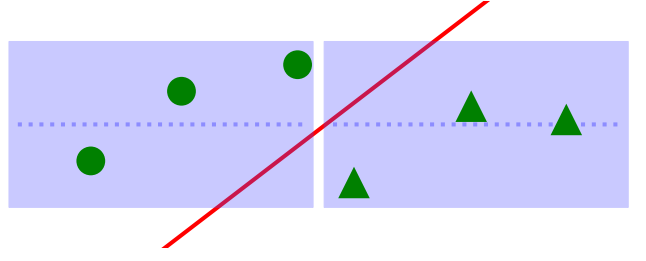

Fig. 1. This example shows that maximizing the margin on training data can reduce classification accuracy since it does not use data far from the margin boundary. The marks are sampled data, the two supports of the data distributions for the two labels $y \in \{-1, +1\}$ are the two shaded rectangles, and the dashed line is the maximum margin linear classifier.

Figure 1 shows an example where the usual margin classifier has issues. Maximizing the minimum margin leads to a classifier with high expected classification error because only a small amount of the data lies at the boundaries of the two classes. The manifold-like structure of the two classes leads to a situation where much of the data lies away from the boundary. A natural question to ask is how margin classifiers may be modified in order to better use data away from the boundary to improve predictions in situations similar to the above shown example.

Given the manifold-like example above, one possiblity is to use manifold regularization techniques. In fact, manifold regularization can be useful for regression [42]–[45]. However, a disadvantage of manifold regularization is it requires the indirect step of first estimating the manifold, and then using the estimated manifold for regularization. This indirect step can increase estimation error in a way that often reverses its regularizing effect. Our goal then is to design a regularizer that provides benefits in the manifold-like setting but does not require estimation of any manifolds.

In this section, we develop and study a new regularization for margin classifiers. We begin by proving a new generalization bound that demonstrates how maximizing the minimum margin does not always lead to minimal expected classification error. This generalization bound is used to motivate our new regularization for any margin classifier, which we call the average margin (AM) regularization. Next, we provide a probabilistic interpretation of this regularization in the context of deep learning. We conclude the section by discussing AM regularization in the special context of SVM. It is shown how this regularization can be used for kernel SVM, and then a result is given showing how AM regularization can simultaneously improve expected classification error and robustness to adversarial perturbations; this is significant because it is in direct contrast to results on adversarial training [26] that find a strict trade-off between classifier accuracy and robustness to adversarial perturbations.

A. Average Margin Generalization Bound

Classical results on the generalization error of classifiers [30] provide justification for maximizing the minimum margin of classifiers. Below, we present a new generalization bound in terms of the average margin of a classifier.

Theorem 1: Let $\mathcal{L}(h) = \mathbb{E}(\mathbf{1}(y \neq \text{sign}(h(x))))$ be the expected classification error rate of h, $K_\gamma(h) = \#\{i : y_i h(x_i) \leq \gamma\}/n$ be the fraction of data with γ-margin mistakes, $J(h) = \frac{1}{n} \sum_{i=1}^{n} y_i h(x_i)$ be the average classification margin, and suppose that $\sup_x |h(x)| \leq c$ for all $h \in \mathcal{H}$. Then for any $\zeta \in [0, 1]$ we have with probability at least $1 - 2\delta$ that

$$\mathcal{L}(h) \leq \zeta \cdot (1 - J(h)/c) + (1 - \zeta) \cdot K_\gamma(h) +$$
$$4 \frac{\mathcal{R}_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\log(\log_2 \frac{4c}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (1)$$

for all $\gamma \in (0, c]$ and all $h \in \mathcal{H}$.

Proof: This proof uses a similar argument to the proof of Theorem 2 by [46], with suitable modifications made to apply to our setting. Define the functions

$$l_\gamma(u) = \begin{cases} 1, & u \leq 0 \\ 1 - u/\gamma, & 0 < u < \gamma \\ 0, & u \geq \gamma \end{cases} \quad (2)$$

and $\ell_\gamma(u) = \zeta \cdot (1 - u/c) + (1 - \zeta) \cdot l_\gamma(u)$. Let $\mathcal{L}_\gamma(h) = \mathbb{E}(\ell_\gamma(yh(x)))$ and $\hat{\mathcal{L}}_\gamma(h) = \frac{1}{n} \sum_{i=1}^{n} \ell_\gamma(y_i h(x_i))$. We will consider the values $\gamma_k = c/2^k$ and $\delta_k = \delta/(k+1)^2$ for $k \in \{0, 1, \ldots\}$. Since $\ell_{\gamma_k}(u)$ is Lipschitz with constant $\zeta/c + (1 - \zeta)/\gamma_k \leq 1/\gamma_k$, applying Theorem 7 from [30] gives that

$$\mathcal{L}_{\gamma_k}(h) \leq \hat{\mathcal{L}}_{\gamma_k}(h) + \frac{2}{\gamma_k} \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta_k)}{2n}} \quad (3)$$

holds with probability at least $1 - \delta_k$ for all $h \in \mathcal{H}$. Next observe that for any $\zeta \in [0, 1]$, $\gamma > 0$, and any $h \in \mathcal{H}$; we have $\mathcal{L}(h) \leq \mathcal{L}_\gamma(h)$ and $\hat{\mathcal{L}}_\gamma \leq \zeta \cdot (1 - J(h)/c) + (1 - \zeta) \cdot K_\gamma(h)$. Thus with probability at least $1 - \delta_k$ we have

$$\mathcal{L}(h) \leq \zeta \cdot (1 - J(h)/c) + (1 - \zeta) \cdot K_{\gamma_k}(h) +$$
$$\frac{2}{\gamma_k} \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta_k)}{2n}} \quad (4)$$

for all $h \in \mathcal{H}$. Applying the union bound over all $k \in \{0, 1, \ldots\}$ gives that (4) holds with probability at least $1 - \pi^2 \delta/6 \geq 1 - 2\delta$ for all $k \in \{0, 1, \ldots\}$ and $h \in \mathcal{H}$. Now we will assume that this union event occurs. This implies (1) holds for $\gamma = c$. Next consider $\gamma \in (0, c)$, and choose the k such that $\gamma_k \leq \gamma < \gamma_{k-1}$. Note $k \leq \log_2(c/\gamma) + 1$, and so

$$\mathcal{L}(h) \leq \zeta \cdot (1 - J(h)/c) + (1 - \zeta) \cdot K_\gamma(h) +$$
$$\frac{4}{\gamma} \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta) + 2\log(\log_2(4c/\gamma))}{2n}} \quad (5)$$

since $K_{\gamma_k}(h) \leq K_\gamma(h)$, $1/\gamma_k \leq 2/\gamma$, and $\log(1/\delta_k) \leq \log(1/\delta) + 2\log(\log_2(4c/\gamma))$. ∎

B. Average Margin Regularization

Given the above intuition, we propose that the average margin $\frac{1}{n} \sum_{i=1}^{n} y_i h(x_i)$ can be used as a regularization term for any margin classifier. Specifically, an AM-regularized classifier can be computed by solving

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i h(x_i)) - \mu \cdot \frac{1}{n} \sum_{i=1}^{n} y_i h(x), \quad (6)$$

where $\mu \geq 0$ is a tuning parameter. Note that we subtract the average margin term because we are minimizing.

Next, we consider deep learning with activation function $\exp(h(x))/(1 + \exp(h(x)))$. The logistic loss is often used to construct classifiers, and the corresponding AM-regularized network is computed by solving

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i h(x_i))) - \mu \cdot \frac{1}{n} \sum_{i=1}^{n} y_i h(x_i), \quad (7)$$

where $\mu \geq 0$ is again a tuning parameter. Now suppose we use the labels $t = (1 + y)/2 \in \{0, 1\}$, and we make the identifications that $\hat{p}_0(x) = 1/(1 + \exp(h(x)))$ and $\hat{p}_1(x) = \exp(h(x))/(1 + \exp(h(x)))$. Then training the AM-regularized network is equivalent to solving

$$\min_{h \in \mathcal{H}} -\frac{1}{n} \sum_{i=1}^{n} \left[t_i \log(\hat{p}_1(x)) + (1 - t_i) \log(\hat{p}_0(x)) \right] +$$
$$- \mu \cdot \frac{1}{n} \sum_{i=1}^{n} t_i \log \frac{\hat{p}_1(x)}{\hat{p}_0(x)} +$$
$$- \mu \cdot \frac{1}{n} \sum_{i=1}^{n} (1 - t_i) \log \frac{\hat{p}_0(x)}{\hat{p}_1(x)}. \quad (8)$$

The first term is cross-entropy, while the last two terms are negative log-likelihood ratios. Thus, for deep learning our AM regularization encourages larger log-likelihood ratios. Recalling hypothesis testing, a larger log-likelihood ratio makes it easier to distinguish between classes.

C. Special Case of SVM

Here, we consider AM regularization in the special case of SVM. First, consider linear SVM with $h(x) = x^\mathsf{T}\beta + b$. Then the AM-regularized linear SVM is given by

$$\min \lambda\|\beta\|^2 + \frac{1}{n} \sum_{i=1}^{n} s_i - \mu \cdot \frac{1}{n} \sum_{i=1}^{n} y_i(x_i^\mathsf{T}\beta + b)$$
$$\text{s.t. } y_i(x_i^\mathsf{T}\beta + b) \geq 1 - s_i, \quad \text{for } i = 1, \ldots, n \quad (9)$$
$$s_i \geq 0, \quad \text{for } i = 1, \ldots, n$$

As seen above, one advantage is that AM regularization is simply a linear term in the objective function. Thus, AM regularization does not significantly affect the computational complexity of solving the linear SVM optimization problem. Another benefit of AM regularization is it can be dualized, which allows for its use in kernel SVM. A standard argument using the KKT conditions shows that the kernel SVM with AM regularization is computed by solving

$$\min z^\mathsf{T}(yy^\mathsf{T} \circ K(X, X))z - \mathbf{1}_n z^\mathsf{T}$$
$$\text{s.t. } yz^\mathsf{T} = 0 \quad (10)$$
$$\frac{\mu}{1+\mu} \cdot \frac{1}{\lambda n} \leq z \leq \frac{1}{\lambda n}$$

This kernel SVM formulation provides further intuition about AM regularization: It shows that increasing μ increases the impact of points away from the margin, thereby mixing the original support vectors with an average over all data points.

Last, we show that AM regularization can generate classifiers that both improve classification accuracy and robustness. Let $\mathcal{L}(h, e) = \mathbb{E}(\max_{\delta:\|\delta\|\leq e} \mathbf{1}(y \neq h(x + \delta)))$ be the generalization error of classifier h when the adversary can perturb data by e magnitude. We specifically prove a result that formalizes the intuition shown in Figure 1.

Proposition 1: Let $\hat{h}_{AM}$ and $\hat{h}_{L2}$ be the linear classifiers computed by SVM with and without AM regularization, respectively, using $n = 4k \geq 4$ points sampled from a data distribution. Recalling (9), we assume λ is chosen so that all sampled points lie beyond the margin, that μ can be chosen based on the sampled data, and that without loss of generality the linear classifier has no intercept term (i.e. $h(x) = x^\mathsf{T}\beta$). There exists a data distribution such that

$$\mathcal{L}(\hat{h}_{AM}, e) < \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [0, 7\sqrt{5}/25)$$
$$\mathcal{L}(\hat{h}_{AM}, e) = \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [7\sqrt{5}/25, 15\sqrt{5}/25)$$
$$\mathcal{L}(\hat{h}_{AM}, e) > \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [15\sqrt{5}/25, \sqrt{5})$$
$$\mathcal{L}(\hat{h}_{AM}, e) = \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [\sqrt{5}, 2\sqrt{5})$$
$$\mathcal{L}(\hat{h}_{AM}, e) < \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [2\sqrt{5}, 110\sqrt{5}/25)$$
$$\mathcal{L}(\hat{h}_{AM}, e) = \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [110\sqrt{5}/25, \infty)$$

Proof: We consider a balanced data distribution with $y_{2i-1} = +1$ and $y_{2i} = -1$ for $i = 1, \ldots, 2k$. Suppose $x_{2i-1} = (10, 0)$ and $x_{2i} = (-10, 0)$ for $i = k+1, \ldots, 2k$. For $i = 1, \ldots, k$: let $a_i = +1$ be the event that $x_{2i-1} = (1, 2)$ and $x_{2i} = (-1, -2)$, and let $a_i = -1$ be the event that $x_{2i-1} = (1, -2)$ and $x_{2i} = (-1, 2)$. We assume the a_i are independent Rademacher random variables.

Let $I = \{1, \ldots, k\}$, and note the margin assumption on λ implies the classifier satisfies $y_i x_i^\mathsf{T}\beta \geq 1$ for all $i = 1, \ldots, n$. Define the conditional expectation $\mathcal{L}(h, e, \mathcal{A}) = \mathbb{E}[\max_{\delta: \|\delta\| \leq e} \mathbf{1}(y \neq h(x+\delta)) | \mathcal{A}]$ for some event $\mathcal{A}$. Clearly, choosing $\mu = 0$ removes the effect of AM regularization and ensures that $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{A}) \leq \mathcal{L}(\hat{h}_{L2}, e, \mathcal{A})$.

Now consider the event $\mathcal{B}$ where $a_i = +1$ for all $i \in I$. Here, the margin assumption means the classifier must satisfy $10\beta_1 \geq 1$, $\beta_1 + 2\beta_2 \geq 1$. A straightforward calculation gives that $\hat{h}_{L2}$ has $\hat{\beta}_{L2} = (\frac{1}{5}, \frac{2}{5})$ and $\mathcal{L}(\hat{h}_{L2}, e, \mathcal{B}) = \frac{1}{4} + \frac{1}{4}\mathbf{1}(e \geq \sqrt{5}) + \frac{1}{2}\mathbf{1}(e \geq 2\sqrt{5})$. The AM-regularized SVM is

$$\min \lambda\|\beta\|^2 - 2\mu k \cdot (\beta_1 + 2\beta_2) - 2\mu k \cdot (10\beta_1)$$
$$\text{s.t. } 10\beta_1 \geq 1 \quad\quad\quad (11)$$
$$\beta_1 + 2\beta_1 \geq 1$$

Now suppose we choose the largest μ such that the optimal solution satisfies $10\beta_1 > 1$ and $\beta_1 + 2\beta_1 = 1$. Then it can be easily verified that this largest value is $\mu = \lambda/(15k)$, and so a straightforward calculation gives that $\hat{h}_{AM}$ has $\hat{\beta}_{AM} = (\frac{11}{15}, \frac{2}{15})$ and $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{B}) = \frac{1}{4}\mathbf{1}(e \geq 7\sqrt{5}/25) + \frac{1}{4}\mathbf{1}(e \geq 15\sqrt{5}/25) + \frac{1}{2}\mathbf{1}(e \geq 110\sqrt{5}/25)$. This means that $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{B}) < \mathcal{L}(\hat{h}_{L2}, e, \mathcal{B})$ for $e \in [0, \sqrt{5})$ and $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{B}) = \mathcal{L}(\hat{h}_{L2}, e, \mathcal{B})$ for $e \geq \sqrt{5}$. And as discussed earlier, choosing $\mu = 0$ for the event $\neg\mathcal{B}$ ensures that $\mathcal{L}(\hat{h}_{AM}, e, \neg\mathcal{B}) \leq \mathcal{L}(\hat{h}_{L2}, e, \neg\mathcal{B})$. Next note $\mathcal{B}, \neg\mathcal{B}$ partition the sample space, and that $\mathbb{P}(\mathcal{B}) > 0$. Hence the result follows from the law of total expectation. ∎

This is a more subtle result than that of [26], which considers $\mathcal{L}(\cdot, e)$ for adversarially trained classifiers at only two discrete values of e. Our analysis shows that AM regularization can both increase classifier accuracy (i.e., at $e = 0$) and robustness to adversarial perturbations for specific data distributions. This is seen because AM regularization generally has lower expected classification error over the whole range of e except for a very narrow range. The AM regularization is less robust when $e \in [15\sqrt{5}/25, \sqrt{5})$ because it has made a careful tradeoff between maximizing the margin and maximizing the average margin.

Another point to note is that for the data distribution in the above proposition, the adversarially trained SVM does not improve upon the standard SVM.

Proposition 2: Let $\hat{h}_{D\gamma}$ be the linear classifier computed by adversarially trained SVM where the adversary can perturb data by $\gamma \in [0, \sqrt{5}]$ and let $\hat{h}_{L2}$ be the linear classifier computed by SVM, using $n = 4k \geq 4$ points sampled from a data distribution. Recalling (9), we assume λ is chosen so all sampled (and perturbed) points lie beyond the margin, and that without loss of generality the linear classifier has no intercept term (i.e. $h(x) = x^\mathsf{T}\beta$). Then, for the distribution that is considered in the proof of Proposition 1, we have that $\mathcal{L}(\hat{h}_{D\gamma}, e) = \mathcal{L}(\hat{h}_{L2}, e)$ for all $e \geq 0$.

Proof: The proof of this proposition follows a similar argument to that for Proposition 1. ∎

IV. EMPIRICAL RESULTS

Here, we use synthetic and real data to compare AM regularization to other regularizations, including adversarial training, for linear SVM. We first present the benchmark regularization methods that we compare AM regularization to. Next, we present empirical results for linear SVM using synthetic data, and we conclude this section by presenting numerical results for linear SVM applied to a real dataset.

A. Benchmarks

We compare linear SVM with AM regularization to: SVM with ℓ_2 regularization, SVM with ℓ_1 regularization, the robust SVM training method of [33], and the adversarial training method outlined by [10], [11], [14].

a) Song et al.: This method relies on minimizing the impact of outliers on the design of the separator. Specifically, it avoids the over-reliance on such outliers by shifting the loss according to the distance of a point to its respective centroid:

$$\min \lambda\|\beta\|^2 + $$
$$\frac{1}{n}\sum_{i=1}^{n} \left(1 - y_i(\beta^\mathsf{T} x_i + b) - \gamma\|x_i - \mu_{y_i}\|_2^2\right)_+, \quad (12)$$

where μ_{y_i} is the centroid for class y_i, and $\lambda, \gamma \geq 0$ are tuning parameters.

b) Adversarial Training: Adversarial training formulates SVM training as a bilevel program, where the loss is minimized not over the original data $\{x_i\}_{i=1}^n$, but rather over worst-case perturbations of the data $\{x_i + \delta_i\}_{i=1}^n$, where the perturbations δ_i are restricted to be in some space. If we restrict $\|\delta_i\|_q \leq \gamma$ for some norm $\|\cdot\|_q$, this is modeled as

$$\min \lambda\|\beta\|^2 + \frac{1}{n}\sum_{i=1}^{n} s_i$$
$$\text{s.t. } y_i(\beta^\mathsf{T} x_i + b) - \gamma\|\beta\|_{q^*} \geq 1 - s_i, \text{ for } i = 1, \ldots, n$$
$$s_i \geq 0, \quad\quad\quad\quad\quad\quad\quad\quad \text{for } i = 1, \ldots, n$$
$$(13)$$

where $\|\cdot\|_{q^*}$ is the dual norm of $\|\cdot\|_q$. In this section, we consider the case where $q = 2$.
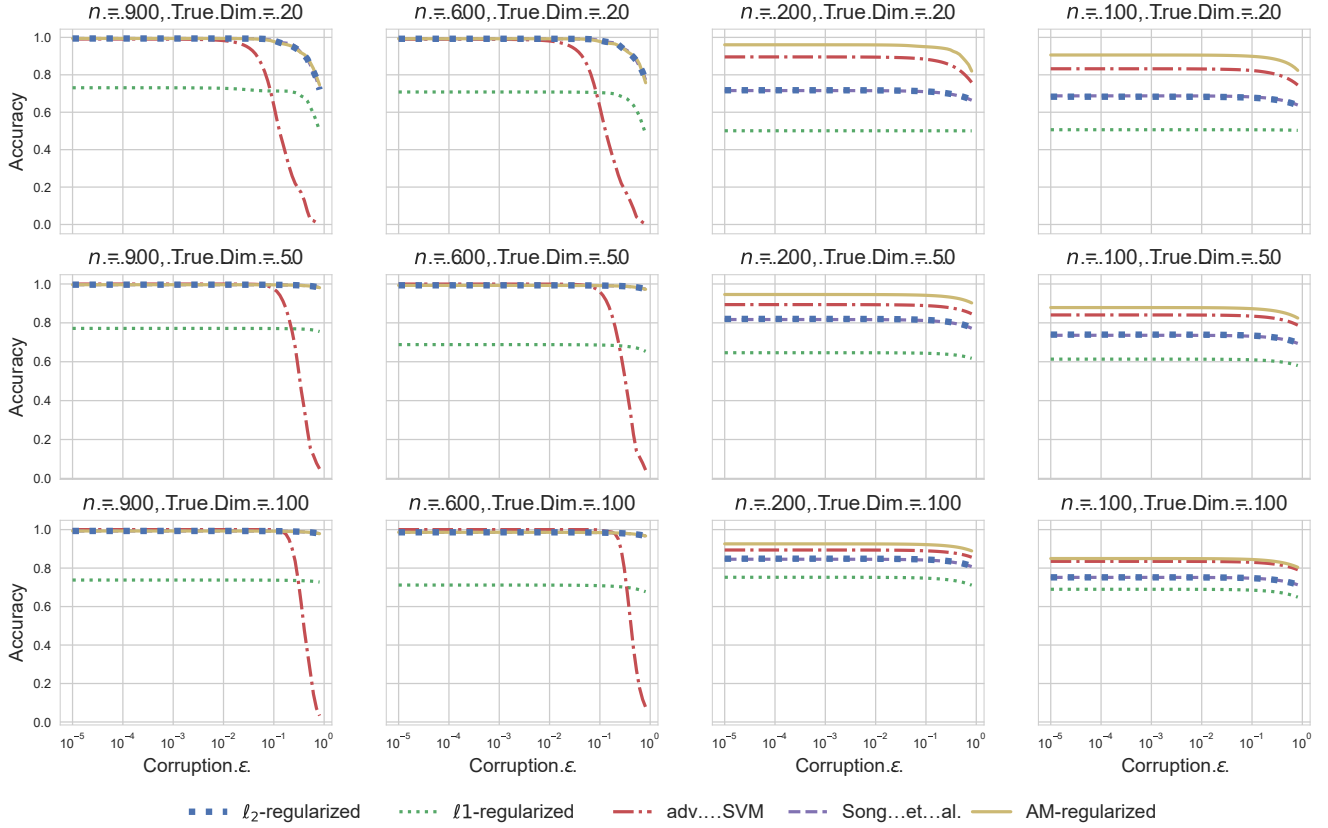
Fig. 2. Robustness to adversarial corruptions of classifiers trained on synthetic data, with accuracy measured as the fraction correctly classified.

B. Synthetic Data

We next run randomized experiments to demonstrate situations where AM regularization is beneficial. We generate U_+, U_- by taking the Q from a QR decomposition done on a matrix of size $d \times m$ with standard normal entries, and $\mu_+, \mu_- \in \mathbb{R}^d$ to have uniformly-distributed entries between 0 and 2. Then, we set $x = \Pi_{U_+}\mu_+ + U_+v_+ + \varepsilon_+$ when $y = +1$, and $x = \Pi_{U_-}\mu_- + U_-v_- + \varepsilon_-$ when $y = -1$, where $v_+ \sim \mathcal{N}(\mu_+, I_m)$, $\varepsilon_+ \sim \mathcal{N}(\mathbf{0}_d, \varepsilon I_d)$, $v_- \sim \mathcal{N}(\mu_-, I_m)$, and $\varepsilon_- \sim \mathcal{N}(\mathbf{0}_d, \varepsilon I_d)$. A training set of n samples was created. All models were then tested using 10,000 samples with adversarial corruption of various magnitude. This procedure was repeated 100 times with $\varepsilon = 0.01$ and $d = 200$, and the results for various training-set sizes and values of n are presented in Figure 2. Our method improves both robustness and accuracy in low-data and low-dimensionality regimes.

C. MNIST Experiments

Next, we use the classic MNIST dataset [47] to evaluate AM regularization. In Table I, we compare AM-regularization to the benchmark methods on the fraction of test points correctly classified. For simplicity, we focus on the binary classification problem of separating hand-drawn 0's from hand-drawn 1's, and report results from training on

TABLE I
FRACTION CORRECTLY CLASSIFIED WHEN CLASSIFYING 0 VS. 1 IN MNIST DATASET

CORRUPTION	0.01	0.2	1.0
ℓ_2-REGULARIZED	0.940	0.938	0.921
ℓ_1-REGULARIZED	0.997	0.995	0.714
AM-REGULARIZED	0.998	0.997	0.983
ADVERSARIAL	0.996	0.995	0.979
SONG ET AL.	0.994	0.992	0.974

10% of available training data. All hyperparameter values were set using 5-fold cross-validation on the training set. We repeat this process 50 times for each model, and report the accuracy results on a common testing set corrupted with various degrees of adversarial perturbations. We note that our method achieves better or almost equivalent accuracy with low corruption, and is best able to retain that level of accuracy at high levels of corruption.

V. CONCLUSION

Based on a novel generalization bound, we have proposed in this paper a new form of regularization for margin clas-

sifiers that we call average margin (AM) regularization. Our theoretical and empirical results support its use by showing that AM regularization can increase both classifier accuracy and adversarial robustness. Taken together, our results suggest that adversarial fragility is an issue of overfitting, rather than a fundamental uniqueness of adversarial perturbations.

One future topic is to better understand AM regularization's theoretical properties. We believe AM regularization works best in the finite sample regime and that its asymptotic behavior when $\mu \not\to 0$ may be poor. Another topic is modifications of AM regularization. For instance, consider a hinged average margin (HAM) regularization that adds $-\mu \times \frac{1}{n}\sum_{i=1}^{n}(\gamma - y_i h(x_i))^+$, where $\mu, \gamma \geq 0$ are tuning parameters. HAM regularization may improve AM regularization by providing saturation, whereby points *very* far (specifically γ distance away) from the margin are not considered in the average margin calculation. Promisingly, this is convex in $y_i h(x_i)$. Furthermore, we would like to investigate the efficacy of AM regularization for deep classifiers.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[2] A. Fawzi, O. Fawzi, and P. Frossard, "Fundamental limits on adversarial robustness," in *ICML Workshop on Deep Learning*, 2015.

[3] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *KDD*. ACM, 2004, pp. 99–108.

[4] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[5] M. Abrams and J. Weiss, "Malicious control system cyber security attack case study–Maroochy water services, australia," *MITRE*, 2008.

[6] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[7] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

[8] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke, "Security for 4g and 5g cellular networks: A survey of existing authentication and privacy-preserving schemes," *Journal of Network and Computer Applications*, vol. 101, pp. 55–82, 2018.

[9] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *Journal of Machine Learning Research*, vol. 3, no. Dec, pp. 555–582, 2002.

[10] T. B. Trafalis and R. C. Gilbert, "Robust support vector machines for classification and computational issues," *Optimisation Methods and Software*, vol. 22, no. 1, pp. 187–198, 2007.

[11] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo, "Robust classification," *J. Mach. Learn. Res*, 2017.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[13] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and lasso," in *NeurIPS*, 2009, pp. 1801–1808.

[14] ——, "Robustness and regularization of support vector machines," *J. Mach. Learn. Res,*, vol. 10, no. Jul, pp. 1485–1510, 2009.

[15] R. Livni, K. Crammer, and A. Globerson, "A simple geometric interpretation of SVM using stochastic adversaries," in *AISTATS*, 2012, pp. 722–730.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, no. 2. ACM, 2001, pp. 37–46.

[18] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[19] J. Weston and R. Herbrich, "Adaptive margin support vector machines," *NeurIPS*, pp. 281–296, 1999.

[20] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," in *NeurIPS*, 2005, pp. 161–168.

[21] S. Gerber, T. Tasdizen, S. Joshi, and R. Whitaker, "On the manifold structure of the space of brain images," in *MICCAI*, 2009, pp. 305–312.

[22] R. Pless and R. Souvenir, "A survey of manifold learning for images," *IPSJ Transactions on Computer Vision and Applications*, vol. 1, pp. 83–94, 2009.

[23] G. Peyré, "Manifold models for signals and images," *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 249–260, 2009.

[24] A. Roberts, "The utility of an invariant manifold description of the evolution of a dynamical system," *SIAM journal on mathematical analysis*, vol. 20, no. 6, pp. 1447–1458, 1989.

[25] L.-C. Shi and B.-L. Lu, "Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning," in *EMBC*, 2010, pp. 6587–6590.

[26] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "There is no free lunch in adversarial robustness (but there are unexpected benefits)," *arXiv preprint arXiv:1805.12152*, 2018.

[27] N. Krause and Y. Singer, "Leveraging the margin more carefully," in *ICML*, 2004, p. 63.

[28] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Trading convexity for scalability," in *ICML*, 2006, pp. 201–208.

[29] Y. Liu and X. Shen, "Multicategory ψ-learning," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 500–509, 2006.

[30] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[31] S. Suzumura, K. Ogawa, M. Sugiyama, and I. Takeuchi, "Outlier path: A homotopy algorithm for robust SVM," in *ICML*, 2014, pp. 1098–1106.

[32] Y.-l. Yu, M. Yang, L. Xu, M. White, and D. Schuurmans, "Relaxed clipping: A global training method for robust regression and classification," in *NeurIPS*, 2010, pp. 2532–2540.

[33] Q. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 32, no. 4, pp. 440–448, 2002.

[34] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos, "On the design of robust classifiers for computer vision," in *CVPR*, 2010, pp. 779–786.

[35] L. Xu, K. Crammer, and D. Schuurmans, "Robust support vector machine training via convex outlier ablation," in *AAAI*, vol. 6, 2006, pp. 536–542.

[36] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Asian Conference on Machine Learning*, 2011, pp. 97–112.

[37] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *NeurIPS*, 2013, pp. 1196–1204.

[38] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 3–14.

[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[40] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018.

[41] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," *arXiv preprint arXiv:1801.09344*, 2018.

[42] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *JMLR*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[43] P. J. Bickel and B. Li, "Local polynomial regression on unknown manifolds," in *Complex Datasets and Inverse Problems*. Institute of Mathematical Statistics, 2007, pp. 177–186.

[44] A. Aswani, P. Bickel, and C. Tomlin, "Regression on manifolds: Estimation of the exterior derivative," *The Annals of Statistics*, vol. 39, no. 1, pp. 48–81, 2011.

[45] M.-Y. Cheng and H.-T. Wu, "Local linear regression on manifolds and its geometric interpretation," *Journal of the American Statistical Association*, vol. 108, no. 504, pp. 1421–1434, 2013.

[46] S. M. Kakade, K. Sridharan, and A. Tewari, "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization," in *NeurIPS*, 2009, pp. 793–800.

[47] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.