Guarantees on Learning Depth-2 Neural Networks Under a Data Poisoning Attack

Sayar Karmakar[1] [0000−0002−4841−2069], Anirbit Mukherjee[2,3] [0000−0001−5189−8939], and Ramchandran Muthukumar[3] [0000−0003−0370−7987]

[1] University of Florida
sayarkarmakar@ufl.edu
[2] University of Pennsylvania
anirbit@wharton.upenn.edu
[3] Johns Hopkins University
rmuthuk1@jhu.edu

Abstract. In this work, we study the possibility of defending against "data-poisoning" attacks while learning a neural net. We focus on the supervised learning setup for a class of finite-sized depth-2 nets - which include the standard single filter convolutional nets. For this setup we attempt to learn the true label generating weights in the presence of a malicious oracle doing stochastic bounded and additive adversarial distortions on the true labels being accessed by the algorithm during training. For the non-gradient stochastic algorithm that we instantiate we prove (worst case nearly optimal) trade-offs among the magnitude of the adversarial attack, the accuracy, and the confidence achieved by the proposed algorithm. Additionally, our algorithm uses mini-batching and we keep track of how the mini-batch size affects the convergence.

Keywords: Adversarial attack · Neural network · non-gradient iterative algorithms · stochastic algorithms · non-smooth non-convex optimization

1 Introduction

The seminal paper [36] was among the first to highlight a key vulnerability of state-of-the-art network architectures like GoogLeNet, that adding small imperceptible adversarial noise to *test data* can dramatically impact the performance of the network. In these cases despite the vulnerability of the predictive models to the distorted input, human observers are still able to correctly classify this adversarially corrupted data.

In the last few years, experiments with adversarially attacked test data have been replicated on several state-of-the-art neural network implementations [13,28,3,14]. This phenomenon has also resulted in new *adversarial defenses* being proposed to counter the attacks. Such empirical observations have been systematically reviewed in [1,29].

But on the other hand the case of "data poisoning" or *adversarially attacked training data* [37], [42], [19] has received much less attention from theoreticians and in this work we take some steps towards bridging that gap.

Here we quickly review the conventional and the more studied mathematical setup of *adversarial risk*. Suppose $\mathcal{Z}$ is the measure space where the data lives, $\mathcal{H}$ is the hypothesis space where the predictor is being searched, and suppose the loss function is mapping, $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$. Then an *adversarial attack* is a set map $\boldsymbol{A} : \mathcal{Z} \to 2^{\mathcal{Z}}$. If $\mathcal{D}$ is the data distribution/probability measure on $\mathcal{Z}$ then given an adversarial attack $\boldsymbol{A}$, the *adversarial risk* of a hypothesis $h \in \mathcal{H}$ is defined as, $\mathcal{R}(h; \mathcal{D}, \ell, \boldsymbol{A}) := \mathbb{E}_{z \sim \mathcal{D}} \left[\sup_{\tilde{z} \in \boldsymbol{A}(z)} \ell(h, \tilde{z}) \right]$

The adversarial learning task is to find $h_* \in \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{R}(h; \mathcal{D}, \ell, \boldsymbol{A})$. Any hypothesis h with low adversarial risk will be stable to attacks of type $\boldsymbol{A}$ on the test data. This optimization formulation of adversarial robustness has been extensively explored in recent years. : multiple attack strategies have been systematically catalogued in [7,23,34], computational hardness of finding an adversarial risk minimizing hypothesis has been analyzed in [4,6,33,25], the issue of certifying adversarial robustness of a given predictor has been analyzed in [30,31] and bounds on the Rademacher complexity of adversarial risk have been explored in [41,17].

To the best of our knowledge theoretical progress with understanding the limits of training of a neural network under either adversarial attacks on test data or training data, have been restricted to the kernel regime of deep-learning (i.e when the nets are asymptotically large). See [10,21].

Thus it still remains an open challenge to demonstrate an example of provably robust training when **(a)** the neural network is realistically large/finite, **(b)** the algorithm has the common structure of being iterative and stochastic in nature and **(c)** the training data is being adversarially attacked. In this work we take a few steps towards this goal by working in a framework inspired by the *causative attack model* [35,2]. The learning task in this attack model is framed as a game between a *defender* who seeks to learn and an *attacker* who aims to prevent this. In a typical scenario the defender draws a finite number of samples from the true distribution (say S_c) and for some $\epsilon \in (0, 1)$ the attacker mixes into the training data a set S_p of (maybe adaptively) corrupted training samples such that $|S_p| = \epsilon |S_c|$. Now the defender has to train on the set $S_p \cup S_c$. *We note that our model differs from this causative attack model because we allow for any arbitrary fraction (including all) of the training data to be corrupted in an online fashion by the bounded additive adversarial attack on the true labels*

We also note that we consider the case of *regression* on nets, which is far less explored than the case of adversarially robust classification by neural nets. To the best of our knowledge previous work on robust regression have been limited to linear functions and have either considered corruptions that are limited to a small subset [15], [38] of the input space/feature predictors or make structural assumptions on the corruption. Despite the substantial progress with understanding robust linear regression [39], [40], [8], [5], [24], [22], [20], [26], the corresponding questions have remained open for even simple neural networks.

Our first key step is to make a careful choice of the neural network class to work with, as given in Definition 1. Secondly, for the optimization algorithm, we draw inspiration from the different versions of the iterative stochastic non-gradient *Tron* algorithms analyzed in the past, [32], [27], [9], [16], [18], [11], [12]. We generalize this class of algorithms to a form as given in Algorithm 1 and we run it in the presence of an adversarial oracle that we instantiate which is free to make any additive bounded perturbation to the *true* output generated by a network of the same architecture as the one being trained on. *Beyond boundedness, we impose no other distributional constraint on how the adversary additively distorts the true label.*

Outline of the paper In section 1.1 we give the mathematical setup of the kind of nets, distributions and data-poisoning adversary that we will use and define our algorithm.

We state our main result, Theorem 1 in section 2. The theorem shows that against our adversary this algorithm while trying to recover the original net's parameters, achieves a certain trade off between accuracy, confidence and the maximum allowed perturbation that the adversary is allowed to make. The proof of Theorem 1 is given in section 3 and the appendices contain multiple lemmas that are needed in the proof. For restriction on space we post the appendix to `https://sayarkarmakar.github.io/publications/guaranteespapersupp.pdf`

We posit that an important question about any such proof of defense against an adversarial attack is to want to know if the risk of the learnt predictor can be better than the maximum per-data damage that the adversary was allowed to do. We investigate this issue for our context in section 2.1 and demonstrate this as being possible within our framework.

Further in section 2.2 we explain that the accuracy-confidence-attack trade-off we obtain is nearly optimal in the worst-case.

We conclude in section 4 by motivating certain directions of future research that can be taken up in this theme.

1.1 The Mathematical Setup

As alluded to earlier we move away from the adversarial risk framework. We work in the supervised learning framework where we observe (input,output) data pairs $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are measure spaces. Let $\mathcal{D}$ be the distribution over the measure space $\mathcal{Z}$ and let $\mathcal{D}_{\text{in}}$ be the marginal distribution over the input space $\mathcal{X}$. If $\mathcal{H}$ is the hypothesis space for our learning task then let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ be a loss function. We model an adversarial oracle as a map $\boldsymbol{O_A} : \mathcal{Z} \mapsto \mathcal{Z}$, that corrupts the data $\mathbf{z} \in \mathcal{Z}$ with the intention of making the learning task harder. The learner observes only the corrupted data $\mathbf{z}_{\text{adv}} := \boldsymbol{O_A}(\mathbf{z})$ where $\mathbf{z} \sim \mathcal{D}$ and aims to find a hypothesis $h \in \mathcal{H}$ that minimizes the *true* risk, $\mathcal{R}(h; \mathcal{D}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}} \left[\ell(h, z) \right]$

In this work, we specifically consider the instance when $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}$, l is the square loss function and the hypothesis space $\mathcal{H} := \mathcal{F}_{k,\alpha,\mathcal{A},\mathcal{W}}$, the class of depth-2, width-k neural networks defined as follows,

Definition 1 (Single Filter Neural Nets of Depth-2 and Width-k). *Given a set of k sensing matrices $\mathcal{A} = \{A_i \in \mathbb{R}^{r \times n} \mid i = 1, \ldots, k\}$, an α-leaky ReLU activation mapping, $\mathbb{R} \ni y \mapsto \sigma(y) = y\mathbf{1}_{y \geq 0} + \alpha y \mathbf{1}_{y<0} \in \mathbb{R}$ and a filter space $\mathcal{W} \subseteq \mathbb{R}^r$, we define the function class $\mathcal{F}_{k,\alpha,\mathcal{A},\mathcal{W}}$ as,*

$$\mathcal{F}_{k,\alpha,\mathcal{A},\mathcal{W}} = \left\{ f_{\mathbf{w}} : \mathbb{R}^n \ni \mathbf{x} \mapsto \frac{1}{k} \sum_{i=1}^{k} \sigma\Big(\mathbf{w}^\top A_i \mathbf{x}\Big) \in \mathbb{R} \mid \mathbf{w} \in \mathcal{W} \right\}$$

Note that the above class of neural networks encompasses the following common instances, **(a) single ReLU gates** as $\mathcal{F}_{1,0,\{\mathrm{I}_{n \times n}\},\mathbb{R}^n}$ and **(b) Depth-2, Width-k Convolutional Neural Nets** when the sensing matrices A_i are such that each have exactly one 1 in each row and at most one 1 in each column and rest entries being zero.

Optimization with an Adversarial Oracle We assume that $\exists\ \mathbf{w}^* \in \mathbb{R}^r$ such that the adversarial oracle acts as, $\mathbb{R}^n \ni \mathbf{x} \mapsto \boldsymbol{O_A}(\mathbf{x}) = f_{\mathbf{w}^*}(\mathbf{x}) + \xi_{\mathbf{x}} \in \mathbb{R}$ while $|\xi_{\mathbf{x}}| \leq \theta$ for some fixed θ and our risk minimization optimization problem can be be stated as : $\text{argmin}_{\mathbf{w} \in \mathcal{W}} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{in}}} \left[\left(f_{\mathbf{w}^*}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x}) \right)^2 \right]$ where $\forall \mathbf{x} \in \mathbb{R}^n$ one only has access to $\boldsymbol{O_A}(\mathbf{x})$ as defined above. Towards being able to solve the described optimization problem we make the following assumptions about the data distribution $\mathcal{D}_{\text{in}}$,

Assumptions 2.1 : Parity Symmetry

We assume that the input distribution $\mathcal{D}_{\text{in}}$ is symmetric under the parity transformation i.e if $\mathbf{x}$ is a random variable such that $\mathbf{x} \sim \mathcal{D}_{\text{in}}$ then we would also have $-\mathbf{x} \sim \mathcal{D}_{\text{in}}$.

Assumptions 2.2 : Finiteness of certain expectations

The following expectations under $\mathbf{x} \sim \mathcal{D}_{\text{in}}$ are are assumed to be finite,

$$\mathbb{E}_{\mathbf{x}}\Big[\|\mathbf{x}\|\Big] =: \mathrm{m}_1, \quad \mathbb{E}_{\mathbf{x}}\Big[\|\mathbf{x}\|^2\Big] =: \mathrm{m}_2, \quad \mathbb{E}_{\mathbf{x}}\Big[\|\mathbf{x}\|^3\Big] =: \mathrm{m}_3, \quad \mathbb{E}_{\mathbf{x}}\Big[\|\mathbf{x}\|^4\Big] =: \mathrm{m}_4$$

It follows that for a measurable function $\beta : \mathbb{R}^n \to [0,1]$ the Assumptions 2.2 in particular imply finiteness of the following expectations

$$\mathbb{E}_{\mathbf{x}}\Big[\beta(\mathbf{x})\|\mathbf{x}\|\Big] =: \beta_1, \mathbb{E}_{\mathbf{x}}\Big[\beta(\mathbf{x})\|\mathbf{x}\|^2\Big] =: \beta_2, \mathbb{E}_{\mathbf{x}}\Big[\beta(\mathbf{x})\|\mathbf{x}\|^3\Big] =: \beta_3, \mathbb{E}_{\mathbf{x}}\Big[\beta(\mathbf{x})\|\mathbf{x}\|^4\Big] =: \beta_4,$$

$\beta(\mathbf{x})$ will eventually be the bias of the coin that the adversarial oracle tosses to decide whether or not to attack the true label at $\mathbf{x}$. Note that in the above the adversarial oracle is free to design $\xi_{\mathbf{x}}$ however cleverly as a function of all the data seen so far and $f_{\mathbf{w}^*}$, while obeying the norm bound. For ease of notation we also define the following quantities,

$$\bar{\mathrm{A}} := \frac{1}{k}\sum_{i=1}^{k} \mathrm{A}_i, \ \ \Sigma := \mathrm{E}[\mathbf{x}\mathbf{x}^\top]$$

$$\lambda_1 := \lambda_{\min}(\bar{\mathrm{A}}\Sigma\mathbf{M}^T), \quad \lambda_2 := \sqrt{\lambda_{\max}(\mathbf{M}^T\mathbf{M})}, \quad \lambda_3 := \frac{1}{k}\sum_{i=1}^{k}\lambda_{\max}(\mathrm{A}_i\mathrm{A}_i^\top).$$

In Algorithm 1, we give a stochastic non-gradient algorithm, **Neuro-Tron** inspired by [16],[18],[12],

Algorithm 1 Neuro-Tron (mini-batched, multi-gate, single filter, stochastic)

```
**Input:** Sampling access to the marginal input distribution $\mathcal{D}_{\text{in}}$ on $\mathbb{R}^n$.
**Input:** Access to adversarially corrupted output $y \in \mathbb{R}$ when queried with $\mathbf{x} \in \mathbb{R}^n$
**Input:** Access to the output of any $f_{\mathbf{w}} \in \mathcal{F}_{k,\alpha,\mathcal{A},\mathcal{W}}$ for any $\mathbf{w}$ and input.
**Input:** A sensing matrix $\mathbf{M} \in \mathbb{R}^{r \times n}$ and an arbitrarily chosen starting point of $\mathbf{w}_1 \in \mathbb{R}^r$
**for** $t = 1, \ldots$ **do**
    Sample $\{\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_b}\} \sim \mathcal{D}_{\text{in}}$ and query the (adversarial) oracle with it.
    The Oracle samples $\forall i = 1, \ldots, b, \alpha_{t_i} \sim \{0,1\}$ with probability $\{1 - \beta(x_{t_i}), \beta(x_{t_i})\}$
    The Oracle replies back with $\forall i = 1, \ldots, b$, $y_{t_i} = \alpha_{t_i}\xi_{t_i} + f_{\mathbf{w}^*}(\mathbf{x}_{t_i})$
    Form the Tron-gradient,
        $$\mathbf{g}^{(t)} := \mathbf{M}\left(\frac{1}{b}\sum_{i=1}^{b}\left(\Big(y_{t_i} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i})\Big)\mathbf{x}_{t_i}\right)\right)$$
    $\mathbf{w}^{(t+1)} := \mathbf{w}^{(t)} + \eta\mathbf{g}^{(t)}$
**end for**
```

2 The Main Results

Theorem 1. *Suppose that* **Assumptions 2.1 and 2.2** *are satisfied and* $\lambda_1 > 0$.

Case I : Exactly realizable labels (i.e $\theta = 0$ and hence no noise.) *Suppose the oracle returns faithful output i.e it sends $\mathbf{y} = f_{\mathbf{w}^*}(\mathbf{x})$ when queried with $\mathbf{x}$. Then if we choose step size η in Algorithm 1 as,*

$$\eta = \frac{1}{\gamma} \cdot \frac{\lambda_1}{(1+\alpha)\lambda_2^2 \lambda_3 (\mathrm{m}_4/b + \mathrm{m}_2^2(1-1/b))}, \text{ where } \gamma > \max\Big\{C, 1\Big\}$$

with $C/\lambda_1^2 = (\lambda_2^2\lambda_3(\mathrm{m}_4/b + \mathrm{m}_2^2(1-1/b)))^{-1}$ for all desired accuracy parameter $\epsilon \geq 0$ and failure probability $\delta \geq 0$, for $\mathrm{T} = \mathcal{O}\Big(\log\Big(\frac{\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2}{\epsilon^2 \delta}\Big)\Big)$ we have that,

$$\|\mathbf{w}^{(\mathrm{T})} - \mathbf{w}^*\| \leq \epsilon \textit{ w.p. atleast } (1-\delta)$$

(Case II : Realizable labels additively corrupted by bounded probabilistic adversarial perturbation) *Define $\mathrm{c}_{\text{trade-off}} = \frac{(1+\alpha)\lambda_1}{\beta_1\lambda_2} - 1$. Suppose the distribution $\mathcal{D}_{\text{in}}$, matrix $\mathbf{M}$ in Algorithm 1, noise bound θ^*, target accuracy ϵ and target confidence $\delta > 0$ are such that,*

$$\theta^{*2} = \epsilon^2\delta \cdot \mathrm{c}_{\text{trade-off}} \textit{ and } \epsilon^2\delta < \|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \tag{1}$$

Then if we choose step size η in Algorithm 1 as,

$$\eta = \frac{1}{\gamma} \cdot \frac{\beta_1 \mathrm{c}_{\text{trade-off}}}{(1+\alpha)^2\lambda_2\lambda_3\Big((\beta_1\mathrm{m}_2 + \mathrm{m}_2^2)\left(1 - \frac{1}{b}\right) + \frac{\beta_3 + \mathrm{m}_4}{b}\Big)}$$

where

$$\gamma > \max\left\{\frac{(\beta_1 \mathrm{c}_{\text{trade-off}})^2}{(1+\alpha)^2\lambda_3((\beta_1\mathrm{m}_2 + \mathrm{m}_2^2)\left(1-\frac{1}{b}\right) + \frac{\beta_3+\mathrm{m}_4}{b})}, C_2\right\} > 1$$

with

$$C_2 = \frac{\epsilon^2\delta + \frac{\theta^2((\beta_1^2 + \beta_1\mathrm{m}_2)\left(1-\frac{1}{b}\right) + \frac{\beta_2+\beta_3}{b})}{(1+\alpha)^2\lambda_3(\beta_1\mathrm{m}_2 + \mathrm{m}_2^2)\left(1-\frac{1}{b}\right) + \frac{\beta_3+\mathrm{m}_4}{b}}}{\epsilon^2\delta - \frac{\theta^2}{\mathrm{c}_{\text{trade-off}}}}.$$

Then for,

$$\mathrm{T} = \mathcal{O}\left(\log\left[\frac{\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2}{\epsilon^2\delta - \frac{\theta^2}{\mathrm{c}_{\text{rate}}}}\right]\right), \quad \textit{where } \mathrm{c}_{\text{rate}} = \frac{\gamma - 1}{\frac{\mathrm{m}_2 + \mathrm{m}_3}{(1+\alpha)^2\lambda_3(\mathrm{m}_3 + \mathrm{m}_4)} + \frac{\gamma}{\mathrm{c}_{\text{trade-off}}}}$$

we have,

$$\|\mathbf{w}^{(\mathrm{T})} - \mathbf{w}^*\| \leq \epsilon \textit{ w.p. atleast } (1-\delta)$$

Remark 1. **(a)** Note that, η is an increasing function of b in both cases, the batch-size of the sample. So increasing b would mean faster convergence.

Remark 2. **(b)** If $r \leq n$ and $\bar{\mathrm{A}}\Sigma$ is full rank i.e rank r then in the above we can always chosen $\mathbf{M} = \bar{\mathrm{A}}\Sigma$ Also if Σ is PD then $\mathbf{M} = \bar{\mathrm{A}}$ is also a valid choice. **(b)** It can be easily seen that the term occurring in the expression of T above, "$\epsilon^2\delta - \frac{\theta^2}{c_{\text{rate}}}$" is positive because of the lowerbound imposed on the parameter γ **(c)** In subsection 2.2, we will show that this worst-case trade-off is nearly optimal in the worst-case i.e when the adversary attacks every sampled data.

To develop some feel for the constraint imposed by equation 1 in the main theorem we look at a situation where the input data is being sampled from a Gaussian distribution and the net class being trained over is $\mathcal{F}_{1,0,\{\mathbf{I}_{n\times n}\},\mathbb{R}^n}$, a single ReLU gate neural networks.

Lemma 1. *(Provable training for Gaussian distributions and single-ReLU gate with an adversarial oracle). Suppose $\mathcal{D}_{\text{in}} = \mathcal{N}(0, \sigma^2\mathbf{I}_{n\times n})$ and $\mathcal{H} = \mathcal{F}_{1,0,\{\mathbf{I}_{n\times n}\},\mathbb{R}^n}$. For the choice of $\mathbf{M} = \mathbf{I}_{n\times n}$ in Algorithm 1, the constraint $\mathrm{c}_{\text{trade-off}}$ in Theorem 1 can be written as,*

$$\mathrm{c}_{\text{trade-off}} = \frac{\theta^{*2}}{\epsilon^2\delta} = \frac{\sigma}{\sqrt{2}\beta} \cdot \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\right] - 1 \tag{2}$$

where we assume $\beta(\mathbf{x}) = \beta$ for all $\mathbf{x}$ with some $0 < \beta < 1$.

Proof. Here $\lambda_1 = \sigma^2$, $\lambda_2 = \lambda_3 = 1$ and $\alpha = 0$. We invoke standard results about the Gaussian distribution to see,

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\sigma^2 I)}\left[\|\mathbf{x}\|^k\right] = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I)}\left[\|\sigma\mathbf{x}\|^k\right] = \sigma^k\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I)}\left[\|\mathbf{x}\|^k\right] = \sigma^k 2^{k/2}\frac{\Gamma\left(\frac{n+k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Hence we have,

$$\beta_1 = \beta\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_{n\times n})}\left[\|\mathbf{x}\|\right] = \sqrt{2}\sigma\beta \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \tag{3}$$

Invoking the above the $\mathrm{c}_{\text{trade-off}}$ in Theorem 1 simplifies to the expression in equation 2. □

Hence we conclude that if we want to defend against an adversary with a fixed corruption budget of θ^* with a desired accuracy of ϵ and failure probability of δ then a sufficient condition (a *safe* data distribution) is if the data distribution is $\mathcal{N}(0, \sigma^2\mathbf{I})$ with σ^2 being an increasing function of the data dimension n in an appropriate way s.t the the RHS of equation 2 remains fixed.

2.1 Understanding the risk of the learnt predictor

Recalling that the predictor we are training is $f_{\mathbf{w}} : \mathbb{R}^n \ni \mathbf{x} \mapsto \frac{1}{k}\sum_{i=1}^{k} \sigma\Big(\mathbf{w}^\top \mathrm{A}_i \mathbf{x}\Big) \in \mathbb{R}$, we can write the prediction risk of the learnt predictor at time T as,

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}\Big[\left(f_{\mathbf{w}^*}(\mathbf{x}) - f_{\mathbf{w}^{\mathrm{T}}}(\mathbf{x})\right)^2 \Big] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}\left[\left(\frac{1}{k}\sum_{i=1}^{k} \Big\{ \sigma\Big(\mathbf{w}^{*\top}\mathrm{A}_i\mathbf{x}\Big) - \sigma\Big(\mathbf{w}^{(\mathrm{T})\top}\mathrm{A}_i\mathbf{x}\Big) \Big\} \right)^2 \right]$$

Now we can ask as to whether the above can be below θ^{*2} when Theorem 1 (Case II) guarantees,

$$\|\mathbf{w}^{(\mathrm{T})} - \mathbf{w}^*\|^2 \leq \epsilon^2 = \frac{\theta^{*2}}{\delta \cdot \mathrm{c}_{\text{trade-off}}}$$

where we recall that we have, $\mathrm{c}_{\text{trade-off}} = (-1 + \frac{(1+\alpha)\lambda_1}{\beta_1\lambda_2})$. From Lemma 4 we note that this will be true at iteration T if we can ensure,

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}\left[(1+\alpha)^2 \left(\frac{1}{k}\sum_{i=1}^{k} \lambda_{\max}(\mathrm{A}_i\mathrm{A}_i^\top)\|\mathbf{x}\|^2 \right) \right] \\ &= \left(\frac{(1+\alpha)^2}{k} \cdot \sum_{i=1}^{k} \lambda_{\max}(\mathrm{A}_i\mathrm{A}_i^\top) \right) \cdot \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}\Big[\|\mathbf{x}\|^2 \Big] \cdot \frac{1}{\delta\left(\frac{(1+\alpha)\lambda_1}{\beta_1\lambda_2} - 1 \right)} < 1 \end{aligned}$$

It is easy to demonstrate cases when the above is true : for example consider the situation when one is trying to train a single ReLU gate and $\mathcal{D}_{\text{in}} = \mathcal{N}(0, \mathbf{I}_{n\times n})$ and we choose $\mathbf{M} = \mathbf{I}_{n\times n}$. This corresponds to $\lambda_1 = 1$, $\lambda_2 = 1$ and $\alpha = 0$. Further suppose that the attack probability is $\beta(\mathbf{x}) = \beta$ for all $\mathbf{x}$ for some $\beta \in (0,1)$.

Then the above condition is equivalent to,

$$n = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[\|x\|^2] \leq \delta\left(\frac{1}{\beta_1} - 1\right) \text{ or } \beta_1 \leq \frac{1}{1+n/\delta}.$$

The above inequality on β_1 alongwith (3) gives the dependence on attack probability as

$$\beta < \frac{1}{\sqrt{2}}\left[\frac{\mathit{\Gamma}\left(\frac{n}{2}\right)}{\mathit{\Gamma}\left(\frac{n+1}{2}\right)} \right] \frac{1}{1+n/\delta}.$$

The above equation thus tells us that for learning a ReLU gate with normally distributed data, there is a dimension dependent upperbound on the attacker's probability s.t *post-training* the learnt weights will have better prediction accuracy on an average than the worst distortion the attacker could have made to any particular label.

2.2 Demonstrating the near optimality in the worst case, of the guarantees of Theorem 1

We recall that Case I of Theorem 1 shows that Algorithm 1 recovers the true filter $\mathbf{w}^*$ when it has access to clean/exactly realizable data. For a given true filter $\mathbf{w}^* \in \mathbb{R}^r$ consider another value for the filter $\mathbb{R}^r \ni \mathbf{w}_{\text{adv}} \neq \mathbf{w}^*$ and suppose that $\theta^* = \zeta$ for some $\zeta \geq \sup_{\mathbf{x} \in \text{supp}(\mathcal{D}_{\text{in}})} \vert f_{\mathbf{w}_{\text{adv}}}(x) - f_{\mathbf{w}^*}(x) \vert$. It is easy to imagine cases where the supremum in the RHS exists like when $\mathcal{D}_{\text{in}}$ is compactly supported. Now in this situation equation 1 says that $\epsilon^2 = \frac{\theta^{*2}}{\delta c_{\text{trade-off}}} \implies \epsilon^2 \geq \frac{\zeta^2}{c_{\text{trade-off}}}$, Hence proving optimality of the guarantee is equivalent to showing the existence of an attack within this ζ bound for which the best accuracy possible nearly saturates this lowerbound.

Now note that this choice of θ^* allows for the adversarial oracle $\mathbf{O}_{\text{A}}$ to be such that when queried at $\mathbf{x}$ it replies back with $\xi_{\mathbf{x}} + f_{\mathbf{w}_*}(\mathbf{x})$ where $\xi_{\mathbf{x}} = f_{\mathbf{w}_{\text{adv}}}(\mathbf{x}) - f_{\mathbf{w}_*}(\mathbf{x})$. Hence the data the algorithm receives will be such that it can be exactly realized with the filter choice being $\mathbf{w}_{\text{adv}}$. Hence the realizable case analysis of Theorem 1 will apply showing that the algorithm's iterates are converging in high probability to $\mathbf{w}_{\text{adv}}$. Hence the error incurred is such that $\epsilon \geq \Vert \mathbf{w}_{\text{adv}} - \mathbf{w}_* \Vert$.

Now consider an instantiation of the above attack happening with $\zeta = r \Vert \mathbf{w}_{\text{adv}} - \mathbf{w}_* \Vert$ for $r = \sup_{\mathbf{x} \in \text{supp}(\mathcal{D}_{\text{in}})} \Vert \mathbf{x} \Vert$ and f being a single ReLU gate i.e $\mathbb{R}^n \ni \mathbf{x} \mapsto f_{\mathbf{w}_*}(\mathbf{x}) = \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}) \in \mathbb{R}$. Its easy to imagine cases where $\mathcal{D}_{\text{in}}$ is such that r is finite and it also satisfies **Assumptions 1.1 and 1.2**. Further, this choice of ζ is valid since the following holds,

$$\begin{aligned} \sup_{\mathbf{x} \in \text{supp}(\mathcal{D}_{\text{in}})} \vert f_{\mathbf{w}_{\text{adv}}}(x) - f_{\mathbf{w}^*}(x) \vert &= \sup_{\mathbf{x} \in \text{supp}(\mathcal{D}_{\text{in}})} \vert \text{ReLU}(\mathbf{w}_{\text{adv}}^\top \mathbf{x}) - \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}) \vert \\ &\leq r \Vert \mathbf{w}_{\text{adv}} - \mathbf{w}_* \Vert = \zeta \end{aligned}$$

Thus the above setup invoked on training a ReLU gate with inputs being sampled from $\mathcal{D}_{\text{in}}$ as above while the labels are being additively corrupted by at most $\zeta = r \Vert \mathbf{w}_{\text{adv}} - \mathbf{w}_* \Vert$ demonstrates a case where the *worst case (i.e* $\beta(\mathbf{x}) = 1$ *identically)* accuracy guarantee of $\epsilon^2 \geq \frac{\zeta^2}{c_{\text{trade-off}}}$ is optimal upto a constant $\frac{r^2}{c_{\text{trade-off}}}$. We note that this argument also implies the near optimality of equation 1 for *any* algorithm defending against this attack which also has the property of recovering the parameters correctly when the labels are exactly realizable.

3 Proof of Theorem 1

Proof. Between consecutive iterates of the algorithm we have,

$$\begin{aligned} \Vert \mathbf{w}^{(t+1)} - \mathbf{w}^* \Vert^2 &= \Vert \mathbf{w}^{(t)} + \eta \mathbf{g}^{(t)} - \mathbf{w}^* \Vert^2 \\ &= \Vert \mathbf{w}^{(t)} - \mathbf{w}^* \Vert^2 + \eta^2 \Vert \mathbf{g}^{(t)} \Vert^2 + 2\eta \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{g}^{(t)} \rangle \end{aligned}$$

Let the training data sampled till the iterate t be $S_t := \bigcup_{i=1}^{t} s_i$. We overload the notation to also denote by S_t, the sigma-algebra generated by the samples seen *and the* αs till the t-th iteration. Conditioned on S_{t-1} , $\mathbf{w}_t$ is determined and g_t is random and dependent on the choice of s_t and $\{\alpha_{t_i}, \xi_{t_i} \mid i = 1, \dots, b\}$. We shall denote the collection of random variables $\{\alpha_{t_i} \mid i = 1, \dots, b\}$ as α_t. Then taking conditional expectations w.r.t S_{t-1} of both sides of the above equation we have,

$$\begin{aligned}
&\mathbb{E}_{s_t,\alpha_t}\left[\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 \middle| S_{t-1}\right] \\
&= \underbrace{2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \mathbb{E}_{\mathbf{x}_{t_i},\alpha_{t_i}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{M}\Big(y_{t_i} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i})\Big)\mathbf{x}_{t_i}\right\rangle \middle| S_{t-1}\right]}_{\text{Term 1}} \\
&+ \underbrace{\eta^2 \mathbb{E}_{\mathbf{x}_{t_i},\alpha_{t_i}}\left[\|\mathbf{g}^{(t)}\|^2 \middle| S_{t-1}\right]}_{\text{Term 2}} + \mathbb{E}_{s_t,\alpha_t}\left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \middle| S_{t-1}\right] \qquad (4)
\end{aligned}$$

We provide the bound for Term 1 in the Appendix A.1 [4] and arrive at

$$\begin{aligned}
&\text{Term 1} \qquad (5) \\
&\leq -\eta(1+\alpha) \cdot \lambda_1 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + 2\eta\theta_*\lambda_2 \cdot \mathbb{E}\left[\beta(\mathbf{x}_{t_1})\|\mathbf{x}_{t_1}\| \middle| S_{t-1}\right] \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|.
\end{aligned}$$

Now we split the Term 2 in the RHS of equation 4 as follows:

$$\begin{aligned}
&\mathbb{E}\left[\|\eta\mathbf{g}^{(t)}\|^2 \middle| S_{t-1}\right] \\
&= \frac{\eta^2}{b^2}\Bigg(\mathbb{E}\left[\sum_{i=1}^{b}(y_{t_i} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i}))^2 \cdot \|\mathbf{M}\mathbf{x}_{t_i}\|^2 \middle| S_{t-1}\right] \\
&\quad + \mathbb{E}\left[\sum_{i=1}^{b}\sum_{j=1, j\neq i}^{b}(y_{t_i} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i}))(y_{t_j} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_j})) \cdot \mathbf{x}_{t_j}^\top \mathbf{M}^\top \mathbf{M}\mathbf{x}_{t_i} \middle| S_{t-1}\right] \\
&=: \text{Term 21} + \text{Term 22} \qquad (6)
\end{aligned}$$

[4] For restriction on space we post the appendix to https://sayarkarmakar.github.io/publications/guaranteespapersupp.pdf

We separately upperbound the Term 21 and Term 22 as outlined in the Appendix A.2 and A.3 [5] respectively and arrive at,

$$\text{Term } 21 \leq \tfrac{\eta^2\lambda_2^2}{b}\left(c^2\mathrm{m}_4\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2 + 2c\theta_*\beta_3\|\mathbf{w}^{(t)}-\mathbf{w}^*\| + \theta_*^2\beta_2\right) \tag{7}$$

$$\begin{aligned}&\text{Term } 22 \\ &\leq \frac{\eta^2(b^2-b)}{b^2}\left[\theta_*^2\lambda_2^2\beta_1^2 + 2\theta_*\lambda_2^2\beta_1 c\mathrm{m}_2\|\mathbf{w}^{(t)}-\mathbf{w}^*\| + \lambda_2^2c^2\mathrm{m}_2^2\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2\right].\end{aligned} \tag{8}$$

Next we take total expectations of both sides of equations 5, 7 and 8 recalling that the conditional expectation of functions of $\mathbf{x}_{t_i}$ w.r.t. S_{t-1} are random variables which are independent of the powers of $\|\mathbf{w}^{(t)}-\mathbf{w}^*\|$. Then we substitute the resulting expressions into the RHS of equation 4 to get,

$$\begin{aligned}&\mathbb{E}\Big[\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|^2\Big] \\ &\quad\leq \left[1+\eta^2\lambda_2^2c^2\left(\mathrm{m}_2^2\left(1-\frac{1}{b}\right)+\frac{\mathrm{m}_4}{b}\right)-\eta\lambda_1(1+\alpha)\right]\cdot\mathbb{E}\Big[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2\Big] \\ &\quad+\left[2\eta^2\lambda_2^2c\theta_*\left(\beta_1\mathrm{m}_2\left(1-\frac{1}{b}\right)+\frac{\beta_3}{b}\right)+2\eta\lambda_2\cdot\beta_1\theta_*\right]\cdot\mathbb{E}\Big[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|\Big] \\ &\quad+\eta^2\theta_*^2\lambda_2^2\left(\beta_1^2\left(1-\frac{1}{b}\right)+\frac{\beta_2}{b}\right).\end{aligned} \tag{9}$$

Case I : Realizable, $\theta_* = 0$.

Here the recursion above simplifies to,

$$\begin{aligned}&\mathbb{E}\Big[\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|^2\Big] \\ &\quad\leq \left[1+\eta^2\lambda_2^2c^2\left(\mathrm{m}_2^2\left(1-\frac{1}{b}\right)+\frac{m_4}{b}\right)-\eta\lambda_1(1+\alpha)\right]\cdot\mathbb{E}\Big[\|\mathbf{w}_t-\mathbf{w}^*\|^2\Big]\end{aligned} \tag{10}$$

Let $\kappa = 1+\eta^2\lambda_2^2c^2\left(\mathrm{m}_2^2(1-1/b)+\mathrm{m}_4/b\right)-\eta\lambda_1(1+\alpha)$. Thus, for all $t\in\mathbb{Z}^+$,

$$\mathbb{E}\Big[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2\Big] \leq \kappa^{t-1}\mathbb{E}\Big[\|\mathbf{w}^{(1)}-\mathbf{w}^*\|^2\Big].$$

[5] For restriction on space we post the appendix to https://sayarkarmakar.github.io/publications/guaranteespapersupp.pdf

Recalling that $c^2 = (1+\alpha)\lambda_3$, we can verify that the choice of step size given the Theorem is $\eta = \frac{1}{\gamma} \cdot \frac{\lambda_1(1+\alpha)}{\lambda_2^2 c^2 (\mathrm{m}_4/b + \mathrm{m}_2^2(1-1/b))}$ and the assumption on γ ensures that for this η, $\kappa = 1 - \frac{\gamma-1}{\gamma^2} \frac{\lambda_1^2}{\lambda_2^2 \lambda_3 (\mathrm{m}_4/b + \mathrm{m}_2^2(1-1/b))} \in (0,1)$. Therefore, for $\mathrm{T} = \mathcal{O}\Big(\log\Big(\frac{\|\mathbf{w}^{(1)} - \mathbf{w}^*\|}{\epsilon^2 \delta} \Big) \Big)$, we have

$$\mathbb{E}\Big[\|\mathbf{w}^{(\mathrm{T})} - \mathbf{w}^*\|^2 \Big] \leq \epsilon^2 \delta$$

The conclusion now follows from Markov's inequality.

Case II : Realizable + Adversarial Noise, $\theta^* > 0$.

Note that the linear term $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|$ in equation 9 is a unique complication that is introduced here because of the absence of distributional assumptions on the noise in the labels. We can now upperbound the linear term using the AM-GM inequality as follows, which also helps decouple the adversarial noise terms from the distance to the optima. The equation 9 becomes,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 \right] \\
&\quad \leq \left[\eta^2 \lambda_2^2 c^2 \left((\beta_1 \mathrm{m}_2 + \mathrm{m}_2^2) \left(1 - \frac{1}{b}\right) + \frac{\beta_3 + \mathrm{m}_4}{b} \right) - \eta \lambda_2 \left(\frac{\lambda_1(1+\alpha)}{\lambda_2} - \beta_1 \right) + 1 \right] \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 \right] \\
&\quad + \theta_*^2 \Big(\eta^2 \lambda_2^2 \left((\beta_1^2 + \beta_1 \mathrm{m}_2) \left(1 - \frac{1}{b}\right) + \frac{\beta_2 + \beta_3}{b} \right) + \eta \lambda_2 \beta_1 \Big)
\end{aligned}$$

Let us define $\mathit{\Delta}_t = \mathbb{E}\Big[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \Big]$, $\eta' := \eta \lambda_2$, $b_* := \frac{\lambda_1(1+\alpha)}{\lambda_2} - \beta_1$, $c_1 = c^2 \left((\beta_1 \mathrm{m}_2 + \mathrm{m}_2^2)\left(1 - \frac{1}{b}\right) + \frac{\beta_3 + \mathrm{m}_4}{b} \right)$, $c_2 := \theta_*^2 \cdot \Big(\eta^2 \lambda_2^2 \Big((\beta_1^2 + \beta_1 \mathrm{m}_2)\left(1 - \frac{1}{b}\right) + \frac{\beta_2 + \beta_3}{b} \Big)$ and $c_3 = \theta_*^2 \cdot \beta_1$. Then the dynamics of the algorithm is given by,

$$\mathit{\Delta}_{t+1} \leq (1 - \eta' b_* + \eta'^2 c_1) \mathit{\Delta}_t + \eta'^2 c_2 + \eta' c_3.$$

We note that the above is of the same form as lemma 2 in the Appendix [6] with $\mathit{\Delta}_1 = \|\mathbf{w}_1 - \mathbf{w}^*\|^2$. We invoke the lemma with $\epsilon'^2 := \epsilon^2 \delta$ s.t equation 1 holds. This ensures that, $\frac{c_3}{b_*} = \frac{\theta_*^2}{\mathrm{c}_{\text{trade-off}}} = \epsilon^2 \delta < \mathit{\Delta}_1$ as required by lemma 2. The chosen value of η in the Theorem follows from the sufficient condition specified for η' in the lemma 2.

Recalling the definition of c_{rate} as given in the Theorem statement we can see that $\frac{\frac{c_2}{c_1} + \gamma \cdot \frac{c_3}{b_*}}{\gamma - 1} = \frac{\theta^2}{\mathrm{c}_{\text{rate}}}$ and hence we can read off from lemma 2 that at the value of T as specified in the Theorem statement we have from lemma 2 that,

[6] For restriction on space we post the appendix to https://sayarkarmakar.github.io/publications/guaranteespapersupp.pdf

$$\Delta_{\mathrm{T}} = \mathbb{E}\Big[\|\mathbf{w}_{\mathrm{T}} - \mathbf{w}^*\|^2\Big] \leq \epsilon^2\delta$$

and the needed high probability guarantee follows by Markov inequality. □

4 Conclusion

To the best of our knowledge in this paper, we have provided the first demonstration of a class of provably robustly learnable finitely large neural networks i.e along with the neural network class, we have given a class of adversarial oracles supplying additively corrupted labels and a corresponding stochastic algorithm which up to a certain accuracy and confidence performs supervised learning on our network in the presence of this malicious oracle corrupting the realizable true labels. We have also established as to why our guarantees are nearly optimal in the worst case.

There are a number of exciting open questions that now open up from here. Firstly it remains to broaden the scope of such proofs to more complicated neural networks and to more creative adversaries which can say do more difficult distortions to the labels or corrupt even the samples from the data distribution $\mathcal{D}_{\text{in}}$. Secondly, even while staying within the setup of Theorem 1, it would be interesting to be able to characterize the information-theoretic limits of accuracy and confidence trade-offs as a function of the adversary's probability profile of attacking (as captured by the β function).

Acknowledgements

We would like to thank Amitabh Basu for extensive discussions on various parts of this paper. The second author would like to thank the MINDS Data Science Fellowship of JHU for supporting this work. The second author would also like to acknowledge the extensive discussions on this topic with Anup Rao, Sridhar Mahadevan, Pan Xu and Wenlong Mou when he was interning at Adobe, San Jose during the summer of 2019.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–14430 (2018)
2. Barreno, M., Nelson, B., Joseph, A.D., Tygar, J.D.: The security of machine learning. Machine Learning **81**(2), 121–148 (2010)
3. Behzadan, V., Munir, A.: Vulnerability of deep reinforcement learning to policy induction attacks. In: International Conference on Machine Learning and Data Mining in Pattern Recognition, pp. 262–275. Springer (2017)
4. Bubeck, S., Price, E., Razenshteyn, I.: Adversarial examples from computational constraints. arXiv preprint arXiv:1805.10204 (2018)

5. Chen, Y., Caramanis, C., Mannor, S.: Robust sparse regression under adversarial corruption. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13, p. III–774–III–782. JMLR.org (2013)
6. Degwekar, A., Nakkiran, P., Vaikuntanathan, V.: Computational limitations in robust classification and win-win results. arXiv preprint arXiv:1902.01086 (2019)
7. Dou, Z., Osher, S.J., Wang, B.: Mathematical analysis of adversarial attacks. arXiv preprint arXiv:1811.06492 (2018)
8. Feng, J., Xu, H., Mannor, S., Yan, S.: Robust logistic regression and classification. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14, p. 253–261. MIT Press, Cambridge, MA, USA (2014)
9. Freund, Y., Schapire, R.E.: Large margin classification using the perceptron algorithm. Machine learning **37**(3), 277–296 (1999)
10. Gao, R., Cai, T., Li, H., Hsieh, C.J., Wang, L., Lee, J.D.: Convergence of adversarial training in overparametrized neural networks. In: Advances in Neural Information Processing Systems, pp. 13009–13020 (2019)
11. Goel, S., Klivans, A.: Learning depth-three neural networks in polynomial time. arXiv preprint arXiv:1709.06010 (2017)
12. Goel, S., Klivans, A., Meka, R.: Learning one convolutional layer with overlapping patches. arXiv preprint arXiv:1802.02547 (2018)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Huang, S., Papernot, N., Goodfellow, I., Duan, Y., Abbeel, P.: Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284 (2017)
15. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B.: Manipulating machine learning: Poisoning attacks and countermeasures for regression learning (2018)
16. Kakade, S.M., Kanade, V., Shamir, O., Kalai, A.: Efficient learning of generalized linear and single index models with isotonic regression. In: Advances in Neural Information Processing Systems, pp. 927–935 (2011)
17. Khim, J., Loh, P.L.: Adversarial risk bounds via function transformation. arXiv preprint arXiv:1810.09519 (2018)
18. Klivans, A., Meka, R.: Learning graphical models using multiplicative weights. In: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pp. 343–354. IEEE (2017)
19. Koh, P.W., Steinhardt, J., Liang, P.: Stronger data poisoning attacks break data sanitization defenses (2018)
20. Laska, J.N., Davenport, M.A., Baraniuk, R.G.: Exact signal recovery from sparsely corrupted measurements through the pursuit of justice
21. Li, M., Soltanolkotabi, M., Oymak, S.: Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In: International Conference on Artificial Intelligence and Statistics, pp. 4313–4324. PMLR (2020)
22. Li, X.: Compressed sensing and matrix completion with constant proportion of corruptions (2011)
23. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for improving transferability of adversarial examples. arXiv preprint arXiv:1908.06281 (2019)

24. Liu, C., Li, B., Vorobeychik, Y., Oprea, A.: Robust linear regression against training data poisoning. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, p. 91–102. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3128572.3140447. URL https://doi.org/10.1145/3128572.3140447
25. Montasser, O., Hanneke, S., Srebro, N.: Vc classes are adversarially robustly learnable, but only improperly. arXiv preprint arXiv:1902.04217 (2019)
26. Nguyen, N.H., Tran, T.D.: Exact recoverability from dense corrupted observations via l_1 minimization (2011)
27. Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. IEEE transactions on neural networks **3 5**, 683–97 (1992)
28. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506–519 (2017)
29. Qiu, S., Liu, Q., Zhou, S., Wu, C.: Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences (2076-3417) **9**(5) (2019)
30. Raghunathan, A., Steinhardt, J., Liang, P.: Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344 (2018)
31. Raghunathan, A., Steinhardt, J., Liang, P.S.: Semidefinite relaxations for certifying robustness to adversarial examples. In: Advances in Neural Information Processing Systems, pp. 10877–10887 (2018)
32. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)
33. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: Advances in Neural Information Processing Systems, pp. 5014–5026 (2018)
34. Song, C., He, K., Wang, L., Hopcroft, J.E.: Improving the generalization of adversarial training with domain adaptation. arXiv preprint arXiv:1810.00740 (2018)
35. Steinhardt, J., Koh, P.W.W., Liang, P.S.: Certified defenses for data poisoning attacks. In: Advances in neural information processing systems, pp. 3517–3529 (2017)
36. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
37. Wang, Y., Chaudhuri, K.: Data poisoning attacks against online learning (2018)
38. Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., Roli, F.: Is feature selection secure against training data poisoning? (2018)
39. Xu, H., Caramanis, C., Mannor, S.: Robust regression and lasso (2008)
40. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
41. Yin, D., Ramchandran, K., Bartlett, P.: Rademacher complexity for adversarially robust generalization. arXiv preprint arXiv:1810.11914 (2018)
42. Zhang, X., Zhu, X., Lessard, L.: Online data poisoning attack (2019)

A Bounds needed in the proof in Section 3

In the following sub-sections we provide the upperbounds for Term 1, Term 21 and Term 22 from the main text.

A.1 Upperbound for Term 1

For Term 1 in equation (4) we proceed by observing that conditioned on S_{t-1}, $\mathbf{w}^{(t)}$ is determined while $\mathbf{w}^{(t+1)}$ and $\mathbf{g}^{(t)}$ are random. Thus we compute the following conditional expectation (suppressing the subscripts of $\mathbf{x}_{t_i}, \alpha_{t_i}$),

$$\begin{align}
\text{Term 1} &= \mathbb{E}\left[2\eta \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{g}^{(t)} \rangle \middle| S_{t-1} \right] \notag \\
&= 2\frac{\eta}{b} \sum_{i=1}^{b} \mathbb{E}\left[\Big(f_{\mathbf{w}^*}(\mathbf{x}_{t_i}) + \alpha_{t_i}\xi_{t_i} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i}) \Big) (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M}\mathbf{x}_{t_i} \middle| S_{t-1} \right] \notag \\
&= 2\frac{\eta}{b} \sum_{i=1}^{b} \mathbb{E}\left[\Big(f_{\mathbf{w}^*}(\mathbf{x}_{t_i}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i}) \Big) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M}\mathbf{x}_{t_i} \middle| S_{t-1} \right] \notag \\
&\quad + 2\frac{\eta}{b} \sum_{i=1}^{b} \mathbb{E}\left[\alpha_{t_i}\xi_{t_i} (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M}\mathbf{x}_{t_i} \middle| S_{t-1} \right] \tag{11} \\
&\leq \frac{-2\eta}{bk} \sum_{i=1}^{b} \sum_{j=1}^{k} \mathbb{E}\left[\Big(\sigma({\mathbf{w}^{(t)}}^\top \mathrm{A}_j \mathbf{x}_{t_i}) - \sigma(\mathbf{w}^{*\top} \mathrm{A}_j \mathbf{x}_{t_i}) \Big) (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M}\mathbf{x}_{t_i} \middle| S_{t-1} \right] \notag \\
&\quad + 2\frac{\eta\theta_*}{b} \sum_{i=1}^{b} \mathbb{E}\left[\beta(\mathbf{x}_{t_i}) \cdot |(\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M}\mathbf{x}_{t_i}| \middle| S_{t-1} \right] \tag{12}
\end{align}$$

We simplify the first term above by recalling an identity proven in [12], which we have reproduced here as Lemma 3 . Thus we get,

$$
\begin{aligned}
&\mathbb{E}\left[2\eta\langle \mathbf{w}^{(t)}-\mathbf{w}^*,\mathbf{g}^{(t)}\rangle \middle| S_{t-1}\right]\\
&\le \frac{-\eta(1+\alpha)}{bk}\sum_{i=1}^{b}\sum_{j=1}^{k}\mathbb{E}\left[(\mathbf{w}^{(t)}-\mathbf{w}^*)^\top \mathrm{A}_j\mathbf{x}_{t_i}(\mathbf{w}^{(t)}-\mathbf{w}^*)^\top\mathbf{M}\mathbf{x}_{t_i}\middle| S_{t-1}\right]\\
&\qquad + 2\frac{\eta\theta_*}{b}\sum_{i=1}^{b}\|\mathbf{w}^{(t)}-\mathbf{w}^*\|\cdot\mathbb{E}\left[\beta(\mathbf{x}_{t_i})\|\mathbf{M}\mathbf{x}_{t_i}\|\middle| S_{t-1}\right]\\
&\le -\eta(1+\alpha)(\mathbf{w}^{(t)}-\mathbf{w}^*)^\top\ \bar{\mathrm{A}}\ \mathbb{E}\left[\mathbf{x}_{t_1}\mathbf{x}_{t_1}^\top\middle| S_{t-1}\right]\ \mathbf{M}^\top\ (\mathbf{w}^{(t)}-\mathbf{w}^*)\\
&\qquad + 2\eta\theta_*\|\mathbf{w}^{(t)}-\mathbf{w}^*\|\sqrt{\lambda_{\max}(\mathbf{M}^\top\mathbf{M})}\cdot\mathbb{E}\left[\beta(\mathbf{x}_{t_1})\|\mathbf{x}_{t_1}\|\middle| S_{t-1}\right]\\
&\le -\eta(1+\alpha)\cdot\lambda_{\min}\left(\bar{\mathrm{A}}\mathbb{E}\left[\mathbf{x}_{t_1}\mathbf{x}_{t_1}^\top\middle| S_{t-1}\right]\mathbf{M}^\top\right)\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2\\
&\qquad + 2\eta\theta_*\cdot\mathbb{E}\left[\beta(\mathbf{x}_{t_1})\|\mathbf{x}_{t_1}\|\middle| S_{t-1}\right]\cdot\sqrt{\lambda_{\max}(\mathbf{M}^T\mathbf{M})}\|\mathbf{w}^{(t)}-\mathbf{w}^*\|\\
&\le -\eta(1+\alpha)\cdot\lambda_1\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2\\
&\qquad + 2\eta\theta_*\lambda_2\cdot\mathbb{E}\left[\beta(\mathbf{x}_{t_1})\|\mathbf{x}_{t_1}\|\middle| S_{t-1}\right]\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^*\|
\end{aligned}
$$

We have invoked the i.i.d nature of the data samples to invoke the definition of the λ_1 in above.

A.2 Upperbound for Term 21

For Term 21 in equation (6) we get,

$$
\begin{aligned}
&\text{Term 21} \le \frac{\eta^2\lambda_2^2}{b}\cdot\mathbb{E}\left[\left(f_{\mathbf{w}^*}(\mathbf{x}_{t_1})+\alpha_{t_1}\xi_{t_1}-f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})\right)^2\cdot\|\mathbf{x}_{t_1}\|^2\middle| S_{t-1}\right]\\
&\le \frac{\eta^2\lambda_2^2}{b}\cdot\mathbb{E}\left[\left(\left(f_{\mathbf{w}^*}(\mathbf{x}_{t_1})-f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})\right)^2+2\alpha_{t_1}\xi_{t_1}\left(f_{\mathbf{w}^*}(\mathbf{x}_{t_1})-f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})\right)+\alpha_{t_1}^2\xi_{t_1}^2\right)\cdot\|\mathbf{x}_{t_1}\|^2\middle| S_{t-1}\right]\\
&\le \frac{\eta^2\lambda_2^2c^2}{b}\mathbb{E}\left[\|\mathbf{x}_{t_1}\|^4\middle| S_{t-1}\right]\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2+\frac{2\eta^2\lambda_2^2c\theta_*}{b}\mathbb{E}\left[\beta(\mathbf{x}_{t_1})\|\mathbf{x}_{t_1}\|^3\middle| S_{t-1}\right]\|\mathbf{w}^{(t)}-\mathbf{w}^*\|\\
&\qquad + \frac{\eta^2\lambda_2^2\theta_*^2}{b}\mathbb{E}\left[\beta(\mathbf{x}_{t_1})\|\mathbf{x}_{t_1}\|^2\middle| S_{t-1}\right]\\
&= \frac{\eta^2\lambda_2^2}{b}\left(c^2\mathrm{m}_4\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2+2c\theta_*\beta_3\|\mathbf{w}^{(t)}-\mathbf{w}^*\|+\theta_*^2\beta_2\right)
\end{aligned}
$$

In the above lines we have invoked lemma 4 twice to upperbound the term, $|\big(f_{\mathbf{w}^*}(\mathbf{x}^{(t)}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}^{(t)})\big)|$ and we have defined,

$$c^2 := (1+\alpha)^2\lambda_3 = \frac{(1+\alpha)^2}{k}\Big(\sum_{i=1}^{k}\lambda_{\max}(\mathrm{A}_i\mathrm{A}_i^\top)\Big).$$

Next we proceed with Term 22 keeping in mind the independence of x_{t_i} and x_{t_j} for $i \neq j$,

A.3 Upperbound for Term 22

For Term 22 in equation (6) we get,

$$\begin{aligned}
&\text{Term 22} \\
&= \frac{\eta^2(b^2-b)}{b^2}\mathbb{E}\left[(\alpha_{t_1}\xi_{t_1} + f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1}))(\alpha_{t_2}\xi_{t_2} + f_{\mathbf{w}^*}(\mathbf{x}_{t_2}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_2})) \cdot \mathbf{x}_{t_2}^\top\mathbf{M}^\top\mathbf{M}\mathbf{x}_{t_1}\middle| S_{t-1}\right] \\
&\leq \frac{\eta^2(b^2-b)}{b^2}\Bigg[\theta_*^2\left(\mathbb{E}_{x_{t_1}}\left[\beta(x_{t_1})\|\mathbf{M}x_{t_1}\|\middle| S_{t-1}\right]\right)^2 \\
&\qquad + 2\theta_*\mathbb{E}_{x_{t_1}}\left[(f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1}))\|\mathbf{M}x_{t_1}\|\middle| S_{t-1}\right]\mathbb{E}_{x_{t_1}}\left[\beta(x_{t_1})\|\mathbf{M}x_{t_1}\|\middle| S_{t-1}\right] \\
&\qquad\quad + \mathbb{E}_{x_{t_1}}\left[(f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1}))\|\mathbf{M}x_{t_1}\|\middle| S_{t-1}\right]^2\Bigg] \\
&\leq \frac{\eta^2(b^2-b)}{b^2}\left[\theta_*^2\lambda_2^2\beta_1^2 + 2\theta_*\lambda_2^2\beta_1 c\mathrm{m}_2\|\mathbf{w}^{(t)} - \mathbf{w}^*\| + \lambda_2^2c^2\mathrm{m}_2^2\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2\right]
\end{aligned}$$

B Estimating the necessary recursion

Lemma 2. *Suppose we have a sequence of real numbers $\varDelta_1, \varDelta_2, \ldots$ s.t*

$$\varDelta_{t+1} \leq (1 - \eta' b + \eta'^2 c_1)\varDelta_t + \eta'^2 c_2 + \eta' c_3$$

for some fixed parameters $b, c_1, c_2, c_3 > 0$ s.t $\varDelta_1 > \frac{c_3}{b}$ and free parameter $\eta' > 0$. Then for,

$$\epsilon'^2 \in \Big(\frac{c_3}{b}, \varDelta_1\Big),\ \eta' = \frac{b}{\gamma c_1}, \quad \gamma > \max\left\{\frac{b^2}{c_1}, \left(\frac{\epsilon'^2 + \frac{c_2}{c_1}}{\epsilon'^2 - \frac{c_3}{b}}\right)\right\} > 1$$

it follows that $\varDelta_{\mathrm{T}} \leq \epsilon'^2$ for,

$$\mathrm{T} = \mathcal{O}\left(\log\left[\frac{\varDelta_1}{\epsilon'^2 - \left(\frac{\frac{c_2}{c_1} + \gamma\cdot\frac{c_3}{b}}{\gamma - 1}\right)}\right]\right)$$

Proof. Let us define $\alpha = 1 - \eta' b + \eta'^2 c_1$ and $\beta = \eta'^2 c_2 + \eta' c_3$. Then by unrolling the recursion we get,

$$\Delta_t \leq \alpha \Delta_{t-1} + \beta \leq \alpha(\alpha \Delta_{t-2} + \beta) + \beta \leq ... \leq \alpha^{t-1} \Delta_1 + \beta(1 + \alpha + \ldots + \alpha^{t-2}).$$

Now suppose that the following are true for ϵ' as given and for α & β (evaluated for the range of η's as specified in the theorem),

Claim 1 : $\alpha \in (0, 1)$

Claim 2 : $0 < \epsilon'^2(1 - \alpha) - \beta$

We will soon show that the above claims are true. Now if T is s.t we have,

$$\alpha^{\mathrm{T}-1} \Delta_1 + \beta(1 + \alpha + \ldots + \alpha^{\mathrm{T}-2}) = \alpha^{\mathrm{T}-1} \Delta_1 + \beta \cdot \frac{1 - \alpha^{\mathrm{T}-1}}{1 - \alpha} = \epsilon'^2$$

then $\alpha^{\mathrm{T}-1} = \frac{\epsilon'^2(1-\alpha)-\beta}{\Delta_1(1-\alpha)-\beta}$. Note that **Claim 2** along with with the assumption that $\epsilon'^2 < \Delta_1$ ensures that the numerator and the denominator of the fraction in the RHS are both positive. Thus we can solve for T as follows,

$$(\mathrm{T} - 1) \log\left(\frac{1}{\alpha}\right) = \log\left[\frac{\Delta_1(1 - \alpha) - \beta}{\epsilon'^2(1 - \alpha) - \beta}\right] \implies \mathrm{T} = \mathcal{O}\left(\log\left[\frac{\Delta_1}{\epsilon'^2 - \left(\frac{\frac{c_2}{c_1} + \gamma \cdot \frac{c_3}{b}}{\gamma - 1}\right)}\right]\right)$$

In the second equality above we have estimated the expression for T after substituting $\eta' = \frac{b}{\gamma c_1}$ in the expressions for α and β.

Proof of claim 1 : $\alpha \in (0, 1)$

We recall that we have set $\eta' = \frac{b}{\gamma c_1}$. This implies that, $\alpha = 1 - \frac{b^2}{c_1} \cdot \left(\frac{1}{\gamma} - \frac{1}{\gamma^2}\right)$. Hence $\alpha > 0$ is ensured by the assumption that $\gamma > \frac{b^2}{c_1}$. And $\alpha < 1$ is ensured by the assumption that $\gamma > 1$

Proof of claim 2 : $0 < \epsilon'^2(1 - \alpha) - \beta$

We note the following,

$$\begin{aligned}
-\frac{1}{\epsilon'^2}\cdot\left(\epsilon'^2(1-\alpha)-\beta\right) &= \alpha - \left(1-\frac{\beta}{\epsilon'^2}\right) \\
&= 1-\frac{b^2}{4c_1}+\left(\eta'\sqrt{c_1}-\frac{b}{2\sqrt{c_1}}\right)^2-\left(1-\frac{\beta}{\epsilon'^2}\right) \\
&= \frac{\eta'^2c_2+\eta'c_3}{\epsilon'^2}+\left(\eta'\sqrt{c_1}-\frac{b}{2\sqrt{c_1}}\right)^2-\frac{b^2}{4c_1} \\
&= \frac{\left(\eta'\sqrt{c_2}+\frac{c_3}{2\sqrt{c_2}}\right)^2-\frac{c_3^2}{4c_2}}{\epsilon'^2}+\left(\eta'\sqrt{c_1}-\frac{b}{2\sqrt{c_1}}\right)^2-\frac{b^2}{4c_1} \\
&= \eta'^2\Bigg(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{c_3}{2\eta'\sqrt{c_2}}\right)^2+\left(\sqrt{c_1}-\frac{b}{2\eta'\sqrt{c_1}}\right)^2 \\
&\qquad -\frac{1}{\eta'^2}\left[\frac{b^2}{4c_1}+\frac{1}{\epsilon'^2}\left(\frac{c_3^2}{4c_2}\right)\right]\Bigg)
\end{aligned}$$

Now we substitute $\eta' = \frac{b}{\gamma c_1}$ for the quantities in the expressions inside the parantheses to get,

$$\begin{aligned}
-\frac{1}{\epsilon'^2}\cdot\left(\epsilon'^2(1-\alpha)-\beta\right) = \alpha-\left(1-\frac{\beta}{\epsilon'^2}\right) &= \eta'^2\Bigg(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{\gamma c_1c_3}{2b\sqrt{c_2}}\right)^2 \\
&\qquad +c_1\cdot\left(\frac{\gamma}{2}-1\right)^2-c_1\frac{\gamma^2}{4}-\frac{1}{\epsilon'^2}\cdot\frac{\gamma^2c_1^2c_3^2}{4b^2c_2}\Bigg) \\
&= \eta'^2\left(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{\gamma c_1c_3}{2b\sqrt{c_2}}\right)^2+c_1(1-\gamma)-\frac{1}{\epsilon'^2}\cdot\frac{\gamma^2c_1^2c_3^2}{4b^2c_2}\right) \\
&= \frac{\eta'^2}{\epsilon'^2}\left(c_2+\frac{\gamma c_1c_3}{b}-\epsilon'^2c_1(\gamma-1)\right) \\
&= \frac{\eta'^2c_1}{\epsilon'^2}\left((\epsilon'^2+\frac{c_2}{c_1})-\gamma\cdot\left(\epsilon'^2-\frac{c_3}{b}\right)\right)
\end{aligned}$$

Therefore, $-\frac{1}{\epsilon'^2}\left(\epsilon'^2(1-\alpha)-\beta\right) < 0$ since by assumption $\epsilon'^2 > \frac{c_3}{b}$, and $\gamma > \frac{\left(\epsilon'^2+\frac{c_2}{c_1}\right)}{\epsilon'^2-\frac{c_3}{b}}$

C Lemmas For Theorem 1

Lemma 3 (Lemma 1, [12]). *If $\mathcal{D}$ is parity symmetric distribution on $\mathbb{R}^n$ and σ is an $\alpha-$Leaky ReLU then $\forall$ $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\sigma(\mathbf{a}^\top\mathbf{x})\mathbf{b}^\top\mathbf{x}\Big] = \frac{1+\alpha}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[(\mathbf{a}^\top\mathbf{x})(\mathbf{b}^\top\mathbf{x})\Big]$$

Lemma 4.

$$(f_{\mathbf{w}_*}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x}))^2 \leq (1+\alpha)^2\Big(\frac{1}{k}\sum_{i=1}^{k}\lambda_{\max}(A_iA_i^\top)\Big)\|\mathbf{w}_* - \mathbf{w}\|^2\|\mathbf{x}\|^2$$

Proof.

$$\begin{aligned}
&\Big(f_{\mathbf{w}_*}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})\Big)^2 \leq \Big(\frac{1}{k}\sum_{i=1}^{k}\sigma\Big(\Big\langle \mathrm{A}_i^\top\mathbf{w}_*, \mathbf{x}\Big\rangle\Big) - \frac{1}{k}\sum_{i=1}^{k}\sigma\Big(\Big\langle \mathrm{A}_i^\top\mathbf{w}, \mathbf{x}\Big\rangle\Big)\Big)^2 \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\Big(\sigma\Big(\Big\langle \mathrm{A}_i^\top\mathbf{w}_*, \mathbf{x}\Big\rangle\Big) - \sigma\Big(\Big\langle \mathrm{A}_i^\top\mathbf{w}, \mathbf{x}\Big\rangle\Big)\Big)^2 \\
&\leq \frac{(1+\alpha)^2}{k}\sum_{i=1}^{k}\Big\langle \mathrm{A}_i^\top\mathbf{w}_* - \mathrm{A}_i^\top\mathbf{w}, \mathbf{x}\Big\rangle^2 = \frac{(1+\alpha)^2}{k}\sum_{i=1}^{k}\Big((\mathbf{w}_* - \mathbf{w})^\top\mathrm{A}_i\mathbf{x}\Big)^2 \\
&= \frac{(1+\alpha)^2}{k}\sum_{i=1}^{k}\|\mathbf{w}_* - \mathbf{w}\|^2\|\mathrm{A}_i\mathbf{x}\|^2 \leq \frac{(1+\alpha)^2}{k}\sum_{i=1}^{k}\|\mathbf{w}_* - \mathbf{w}\|^2\lambda_{\max}(\mathrm{A}_i\mathrm{A}_i^\top)\|\mathbf{x}\|^2 \\
&\leq \frac{(1+\alpha)^2}{k}\Big(\sum_{i=1}^{k}\lambda_{\max}(\mathrm{A}_i\mathrm{A}_i^\top)\Big)\|\mathbf{w}_* - \mathbf{w}\|^2\|\mathbf{x}\|^2
\end{aligned}$$