A Study of Neural Training with Iterative Non-Gradient Methods

Sayar Karmakar[†], Anirbit Mukherjee[*], Ramchandran Muthukumar[‡],

Abstract

In this work we demonstrate provable guarantees on the training of depth-2 neural networks in new regimes than previously explored. (1) First we give a simple stochastic algorithm that can train a ReLU gate in the realizable setting in *linear time* while using significantly milder conditions on the data distribution than previous results. Leveraging some additional distributional assumptions we also show approximate recovery of the true label generating parameters when training a ReLU gate while a probabilistic adversary is allowed to corrupt the true labels of the training data. Our guarantee on recovering the true weight degrades gracefully with increasing probability of attack and its nearly optimal in the worst case. Additionally our analysis allows for mini-batching and computes how the convergence time scales with the mini-batch size. (2) Secondly, we exhibit a non-gradient iterative algorithm "Neuro-Tron" which gives a first-of-its-kind poly-time approximate solving of a neural regression (here in the ℓ_∞-norm) problem at finite net widths and for non-realizable data.

CONTENTS

[†] Department of Statistics, University of Florida, sayarkarmakar@ufl.edu
[*] Department of Applied Mathematics and Statistics, Johns Hopkins University, anirbit@wharton.upenn.edu
[‡] Department of Computer Science, Johns Hopkins Univeristy, rmuthuk1@jhu.edu

I. INTRODUCTION

In this paper we prove results about trainability of certain depth 2 neural nets under more general settings than previously analyzed. Specialized to neural nets, *learning theory* seeks to solve the following function optimization/*risk minimization* problem,

$$\min_{\mathbf{N} \in \mathcal{N}} \mathbb{E}_{\mathbf{z} \in \mathcal{D}}[\ell(\mathbf{N}, \mathbf{z})] \tag{1}$$

where ℓ is some lower-bounded non-negative function, members of $\mathcal{N}$ are continuous piecewise linear functions representable by some chosen neural net architecture and we only have sample access to the distribution $\mathcal{D}$. This reduces to the *empirical risk minimization* question when this $\mathcal{D}$ is an uniform distribution on a finite set of points. To the best of our knowledge about the state-of-the-art in deep-learning any of these two optimization problems is typically solvable in either of the following two mutually exclusive scenarios : **(a)** the data comes as tuples $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ with $\mathbf{y}$ being the noise corrupted output of a net (of known architecture on which determines the class $\mathcal{N}$) when given $\mathbf{x}$ as an input. And **(b)** the data comes as tuples $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ with no obvious functional relationship between $\mathbf{x}$ and $\mathbf{y}$ but there could be geometrical or statistical assumptions about the $\mathbf{x}$ and $\mathbf{y}$.

We note that in the fully agnostic setting as in that case training even a single ReLU gate can be SPN-hard as shown in [16]. On the other hand the simplifications that happen for infinitely large networks have been discussed since [28] and this theme has had a recent resurgence in works like [9], [20]. Eventually this lead to an explosion of literature getting linear time training of various kinds of neural nets when their width is a high degree polynomial in training set size, inverse accuracy and inverse confidence parameters (a somewhat *unrealistic* regime), [25], [34], [13], [32], [23], [19], [3], [2], [4], [12], [35], [36], [7], [8], [26], [6]. The essential essential proximity of this regime to kernel methods have been thought of separately in works like [1], [33]

Even in the wake of this progress, it remains unclear as to how to establish rigorous guarantees about constant size neural nets which is a regime closer to what is implemented in the real world. Thus motivated we can summarize what is open about training depth 2 nets into the following two questions,

1) **Question** 1 Can (Stochastic) Gradient Descent or *any* algorithm train a ReLU gate to $\epsilon-$accuracy in poly $\left(\text{input dimension}, \log(\frac{1}{\epsilon})\right)$ time (a) while using random or arbitrary initialization and (b) for a generic class of data distributions which neither assumes symmetry nor compact support?

2) **Question** 2 Can a neural training algorithm work with the following naturally wanted properties being simultaneously true?

   a) Nets of depth 2 with a constant/small number of gates.

   b) The training data instances (and maybe also the noise) would have non-Gaussian non-compactly supported distributions.

   c) Less structural assumptions on the weight matrices than being of the single filter convolutional type.

   d) $\epsilon-$approximate answers be obtainable in at most poly(input$-$dimension, $\frac{1}{\epsilon}$) time.

II. A SUMMARY OF OUR RESULTS

We make progress on some of the above fronts by drawing inspiration from and building on top of the different avatars of the iterative stochastic non-gradient *Tron* algorithms analyzed in the past like, [30], [29], [15], [21], [24], [17], [18]. We will be working exclusively with the function class of neural nets whose activation function can be the ReLU function which maps $\mathbb{R}^n \ni \mathbf{x} \mapsto \max\{0, \mathbf{w}^\top \mathbf{x}\} \in \mathbb{R}$ for $\mathbf{w} \in \mathbb{R}^n$ being its weight. Hence for such nets the corresponding empirical or the population risk is neither convex nor smooth in how it depends on the weights.

Our results in this work can be said to be of two types:

A. **Guarantees for algorithms specific to a single** ReLU **gate**

In the short Section II-D we first do a quick re-analysis of a known algorithm called GLM-Tron under more general conditions than previously to show how well it can do (empirical) risk minimization on any Lipschitz gate with Lipschitz constant < 2 (in particular a ReLU gate) in the noisily realizable setting while no assumptions are being made on the distribution of the noise beyond their boundedness - hence the noise can be *adversarial*. We also point

out how the result can be improved under some assumptions on the noise making it more benign. In Section III we will use a stochastic algorithm while solving a similar regression problem specific to a ReLU gate and there we shall exploit the structure of the ReLU gate (and mild distributional assumptions) to directly achieve parameter recovery. In contrast to the results in Section III, in Section II-D we shall be using full-batch iterative updates to gain these other advantages of being able to handle more general gates while having essentially no distributional assumptions on the training data.

We show 2 kinds of results specific to training a ReLU gate.

Firstly, in Section III we show a very simple iterative stochastic algorithm to recover in linear time the underlying parameter $\mathbf{w}_*$ of the ReLU gate when the data being samples is exactly realizable being of the form $(\mathbf{x}, \max\{0, \mathbf{w}_*^\top \mathbf{x}\})$. That is w.h.p in $\log\left(\frac{1}{\epsilon}\right)$ iterations we get ϵ close to $\mathbf{w}_*$ while starting from any arbitrary initial point. To achieve this we use a mild distributional condition which essentially just captures the intuition that enough of our samples are such that $\mathbf{w}_*^\top \mathbf{x} > 0$. To the best of our knowledge this is the first example of nearly distribution free training of a ReLU gate in linear time.

Secondly, by making a slightly stronger distributional assumption, in Case (II) of the Theorem 3.1 in Section III we also encompass the case when during training the oracle behaves adversarially i.e it tosses a biased coin and decides whether or not to additively distort the true labels by a bounded perturbation. Additionally we also allow for the bias to be data-dependent. This is a "data-poisoning" attack since the adversary corrupts the training data in an online fashion. In this case we show that the accuracy of the algorithm in recovering $\mathbf{w}_*$ is not only worst-case near optimal but is such that the accuracy degrades gracefully as the probability increases for the adversary to act.

To the best of our knowledge this is the first guarantee on training a ReLU gate while under any kind of adversarial attack. Also in both these cases above we allow for mini-batching in the algorithm and keep track of how the mini-batch size affects the convergence time.

B. Regression on certain finite width depth 2 nets with non-realizable data

We generalize the algorithmic ideas from GLM-Tron in Section II-D to multi gate nets by focussing on the following class of depth 2 nets,

*Definition 1 (***Single Filter Neural Nets Of Depth** 2**):** Given a set of k matrices $\mathrm{A}_i \in \mathbb{R}^{r \times n}$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ we call the following depth 2, width k neural net to be a *single filter neural net* defined by the matrices $\mathrm{A}_1, \ldots, \mathrm{A}_k$

$$\mathbb{R}^n \ni \mathbf{x} \mapsto f_{\mathbf{w}}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} \sigma\left(\mathbf{w}^\top \mathrm{A}_i \mathbf{x}\right) \in \mathbb{R}$$

and where σ is the Leaky-ReLU which maps as, $\mathbb{R} \ni y \mapsto \sigma(y) = y\mathbf{1}_{y \geq 0} + \alpha y \mathbf{1}_{y<0}$ for some $\alpha \geq 0$

Note that the above class of nets includes any single ReLU gate for $\alpha = 0, k = 1, \mathrm{A}_1 = I_{n \times n}$ and it also includes any depth 2 convolutional neural net with a single filter by setting the $\mathrm{A}_i's$ to be $0/1$ matrices such that each row has exactly one 1 and each column has at most one 1.

Now given training data $\mathcal{T} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \mathbb{R} \mid i = 1, \ldots, S\}$, we usually pose the problem of empirical risk minimization with a ℓ_2-loss. This is one of the most basic of such neural ERM problems but if the data is not realizable then we are not aware of any poly-time algorithm which can (approximately) solve this ERM or of any result which can give an analytic description of its optima. (*Though note that this question is always exactly solvable for any depth 2 net and any data in time poly(S), at any fixed input dimension and width, as had been shown in [5]*)

In Section IV (in particular in Theorem 4.3), we shall see that such distribution free finite width neural regression with non-realizable data - an otherwise unsolved problem with poly-time restrictions - does get more tractable for the above kinds of nets for certain classes of $(\mathcal{T}, \{\mathrm{A}_i\})$ if we instead work in the infinity norm and pose the following problem,

$$\min_{\mathbf{w} \in \mathbb{R}^r} \max_{i=1,\ldots,S} \left| y_i - \frac{1}{w} \sum_{k=1}^{w} \sigma\left(\langle \mathrm{A}_k^\top \mathbf{w}, \mathbf{x}_i \rangle\right) \right| \tag{2}$$

Hence in above for the given training data we are looking to find the pointwise optimal interpolant in the chosen class. In Theorem 4.3 it will be shown that using our algorithm Neuro-Tron (Algorithm 3) (which is inspired from works like [21], [18]), an approximate interpolant with provable approximation guarantee can be found in poly-time for a class of $(\mathcal{T}, \{A_i\})$.

We also point out in Section IV-A the special cases when Neuro-Tron can be interpreted as a S.G.D on a surrogate risk. Additionally, in the discussion below Theorem 4.3, we point out how the convergence result also gives bounds on the size of the set of optimal interpolants.

Remark 1: A *stochastic* version of the Neuro-Tron algorithm has been analyzed in our companion paper [27] where a data-poisoning attack is allowed to happen on the training data in an online fashion and stochastic Neuro-Tron is shown to be near optimally defending against it under a similar setup as in Section IV here.

C. Comparison with concurrent literature

Firstly, we note that the conditions on the patch matrices P_i and the distribution in [18] is subsumed in the consistency condition we use in Definition 3 with the P_is there being generalized to the A_i matrices here and the average of the P_is there being generalized to the matrix $\mathbf{M}$ here. Thus the algorithm in [18] run in the full-batch setting while the data is exactly realizable is a special case of the $\theta_{\min} = 0$ case of our Theorem 4.3.

Further, in contrast to [18], in Section IV-A1, we point out how one can sample the matrices $\{A_i, i = 1, \ldots, w\}$ and the $\mathbf{M}$ s.t the consistency condition of Definition 3, as required for the Neuro-Tron proof of Theorem 4.3, gets satisfied with high probability.

To the best of our knowledge provable training of a finitely large neural net under agnostic data is still an open question. Theorem 4.3 of ours can be seen as a step in that direction.

Secondly, [18] also contained as a special case a solution to the problem of learning a ReLU gate under realizable settings but only under symmetry assumptions on the distribution. Specific to the marginal distribution on the data being Gaussian, works like [31], [22] had solved the same problem using gradient based methods. A notable recent progress with understanding the behaviour of (stochastic) gradient descent on a ReLU gate was achieved in [14]. Their Theorem D.1 (b) is solving the same question as our Theorem 3.1 Case (I) but our algorithm in this special case not only accounts for the effect of mini-batching on the convergence time but also converges exponentially faster than whats guaranteed in [14].

In contrast to these results cited above, (a) our assumptions on the distribution are significantly milder than all previous attempts - we only make certain non-degeneracy assumptions and (b) most significantly our Theorem 3.1 Case (II) encompasses the situation of a probabilistic adversary causing distortions to the true labels. (And here too we keep track of the subtleties of using mini-batches and how the mini-batch size affects the convergence time) To the best of our knowledge this is the first work to analyze training of a ReLU gate in any kind of adversarial setup - in particular a data-poisoning attack on the training data (labels). We also allow for the adversary to decide to attack or not using a biased coin toss whose bias is allowed to be data-dependent.

In [10] the authors had given algorithms for learning of a ReLU gate in the non-realizable setting for certain nice marginal distributions on the data. We note that such results about risk minimization are incomparable to our goal in Theorem 3.1 Case (II) of recovering weights as close as possible to the true label generating weight (the $\mathbf{w}_*$ therein) of the ReLU gate under adversarial corruption of the true labels. But this result of ours can be seen as a natural regression analogue of the recent result in [11] about learning half-space indicators under a Massart noise.

D. GLM-Tron converges on certain Lipschitz gates with no distributional assumption on the data

---

Algorithm 1 GLM-Tron

1: **Input:** $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,m}$ and an *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$
2: $\mathbf{w}_1 = 0$
3: **for** $t = 1, \ldots$ **do**
4: $\quad \mathbf{w}_{t+1} := \mathbf{w}_t + \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i \rangle) \right) \mathbf{x}_i$
5: **end for**

First we state the following crucial lemma,

Lemma 2.1: Assume that for all $i = 1, \ldots, S$ $\|\mathbf{x}_i\| \leq 1$ and in Algorithm 1, σ is a $L-$Lipschitz non-decreasing function. Suppose $\exists \mathbf{w}$ s.t at iteration t, we have $\|\mathbf{w}_t - \mathbf{w}\| \leq W$ and define $\eta > 0$ s.t $\|\frac{1}{S} \sum_{i=1}^{S} \left(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i \rangle) \right) \mathbf{x}_i\| \leq \eta$. Then it follows that $\forall t = 1, 2, \ldots$,

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 \leq \|\mathbf{w}_t - \mathbf{w}\|^2 - \left(\frac{2}{L} - 1 \right) \tilde{L}_S(h_t) + \left(\eta^2 + 2\eta W(L+1) \right)$$

where we have defined,
$$\tilde{L}_S(h_t) := \frac{1}{S} \sum_{i=1}^{S} \left(h_t(x_i) - \sigma(\langle \mathbf{w}, \mathbf{x}_i \rangle) \right)^2 = \frac{1}{S} \sum_{i=1}^{S} \left(\sigma(\langle \mathbf{w}_t, x_i \rangle) - \sigma(\langle \mathbf{w}, \mathbf{x}_i \rangle) \right)^2$$

The above algorithm was introduced in [21] for bounded activations. Here we show the applicability of that idea for more general activations and also while having adversarial attacks on the labels. We give the proof of the above lemma in Appendix A-A. Now we will see in the following theorem and its proof as to how the above lemma leads to convergence of the *effective-ERM*, $\tilde{L}_S$ by GLM-Tron on a single gate.

Theorem 2.2 (**GLM-Tron (Algorithm 1) solves the effective-ERM on a ReLU gate upto noise bound with minimal distributional assumptions**): Assume that for all $i = 1, \ldots, S$ $\|\mathbf{x}_i\| \leq 1$ and the label of the ith data point y_i is generated as, $y_i = \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) + \xi_i$ s.t $\forall i, |\xi_i| \leq \theta$ for some $\theta > 0$ and $\mathbf{w}_* \in \mathbb{R}^n$. If σ is a $L-$Lipschitz non-decreasing function for $L < 2$ then in at most $T = \frac{\|\mathbf{w}_*\|}{\epsilon}$ GLM-Tron steps we would attain parameter value $\mathbf{w}_T$ s.t,

$$\tilde{L}_S(h_T) = \frac{1}{S} \sum_{i=1}^{S} \left(\sigma(\langle \mathbf{w}_T, x_i \rangle) - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) \right)^2 < \frac{L}{2-L} \left(\epsilon + (\theta^2 + 2\theta W(L+1)) \right)$$

Remark 2: Firstly, note that in the realizable setting i.e when $\theta = 0$, the above theorem is giving an upperbound on the number of steps needed to solve the ERM on say a ReLU gate to $O(\epsilon)$ accuracy. *Secondly,* observe that the above theorem does not force any distributional assumption on the ξ_i beyond the assumption of its boundedness. Thus the noise could as well be chosen *adversarially* upto the constraint on its norm.

The above theorem is proven in Appendix A-B. If we make some assumptions on the noise being somewhat benign then we can get the following.

Theorem 2.3 (**Performance guarantees on the GLM-Tron (Algorithm 1) when solving ERM**): Assume that the noise random variables $\xi_i, i = 1, \ldots, S$ are identically distributed as a centered random variable say ξ. Then for $T = \frac{\|\mathbf{w}\|}{\epsilon}$, we have the following guarantee on the (true) empirical risk after T iterations of GLM-Tron (say $\tilde{L}_S(h_T)$),
$$\mathbb{E}_{\{(\mathbf{x}_i, \xi_i)\}_{i=1,\ldots S}} \left[L_S(h_T) \right] \leq \mathbb{E}_\xi[\xi^2] + \frac{L}{2-L} \left(\epsilon + (\theta^2 + 2\theta W(L+1)) \right)$$

The above is proven in Appendix A-C. Here we note a slight generalization of the above that can be easily read off from the above.

Corollary 2.4: Suppose that instead of assuming $\forall i = 1, \ldots, S$ $|\xi_i| \leq \theta$ we instead assume that the joint distribution of $\{\xi_i\}_{i=1,\ldots,S}$ is s.t $\mathbb{P}\left[|\xi_i| \leq \theta \ \forall i \in \{1, \ldots, S\} \right] \geq 1 - \delta$ Then it would follow that the guarantee of the above Theorem 2.3 still holds but now with probability $1 - \delta$ over the noise distribution.

In Section IV we will see a significant generalization of the above to multi-gate settings with non-realizable data - but while making some mild symmetry assumptions about the data and constraining the activation σ to be of the form of a Leaky-ReLU. *We will see that the multi-gate version of the above analysis leads to a much more complicated dynamical system to be analyzed and leads to much richer insights into the behaviour of large neural nets.*

In the next section we shall continue with the current theme of training a single neuron and see how a stochastic algorithm can be designed to get stronger training guarantees specific to a ReLU gate.

III. Learning a ReLU gate in the realizable setting under a data-poisoning attack

In this section we instantiate the following model of an adversary doing a data-poisoning attack on an iterative stochastic learning algorithm using a mini-batch size of b. Given a marginal distribution $\mathcal{D}$ on the inputs $\mathbf{x}$, suppose the corresponding true labels are generated as $y = \text{ReLU}(\mathbf{w}_*^\top \mathbf{x})$ for some unknown $\mathbf{w}_* \in \mathbb{R}^n$. We assume that we have sampling access to $\mathcal{D}$ and an adversarial label oracle that on the $t^{th}-$iterate gets queried with b inputs $\{\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_b}\}$ drawn uncorrelatedly from $\mathcal{D}$. The oracle then flips a coin for each of these minibatch data with probability of the coin returning 0 being $1 - \beta(\mathbf{x}_{t_i})$ for some fixed function $\beta : \mathbb{R}^n \to [0,1]$. We assume that these coin flips are uncorrelated to each other and the mini-batch sample and if the coin flips gives 1 only then the adversary does a bounded (by a constant θ_*) additive distortion to the true label of the corresponding data. We precisely instantiate this model as given in Algorithm 2

To learn the true labeling function $\mathbb{R}^n \ni \mathbf{y} \mapsto \text{ReLU}(\mathbf{w}_*^\top \mathbf{y}) \in \mathbb{R}$ in this adversarially corrupted realizable setting we try to solve the following optimization problem, $\min_{\mathbf{w} \in \mathbb{R}^n} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\left(\text{ReLU}(\mathbf{w}^\top \mathbf{x}) - y \right)^2 \right]$

In contrast to previous work we show that the simple algorithm given below solves this learning problem by leveraging the intuition that if we see enough labels $y = \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}) + \xi$ where $y > \theta_*$, then solving the linear regression problem on this subset of samples, gives a $\tilde{\mathbf{w}}_*$ which is close to $\mathbf{w}_*$. In the situation with adversarial corruption ($\theta_* > 0$) we show in subsection III-A that our recovery guarantee is nearly optimal. Additionally in the realizable case ($\theta_* = 0$), our setup learns to arbitrary accuracy the true weight $\mathbf{w}_*$ using much milder distributional constraints than previous such results that we are aware of.

Algorithm 2

Modified mini-batch SGD for training a ReLU gate with adversarially perturbed realizable labels.

1: **Input:** Sampling access to a distribution $\mathcal{D}$ on $\mathbb{R}^n$ and a function $\beta : \mathbb{R}^n \to [0,1]$
2: **Input:** Oracle access to labels $y \in \mathbb{R}$ when queried with some $\mathbf{x} \in \mathbb{R}^n$
3: **Input:** An arbitrarily chosen starting point of $\mathbf{w}_1 \in \mathbb{R}^n$
4: **for** $t = 1, \ldots$ **do**
5:     Sample independently $s_t := \{\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_b}\} \sim \mathcal{D}$ and query the oracle with this set.
6:     The Oracle samples $\forall i = 1, \ldots, b, \alpha_{t_i} \sim \{0,1\}$ with probability $\{1 - \beta(x_{t_i}), \beta(x_{t_i})\}$
7:     The Oracle replies $\forall i = 1, \ldots, b, y_{t_i} = \alpha_{t_i} \cdot \xi_{t_i} + \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i})$ s.t $|\xi_{t_i}| \leq \theta_*$
8:     Form the gradient (proxy),

$$\mathbf{g}_t := -\frac{1}{b} \sum_{i=1}^b \mathbb{1}_{\{y_{t_i} > \theta_*\}} (y_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}) \mathbf{x}_{t_i}$$

9:     $\mathbf{w}_{t+1} := \mathbf{w}_t - \eta \mathbf{g}_t$
10: **end for**

We note that the choice of $\mathbf{g}_t$ in Algorithm 2 resembles the stochastic gradient that is commonly used and is known to have great empirical success. In a true S.G.D the the indicator occurring in $\mathbf{g}_t$ would have been $\mathbb{1}_{\{\mathbf{w}_t^\top \mathbf{x}_{t_i} > 0\}}$ for each i

Towards stating the theorem we define the following notation.

Definition 2: Given $\mathbf{w}_* \in \mathbb{R}^n, \theta_* \in \mathbb{R}^+$, a distribution $\mathcal{D}$ on $\mathbb{R}^n$ and a function $\beta : \mathbb{R}^n \to [0,1]$, we define the following constants associated to them (assuming they are finite),

$$a_i := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x} > 0} \|\mathbf{x}\|^i \right], \text{ for } i = 2, 4$$

$$\beta_j := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\beta(\mathbf{x}) \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x} > 0} \|\mathbf{x}\|^j \right], \text{ for } j = 1, 2, 3$$

$$\lambda_1(\theta_*) := \lambda_{\min}\left(\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x} > 2\theta_*} \mathbf{x}\mathbf{x}^\top \right] \right)$$

*Theorem 3.1 (***Training a ReLU gate with realizable data and a probabilistic data-poisoning adversary, Proof in Subsection III-B**):*

In Algorithm 2 we will assume that (a) for $i \neq j$ and for all t, the random variables/data samples $\mathbf{x}_{t_i}$ and $\mathbf{x}_{t_j}$ are uncorrelated and (b) that the random variables α_{t_i} and α_{t_j} are mutually uncorrelated and also uncorrelated with the the mini-batch choice s_t.

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

Suppose (a) $\mathbb{E}\left[\|\mathbf{x}\|^4\right]$ and the covariance matrix $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$ exist and (b) $\mathbf{w}_*$ is s.t a_4 exists and $\mathbb{E}\left[\mathbf{1}_{\mathbf{w}^\top\mathbf{x}>0}\mathbf{x}\mathbf{x}^\top\right]$ is positive definite and hence let $\lambda_1 := \lambda_1(0)$ is well defined. Then if $\lambda_1 < \infty$, one can find a suitable step-size $\eta > 0$ and run **Algorithm 2** starting from arbitrary $\mathbf{w}_1 \in \mathbb{R}^n$ so that $\forall \epsilon > 0$, $\delta \in (0,1)$, after $\mathrm{T} = O\left(\log \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{\epsilon^2 \delta}\right)$ iterations we have

$$\mathbb{P}\left[\|\mathbf{w}_\mathrm{T} - \mathbf{w}_*\|^2 \leq \epsilon^2\right] \geq 1 - \delta$$

Case II : With bounded adversarial corruption of the true labels, $\theta_* > 0$

Suppose $\mathbf{w}_*$ and θ_* are s.t (a) $a_2, a_4, \beta_1, \beta_2, \beta_3$ exist and (b) $\lambda_1(\theta_*) > 0$. Then there exists constants b_1', c_1', c_2', c_3' (to be defined below) s.t. one can choose $\eta = \frac{b_1'}{\gamma c_1'}$ and run **Algorithm 2** starting from arbitrary $\mathbf{w}_1 \in \mathbb{R}^n$ so that,

after $\mathrm{T} = O\left(\log \dfrac{\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{\epsilon^2 \delta - \theta_*^2 \cdot \left(\frac{\frac{c_2'}{c_1'} + \gamma \cdot \frac{c_3'}{b_1'}}{\gamma - 1}\right)}\right)$ iterations we have

$$\mathbb{P}\left[\|\mathbf{w}_\mathrm{T} - \mathbf{w}_*\|^2 \leq \epsilon^2\right] \geq 1 - \delta$$

were $\epsilon > 0$ and $\delta \in (0,1)$ are s.t. $\epsilon^2 \delta = \beta_1^2 \cdot \frac{K \cdot \theta_*^2}{(2\lambda_1(\theta_*) - \frac{1}{K})}$ and $K > 0$ large enough s.t $2\lambda_1(\theta_*) - \frac{1}{K} > 0$, and

$$b_1' = 2\lambda_1(\theta_*) - \frac{1}{K}, c_1' = \frac{1 + a_4 + (1 + a_2^2)(b - 1)}{b}$$

$$c_2' = \frac{1}{\beta_1}\left(\beta_3^2 + (\beta_2 \cdot a_1)^2 \cdot (b - 1) + (\beta_2 + (b - 1) \cdot \beta_1^2)\right), c_3' = K \cdot \beta_1^2,$$

$$\text{and } \gamma > \max\left(\frac{b_1'^2}{c_1'}, \frac{\epsilon^2 \delta + \theta_*^2 \cdot \frac{c_2'}{c_1}}{\epsilon^2 \delta - \theta_*^2 \cdot \frac{c_3'}{b_1}}\right).$$

Remark 3: We note the following salient points about the structure of Theorem 3.1 :

(a) Note that for any fixed δ, the ϵ error guaranteed by the theorem approaches 0 as $\sup_\mathbf{x} \beta(\mathbf{x}) \to 0$. Thus we have continuous improvement of the minimum achievable error as the likelihood of the data-poisoning attack decreases.
(b) $\|\mathbf{w}_\mathrm{T} - \mathbf{w}_*\|^2 \leq \epsilon^2 \implies \mathbb{E}_\mathbf{x}\left[\left(\mathrm{ReLU}(\mathbf{w}_\mathrm{T}^\top\mathbf{x}) - \mathrm{ReLU}(\mathbf{w}_*^\top\mathbf{x})\right)^2\right] \leq \epsilon^2 \mathbb{E}\left[\|\mathbf{x}\|^2\right]$ and hence Algorithm 2 solves the risk minimization problem for $\theta = 0$ to any desired accuracy and in linear time. **(c)** Note that the above convergence holds starting from an arbitrary initialization $\mathbf{w}_1$. **(d)** In subsection III-A we shall see how the above theorem gives a *worst-case* near-optimal trade-off between ϵ (the accuracy) and δ (the confidence) that can be achieved when training against a θ^* (a *constant*) additive norm bounded adversary corrupting the true output.

(e) *Convergence speed increases with the minibatch size b* In the Case (I) above i.e when $\theta_* = 0$, one can read off from the proof that upon defining $b_1 = 2\lambda_1$ & $c_1 = \frac{a_4 + a_2^2(b - 1)}{b}$, one can find δ_0 so that $c_1 > \frac{b_1^2 \delta_0}{(1 + \delta_0)^2}$ and choose $\eta = b_1/(c_1(1 + \delta_0))$ to obtain

$$\mathrm{T} = 1 + \left(\frac{\log \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{\epsilon^2 \delta}}{\log \frac{1}{\alpha}}\right)$$

$$\text{where } \alpha = 1 - \frac{4\lambda_1^2 \delta_0}{\left(a_2^2 + \frac{(a_4 - a_2^2)}{b}\right) \cdot (1 + \delta_0)^2}$$

Note that this T is a decreasing function of the batchsize b and hence quantifies the intuition that a pre-specified level of precision to be achieved earlier when using larger batch-sizes.

A similar conclusion prevails in the $\theta_* > 0$ case as well.

(f) *The distributional condition is mild*

Corresponding to both the situations, $\theta_* = 0$ and $\theta_* > 0$, here we provide simple examples that satisfy the condition of $\lambda_1(\theta_*) > 0$,

Example 1: Compact multivariate distribution

Suppose $n = 2$ and $\mathbf{x} \sim \text{Unif}[-1, 1] \times [-1, 1]$ and suppose $\mathbf{w}_* = (-1, 1)$. Hence we can define,

$$d_1(\theta_*) := \mathbb{E}(\mathbf{1}_{-x_1 + x_2 > 2\theta_*} x_1^2) = \mathbb{E}(\mathbf{1}_{x_1 + x_2 > 2\theta_*} x_2^2) = \frac{1}{48}(7 - 8\theta_* + (2\theta_* - 1)^4)$$

$$d_2(\theta_*) := \mathbb{E}(\mathbf{1}_{-x_1 + x_2 > 2\theta_*} x_1 x_2) = \frac{1}{32} - \frac{4\theta_*}{24} + \frac{4\theta_*^2 - 1}{16} - \frac{(2\theta_* - 1)^4}{32}$$
$$+ \frac{4\theta_*(2\theta_* - 1)}{24} - \frac{(4\theta_*^2 - 1)(2\theta_* - 1)^2}{16}$$

Then we have $\lambda_1(\theta_*) := \lambda_{\min}\left(\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\mathbf{1}_{-x_1 + x_2 > 2\theta_*} \mathbf{x}\mathbf{x}^\top\right]\right) = d_1(\theta_*) - |d_2(\theta_*)|$

Hence ensuring convergence needs, $d_1(\theta_*) > |d_2(\theta_*)|$ and this is satisfied for examples such as : (a) $\theta_* = 0$, $\lambda_1(0) = \frac{1}{6} - 0 = \frac{1}{6}$ (b) $\theta_* = 1$, $\lambda_1(1) = \frac{1}{16} - \frac{5}{96} = \frac{1}{96}$.

Example 2: Non-compact univariate distribution

Suppose $n = 1, x \sim \mathcal{N}(0, 1)$. Then for any w_* we have,

$$0 < \lambda_1(\theta_*) = \mathbb{E}(\mathbf{1}_{w_* x > 2\theta_*} x^2) \leq \int_{-\infty}^{\infty} x^2 \phi(x) dx = 1$$

where $\phi(x)$ is the standard normal pdf. This implies $\lambda_1(\theta_*)$ is finite and positive and thus ensures convergence.

It is easy to demonstrate further examples in other univariate/multivariate and compact/non-compact distributions as well and see that the convergence conditions are not very strong.

A. Near-optimality of Theorem 3.1

We consider the "worst case" situation of Theorem 3.1 i.e when $\beta = 1$ identically and hence the adversary always acts. Now consider another value for the filter $\mathbb{R}^r \ni \mathbf{w}_{\text{adv}} \neq \mathbf{w}^*$ being chosen by this adversary and suppose that $\theta^* = \theta_{\text{adv}}$ s.t

$$\theta_{\text{adv}} \geq \sup_{\mathbf{x} \in \text{supp}(\mathcal{D})} |\text{ReLU}(\mathbf{w}_{\text{adv}}^\top \mathbf{x}) - \text{ReLU}(\mathbf{w}_*^\top \mathbf{x})| \tag{3}$$

It is easy to imagine cases where the supremum in the RHS above exists like when $\mathcal{D}$ is compactly supported. Now in this situation we define $\mathbf{c}_{\text{bound}} := \frac{(2\lambda_1(\theta_*) - \frac{1}{K})}{\beta_1^2 \cdot K}$ and hence Theorem 3.1 says that the lowest value of the parameter error achievable is,

$$\epsilon^2 = \frac{\theta^{\star 2}}{\delta c_{\text{bound}}} \implies \epsilon^2 \geq \frac{\theta^2_{\text{adv}}}{c_{\text{bound}}} \tag{4}$$

Hence proving optimality of this guarantee is equivalent to showing the existence of an attack within this θ_{adv} bound for which the best accuracy possible nearly saturates the lowerbound in equation 4

We note that for the choice of corruption bound θ_{adv}, the adversarial oracle when queried with $\mathbf{x}$ can respond with $\xi_{\mathbf{x}} + \text{ReLU}(\mathbf{w}_*^\top \mathbf{x})$ where $\xi_{\mathbf{x}} = \text{ReLU}(\mathbf{w}_{\text{adv}}^\top \mathbf{x}) - \text{ReLU}(\mathbf{w}_*^\top \mathbf{x})$. Hence the data received by the algorithm can be exactly realized with the filter choice $\mathbf{w}_{\text{adv}}$. In that case the analysis of Theorem 3.1, Case (I) shows that Algorithm 2 will converge in high probability to $\mathbf{w}_{\text{adv}}$. Thus the error incurred is $\epsilon \geq \|\mathbf{w}_{\text{adv}} - \mathbf{w}_*\|$.

An instantiation of the above attack happening is when $\theta_{\text{adv}} = r\|\mathbf{w}_{\text{adv}} - \mathbf{w}_*\|$ for $r = \sup_{\mathbf{x} \in \text{supp}(\mathcal{D})} \|\mathbf{x}\|$. Its easy to imagine cases where $\mathcal{D}$ is s.t r defined above is finite. Further, this choice of θ_{adv} is valid since the following holds as required by equation 3,

$$\sup_{\mathbf{x} \in \text{supp}(\mathcal{D})} |\text{ReLU}(\mathbf{w}_{\text{adv}}^\top \mathbf{x}) - \text{ReLU}(\mathbf{w}_*^\top \mathbf{x})| \leq r\|\mathbf{w}_{\text{adv}} - \mathbf{w}_*\| = \theta_{\text{adv}}$$

Thus the above setup invoked on training a ReLU gate with inputs being sampled from $\mathcal{D}$ as above while the labels are being additively corrupted by at most $\theta_*(= \theta_{\text{adv}}) = r\|\mathbf{w}_{\text{adv}} - \mathbf{w}_*\|$ demonstrates a case where the *worst case* accuracy guarantee of $\epsilon^2 \geq \frac{\theta^2_{\text{adv}}}{c_{\text{bound}}}$ is optimal upto a constant $\frac{r^2}{c_{\text{bound}}}$. We note that this argument also implies the worst-case near optimality of guarantees like equation 4 for *any* algorithm defending against this attack which also has the property of recovering the parameters correctly when the labels are exactly realizable.

B. Proof of Theorem 3.1

Proof:

Here we analyze the dynamics of the Algorithm 2.

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 = \|\mathbf{w}_t - \eta \mathbf{g}_t - \mathbf{w}_*\|^2 = \|\mathbf{w}_t - \mathbf{w}_*\|^2 + \eta^2 \|\mathbf{g}_t\|^2 - 2\eta \langle \mathbf{w}_t - \mathbf{w}_*, \mathbf{g}_t \rangle$$

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 $\mathbf{s}_t$ 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,

$$\mathbb{E}_{s_t, \alpha_t}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \Big| S_{t-1} \right]$$

$$= \mathbb{E}_{s_t, \alpha_t}\left[\|\mathbf{w}_t - \mathbf{w}_*\|^2 \Big| S_{t-1} \right] + 2\frac{\eta}{b} \cdot \underbrace{\sum_{i=1}^b \mathbb{E}_{\mathbf{x}_{t_i}, \alpha_{t_i}}\left[\left\langle \mathbf{w}_t - \mathbf{w}_*, \mathbf{1}_{y_{t_i} > \theta_*} \left(y_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i} \right) \mathbf{x}_{t_i} \right\rangle \Big| S_{t-1} \right]}_{\text{Term 1}}$$

$$+ \underbrace{\eta^2 \mathbb{E}_{\mathbf{x}_{t_i}, \alpha_{t_i}}\left[\|\mathbf{g}_t\|^2 \Big| S_{t-1} \right]}_{\text{Term 2}} \tag{5}$$

Now we simplify the last two terms of the RHS above, starting from the rightmost,

$$\text{Term 2} = \eta^2 \cdot \mathbb{E}\left[\|\mathbf{g}_t\|^2 \mid S_{t-1}\right]$$

$$= \frac{\eta^2}{b^2} \sum_{i,j=1}^{b} \mathbb{E}\left[\mathbf{1}_{y_{t_i} > \theta_*} \mathbf{1}_{y_{t_j} > \theta_*} \cdot (y_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}) \cdot (y_{t_j} - \mathbf{w}_t^\top \mathbf{x}_{t_j}) \cdot \langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j} \rangle \Big| S_{t-1}\right]$$

$$= \frac{\eta^2}{b^2} \sum_{i,j=1}^{b} \mathbb{E}\left[\mathbf{1}_{y_{t_i} > \theta_*} \mathbf{1}_{y_{t_j} > \theta_*} \langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j} \rangle \cdot \Big[\alpha_{t_i} \alpha_{t_j} \xi_{t_i} \xi_{t_j} \right.$$

$$+ \left(\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i} \right) \left(\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_j}) - \mathbf{w}_t^\top \mathbf{x}_{t_j} \right)$$

$$\left. + \alpha_{t_i} \xi_{t_i} \left(\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_j}) - \mathbf{w}_t^\top \mathbf{x}_{t_j} \right) + \alpha_{t_j} \xi_{t_j} \left(\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i} \right) \Big] \Big| S_{t-1}\right]$$

$$\leq \frac{\eta^2}{b^2} \sum_{i,j=1}^{b} \left(\mathbb{E}\left[\mathbf{1}_{y_{t_i} > \theta_*} \mathbf{1}_{y_{t_j} > \theta_*} |\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j} \rangle| \right.\right.$$

$$\times \left[\alpha_{t_i} \alpha_{t_j} \theta_*^2 + |\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i}| \cdot |\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_j}) - \mathbf{w}_t^\top \mathbf{x}_{t_j}| \right.$$

$$\left.\left.\left. + \theta_* \left(\alpha_{t_i} \left| \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_j}) - \mathbf{w}_t^\top \mathbf{x}_{t_j} \right| + \alpha_{t_j} \left| \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i} \right| \right) \right] \Big| S_{t-1}\right]\right)$$

As events we have for, $k = i, j$, $\mathbf{1}_{y_{t_k} > \theta_*} \subset \mathbf{1}_{\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_k}) > 0} = \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_k} > 0}$

And hence we can simplify as follows,

$$\leq \frac{\eta^2}{b^2} \sum_{i,j=1}^{b} \left\{ \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j} > 0} |\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j} \rangle| \right.\right.$$

$$\cdot \left[\alpha_{t_i} \alpha_{t_j} \theta_*^2 + |\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i}| \cdot |\text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_j}) - \mathbf{w}_t^\top \mathbf{x}_{t_j}| \right.$$

$$\left.\left.\left. + \theta_* \left(\alpha_{t_i} \left| \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_j}) - \mathbf{w}_t^\top \mathbf{x}_{t_j} \right| + \alpha_{t_j} \left| \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i} \right| \right) \right] \Big| S_{t-1}\right]\right\}$$

$$\leq \frac{\eta^2}{b^2} \sum_{i,j=1}^{b} \left\{ \theta_*^2 \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j} > 0} |\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j} \rangle| \cdot \left[(\beta(\mathbf{x}_{t_i}) \mathbf{1}_{i=j} + \beta(\mathbf{x}_{t_i}) \beta(\mathbf{x}_{t_j}) \mathbf{1}_{i \neq j}) \right] \Big| S_{t-1}\right] \right.$$

$$+ \mathbf{1}_{i \neq j} \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \|\mathbf{x}_{t_i}\| \cdot |\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}| \Big| S_{t-1}\right] \times \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j} > 0} \|\mathbf{x}_{t_j}\| \cdot |\mathbf{w}_*^\top \mathbf{x}_{t_j} - \mathbf{w}_t^\top \mathbf{x}_{t_j}| \Big| S_{t-1}\right]$$

$$+ \mathbf{1}_{i=j} \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \|\mathbf{x}_{t_i}\|^2 \cdot |\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}|^2 \Big| S_{t-1}\right]$$

$$+ \theta_* \cdot \mathbf{1}_{i \neq j} \cdot \left(\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \cdot \beta(\mathbf{x}_{t_i}) \cdot \|\mathbf{x}_{t_i}\| |\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}| \Big| S_{t-1}\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j} > 0} \|\mathbf{x}_{t_j}\| \Big| S_{t-1}\right] + (i \leftrightarrow j) \right)$$

$$\left. + 2\theta_* \cdot \mathbf{1}_{i=j} \cdot \left(\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \cdot \beta(\mathbf{x}_{t_i}) \cdot \|\mathbf{x}_{t_i}\|^2 |\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}| \Big| S_{t-1}\right] \right) \right\} \tag{6}$$

In the last inequality above we have used the facts that (a) for $i \neq j$, functions of $\mathbf{x}_{t_i}$ are uncorrelated with functions of $\mathbf{x}_{t_j}$ and (b) that the random variables α_{t_i} and α_{t_j} are independent of each other and of the mini-batch choice s_t and hence they can be replaced by their respective expectations $\beta(\mathbf{x}_{t_i})$ and $\beta(\mathbf{x}_{t_j})$. And for the first term we need to note the $i = j$ case that, $\mathbb{E}[\alpha_{t_i}^2] = \beta(\mathbf{x}_{t_i})$

Now we can simplify the first term of the RHS of equation 6 as,

$$\theta_*^2 \cdot \mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*}\mathbf{1}_{y_{t_j}>\theta_*}|\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j}\rangle| \cdot \left[(\beta(\mathbf{x}_{t_i})\mathbf{1}_{i=j} + \beta(\mathbf{x}_{t_i})\beta(\mathbf{x}_{t_j})\mathbf{1}_{i\neq j})\right] \Big| S_{t-1}\right]$$

$$\leq \theta_*^2 \cdot \mathbb{E}_{\mathbf{x}_{t_i}}\left[\beta(\mathbf{x}_{t_i})\|\mathbf{x}_{t_i}\|^2 \mathbf{1}_{y_{t_i}>\theta_*}\Big| S_{t-1}\right]\mathbf{1}_{i=j}$$

$$+ \theta_*^2 \cdot \mathbb{E}_{\mathbf{x}_{t_i}}\left[\beta(\mathbf{x}_{t_i})\|\mathbf{x}_{t_i}\|\mathbf{1}_{y_{t_i}>\theta_*}\Big| S_{t-1}\right] \cdot \mathbb{E}_{\mathbf{x}_{t_j}}\left[\beta(\mathbf{x}_{t_j})\|\mathbf{x}_{t_j}\|\mathbf{1}_{y_{t_j}>\theta_*}\Big| S_{t-1}\right]\mathbf{1}_{i\neq j}$$

$$\leq \theta_*^2 \cdot \mathbb{E}_{\mathbf{x}_{t_i}}\left[\beta(\mathbf{x}_{t_i})\|\mathbf{x}_{t_i}\|^2 \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0}\Big| S_{t-1}\right]\mathbf{1}_{i=j}$$

$$+ \theta_*^2 \cdot \mathbb{E}_{\mathbf{x}_{t_i}}\left[\beta(\mathbf{x}_{t_i})\|\mathbf{x}_{t_i}\|\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0}\Big| S_{t-1}\right] \cdot \mathbb{E}_{\mathbf{x}_{t_j}}\left[\beta(\mathbf{x}_{t_j})\|\mathbf{x}_{t_j}\|\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j}>0}\Big| S_{t-1}\right]\mathbf{1}_{i\neq j}$$

Since $\mathbf{x}_{t_i}$ & $\mathbf{x}_{t_j}$ are identically distributed, we can invoke the constants, β_1 & β_2 and under taking total expectations the above is bounded by $\theta_*^2(\beta_2\mathbf{1}_{i=j} + \beta_1^2\mathbf{1}_{i\neq j})$. Using this we have from taking total expectations on both sides of equation 6,

$$\mathbb{E}\left[\text{Term 2}\right] \leq \frac{\eta^2}{b^2} \cdot \theta_*^2(b \cdot \beta_2 + (b^2 - b) \cdot \beta_1^2)$$

$$+ \frac{\eta^2}{b^2}\sum_{i=1}^{b}\left\{\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0}\|\mathbf{x}_{t_i}\|^2 \cdot |\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}|^2 \Big| S_{t-1}\right]\right]\right.$$

$$+ 2\theta_* \cdot \left(\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0} \cdot \beta(\mathbf{x}_{t_i}) \cdot \|\mathbf{x}_{t_i}\|^2|\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}|\Big| S_{t-1}\right]\right]\right)\right\}$$

$$+ \frac{\eta^2}{b^2}\sum_{i,j=1,i\neq j}^{b}\left\{\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0}\|\mathbf{x}_{t_i}\| \cdot |\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}|\Big| S_{t-1}\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j}>0}\|\mathbf{x}_{t_j}\| \cdot |\mathbf{w}_*^\top \mathbf{x}_{t_j} - \mathbf{w}_t^\top \mathbf{x}_{t_j}|\Big| S_{t-1}\right]\right]\right.$$

$$+ \theta_* \cdot \left(\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0} \cdot \beta(\mathbf{x}_{t_i}) \cdot \|\mathbf{x}_{t_i}\||\mathbf{w}_*^\top \mathbf{x}_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}|\Big| S_{t-1}\right]\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j}>0}\|\mathbf{x}_{t_j}\|\right] + (i \leftrightarrow j)\right)\right\}$$

In the last term of the RHS above we have used the fact that conditioned on S_{t-1} a function of $(\mathbf{w}_t, \mathbf{x}_{t_i})$ is uncorrelated with a function of $\mathbf{x}_{t_j}$ for $i \neq j$. Now we further invoke that for $k = i, j$, conditioned on S_{t-1}, $\mathbf{w}_t$ is uncorrelated with any function of $\mathbf{x}_{t_k}$ to simplify the above as,

$$\mathbb{E}\left[\text{Term 2}\right] \leq \frac{\eta^2}{b^2} \cdot \theta_*^2 (b \cdot \beta_2 + (b^2 - b) \cdot \beta_1^2)$$

$$+ \frac{\eta^2}{b^2} \sum_{i=1}^{b} \left\{ \mathbb{E}\left[\|\mathbf{w}_* - \mathbf{w}_t\|^2\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \|\mathbf{x}_{t_i}\|^4\right]\right.$$

$$+ 2\theta_* \cdot \left(\mathbb{E}\left[\|\mathbf{w}_* - \mathbf{w}_t\|\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \cdot \beta(\mathbf{x}_{t_i}) \cdot \|\mathbf{x}_{t_i}\|^3\right] \right) \right\}$$

$$+ \frac{\eta^2}{b^2} \sum_{i,j=1, i \neq j}^{b} \left\{ \mathbb{E}\left[\|\mathbf{w}_* - \mathbf{w}_t\|^2\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \|\mathbf{x}_{t_i}\|^2\right] \times \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j} > 0} \|\mathbf{x}_{t_j}\|^2\right]\right.$$

$$+ \theta_* \cdot \left(\mathbb{E}\left[\|\mathbf{w}_* - \mathbf{w}_t\|\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \cdot \beta(\mathbf{x}_{t_i}) \cdot \|\mathbf{x}_{t_i}\|^2\right] \cdot \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_j} > 0} \|\mathbf{x}_{t_j}\|\right] + (i \leftrightarrow j) \right) \right\}$$

$$\leq \frac{\eta^2}{b} \cdot \left\{ a_4 \cdot X_t + 2\theta_* \cdot \mathbb{E}\left[\beta_3 \cdot \|\mathbf{w}_* - \mathbf{w}_t\|\right] \right\}$$

$$+ \frac{\eta^2}{b^2} \cdot (b^2 - b) \cdot \left\{ a_2^2 \cdot X_t + 2\theta_* \cdot \mathbb{E}\left[\beta_2 a_1 \cdot \|\mathbf{w}_* - \mathbf{w}_t\|\right] \right\} + \frac{\eta^2}{b^2} \cdot \theta_*^2 (b \cdot \beta_2 + (b^2 - b) \cdot \beta_1^2) \qquad (7)$$

In the last line above we have recalling that $\mathbf{x}_{t_i}$ and $\mathbf{x}_{t_j}$ are identically distributed and the definitions of $a_1, a_2, a_4, \beta_2 \,\&\, \beta_3$ and have defined $X_t := \mathbb{E}\left[\|\mathbf{w}_* - \mathbf{w}_t\|^2\right]$. In the second and the fourth terms of the RHS above we invoke the inequalities,

$$2\theta_* \cdot \mathbb{E}\left[\beta_3 \cdot \|\mathbf{w}_* - \mathbf{w}_t\|\right] \leq (\theta_* \cdot \beta_3)^2 + X_t$$

$$2\theta_* \cdot \mathbb{E}\left[\beta_2 a_1 \cdot \|\mathbf{w}_* - \mathbf{w}_t\|\right] \leq (\theta_* \cdot \beta_2 \cdot a_1)^2 + X_t$$

Thus we have,

$$\mathbb{E}\left[\text{Term 2}\right] \leq \left(\frac{a_4 + 1}{b} + \frac{(a_2^2 + 1)(b^2 - b)}{b^2} \right) \cdot \eta^2 \cdot X_t$$

$$+ \left(\frac{(\theta_* \cdot \beta_3)^2}{b} + \frac{(\theta_* \cdot \beta_2 \cdot a_1)^2 (b^2 - b)}{b^2} + \frac{\theta_*^2 (b \cdot \beta_2 + (b^2 - b) \cdot \beta_1^2)}{b^2} \right) \cdot \eta^2 \qquad (8)$$

$$\text{Term } 1 = 2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \mathbb{E}_{\mathbf{x}_{t_i},\alpha_{t_i}}\left[\left\langle \mathbf{w}_t - \mathbf{w}_*, \mathbf{1}_{y_{t_i}>\theta_*}\left(y_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i}\right)\mathbf{x}_{t_i}\right\rangle \Big| S_{t-1}\right]$$

$$= 2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*}\left(\alpha_{t_i}\xi_{t_i} + \text{ReLU}(\mathbf{w}_*^\top \mathbf{x}_{t_i}) - \mathbf{w}_t^\top \mathbf{x}_{t_i}\right) \times (\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{x}_{t_i}\Big| S_{t-1}\right]$$

Since $|\xi_{t_i}| \leq \theta_*$ it follows that $y_{t_i} > \theta_* \implies \mathbf{w}_*^\top \mathbf{x}_{t_i} > 0$. Hence,

$$= 2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*}\left(\alpha_{t_i}\xi_{t_i} + (\mathbf{w}_* - \mathbf{w}_t)^\top \mathbf{x}_{t_i}\right) \times (\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{x}_{t_i}\Big| S_{t-1}\right]$$

$$= -2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*}(\mathbf{w}_* - \mathbf{w}_t)^\top \cdot \mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top \cdot (\mathbf{w}_* - \mathbf{w}_t)\Big| S_{t-1}\right]$$

$$+ 2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*} \cdot \alpha_{t_i}\xi_{t_i} \cdot (\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{x}_{t_i}\Big| S_{t-1}\right]$$

$$\leq -2\frac{\eta}{b} \cdot \sum_{i=1}^{b} \lambda_{\min}\left(\mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*}\mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top\Big| S_{t-1}\right]\right)\|\mathbf{w}_t - \mathbf{w}_*\|^2$$

$$+ 2\frac{\eta}{b} \cdot \theta_* \cdot \sum_{i=1}^{b} \mathbb{E}\left[\beta(\mathbf{x}_{t_i}) \cdot \mathbf{1}_{y_{t_i}>\theta_*} \cdot \|\mathbf{x}_{t_i}\|\Big| S_{t-1}\right] \cdot \|\mathbf{w}_t - \mathbf{w}_*\|$$

$$\implies \mathbb{E}\left[\text{Term } 1\right] \leq -2\eta\lambda_1(\theta_*) \cdot X_t + 2\eta\theta_*\mathbb{E}\left[\beta_1 \cdot \|\mathbf{w}_t - \mathbf{w}_*\|\right]$$

$$\leq -2\eta\lambda_1(\theta_*) \cdot X_t + \eta\left(K(\theta_* \cdot \beta_1)^2 + \frac{1}{K}X_t\right)\mathbf{1}_{\theta_*>0} \tag{9}$$

In the last line above we used the following argument to write the upperbound in terms of $\lambda_1(\theta_*)$ as given in Definition 2 : We observe that for any i, $\mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*} \cdot \|\mathbf{x}_{t_i}\|\Big| S_{t-1}\right] \leq \mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>0} \cdot \|\mathbf{x}_{t_i}\|\Big| S_{t-1}\right]$. Also note that $y_{t_i} < \theta_* \implies \mathbf{w}_*^\top \mathbf{x}_{t_i} < 2\theta_*$. Hence for any test vector $\mathbf{v}$ we have, $\mathbf{v}^\top\left(\mathbb{E}\left[\left(\mathbf{1}_{y_{t_i}>\theta_*} - \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>2\theta_*}\right)\mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top\Big| S_{t-1}\right]\right)\mathbf{v} \geq 0$ and that in turn implies,

$$\lambda_{\min}\left(\mathbb{E}\left[\mathbf{1}_{y_{t_i}>\theta_*}\mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top\Big| S_{t-1}\right]\right) \geq \lambda_{\min}\left(\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>2\theta_*}\mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top\Big| S_{t-1}\right]\right) = \lambda_{\min}\left(\mathbb{E}\left[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i}>2\theta_*}\mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top\right]\right)$$

a) Case 1 : $\theta_* = 0$: Taking total expectations on both sides of equation 5 and setting $\theta_* = 0$ in the RHS of equations 7 and 9 we have,

$$X_{t+1} \leq \left(1 - 2\eta\lambda_1 + \frac{\eta^2}{b} \cdot (a_4 + a_2^2(b-1))\right)X_t \tag{10}$$

The above recursion is of the same form as analyzed in Lemma C.1 with $b_1 = 2\lambda_1, c_1 = \frac{a_4 + a_2^2(b-1)}{b}$ one can see that $c_1 > 0$ and hence convergence can be ensured if $c_1 > \frac{b_1^2\delta_0}{(1+\delta_0)^2}$ (With $\eta = \frac{b_1}{c_1(1+\delta_0)}$) for any positive δ_0

Thus from Lemma C.1 we have that given any $\epsilon > 0, \delta \in (0,1)$, $X_T \leq \epsilon^2 \cdot \delta$ for,

$$T = 1 + \frac{\log \frac{X_1}{\epsilon^2\delta}}{\log \frac{1}{\alpha}} \text{ with } \alpha = \left(1 - 2\eta\lambda_1 + \frac{\eta^2}{b} \cdot (a_4 + a_2^2(b-1))\right), \eta = \frac{2b\lambda_1}{(a_4 + a_2^2(b-1))(1+\delta_0)}$$

for a suitable $\delta_0 > 0$ as mentioned above.

b) Case 2 : $\theta_* > 0$: Taking total expectations on both sides of equation 5 and invoking the RHS of equations 8 and 9 we have,

$$X_{t+1} \leq \left(1 - 2\eta\lambda_1(\theta_*) + \frac{\eta}{K} + \frac{\eta^2}{b} \cdot ((1 + a_4) + (1 + a_2^2)(b - 1))\right)X_t$$
$$+ K\theta_*^2 \cdot \eta \cdot \beta_1^2 + \theta_*^2 \cdot \frac{\eta^2}{b} \cdot \left(\beta_3^2 + (\beta_2 \cdot a_1)^2 \cdot (b - 1) + (\beta_2 + (b - 1) \cdot \beta_1^2)\right) \tag{11}$$

Now we can invoke Lemma C.2 on the above recursion with the following identifications for the constants therein,

$$b_1 = 2\lambda_1(\theta_*) - \frac{1}{K}, c_1 = \frac{1 + a_4 + (1 + a_2^2)(b - 1)}{b}$$

$$c_3 = K_1\theta_*^2\beta_1^2, c_2 = \frac{\theta_*^2}{\beta_1}\left(\beta_3^2 + (\beta_2 \cdot a_1)^2 \cdot (b - 1) + (\beta_2 + (b - 1) \cdot \beta_1^2)\right)$$

Note that since K is so chosen that $2\lambda_1(\theta_*) > \frac{1}{K}$, we have $b_1 > 0$ and hence the conditions of Lemma C.2

Hence the smallest value of X_t (say $\epsilon^2 \cdot \delta$ for some $\epsilon > 0$ and $\delta \in (0, 1)$) that the Lemma C.2 guarantees to be attained, say at X_T is $\frac{c_3}{b_1} = \frac{K\theta_*^2\beta_1^2}{(2\lambda_1(\theta)-1/K)}$ for

$$\mathrm{T} = \mathcal{O}\left(\log\left[\frac{X_1}{\epsilon^2\delta - \left(\frac{\frac{c_2}{c_1} + \gamma \cdot \frac{c_3}{b_1}}{\gamma - 1}\right)}\right]\right)$$

when we choose $\eta = \frac{b_1}{\gamma c_1}$ for some $\gamma > \max\left(\frac{b_1^2}{c_1}, \frac{\epsilon^2\delta + \frac{c_2}{c_1}}{\epsilon^2\delta - \frac{c_3}{b_1}}\right)$. Now we can invoke Markov inequality to get what we set out to prove,

$$\mathbb{P}\left[\|\mathbf{w}_{\mathrm{T}} - \mathbf{w}_*\|^2\right] \leq \epsilon^2\right] \geq 1 - \delta$$

$\blacksquare$

IV. NEURO-TRON

Definition 3 (**A consistency condition**): Given $w \in \mathbb{Z}^+$ and $\mathcal{A} = \{\mathrm{A}_i \in \mathbb{R}^{r \times n} \mid i = 1, \ldots, w\}$ we define the set of neural nets,

$$\mathcal{N}_{\mathcal{A}} = \left\{ \mathbb{R}^n \ni \mathbf{z} \mapsto f_{\mathbf{w}}(\mathbf{z}) = \frac{1}{w} \sum_{k=1}^{w} \sigma\left(\langle \mathrm{A}_k^\top \mathbf{w}, \mathbf{z} \rangle \right) \in \mathbb{R} \mid \mathbf{w} \in \mathbb{R}^r \right\}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is mapping, $x \mapsto x\mathbf{1}_{x \geq 0} + \alpha x\mathbf{1}_{x \leq 0}$ for some $\alpha \geq 0$ We call a pair of matrices $(\mathbf{P}, \mathbf{M})$ of appropriate dimensions to be "$\mathcal{N}_{\mathcal{A}} - \text{consistent}$" if $\lambda_{\min}\left(\bar{\mathrm{A}} \mathbf{P} \mathbf{M}^\top \right) > 0$, where $\bar{\mathrm{A}} = \frac{1}{w} \sum_{k=1}^{w} \mathrm{A}_k$

Remark 4: Given any $\mathbf{P}$ which is PD and $\mathbf{M}$ which is full-rank, the pair $(\mathbf{P}, \mathbf{M})$ is always $\mathcal{N}_{\mathcal{A}} - \text{consistent}$ for any $\mathcal{A}$ s.t $\bar{\mathrm{A}} = \mathbf{M}$.

Remark 5: Note that there could be multiple nets which are functionally identical. For example suppose $f_1(x) := \frac{1}{2}(\sigma\{0, 2x\} + \sigma\{0.4x\})$ and $f_2(x) = \frac{1}{4}(\sigma\{0, 2x\} + \sigma\{0, 2x\} + \sigma\{0.4x\} + \sigma\{0.4x\})$. Then as functions $f_1 = f_2$ but as nets they are different, f_1 is a width 2 net while f_2 is a width 4 net. This also means that f_1 and f_2 cannot belong to the same class of nets $\mathcal{N}$ as defined above.

A. Defining the Neuro-Tron algorithm

Algorithm 3 Neuro-Tron

1: **Input:** A training set $\mathcal{T}$ of S samples of the form, $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,S}$ and a matrix $\mathbf{M} \in \mathbb{R}^{r \times n}$
2: **Input:** Oracle access to a class $\mathcal{N}_{\mathcal{A}}$ as given in Definition 3 s.t $\left(\tilde{\Sigma} := \frac{1}{S} \sum_{i=1}^{S} \mathbf{x}_i \mathbf{x}_i^\top, \mathbf{M} \right)$ is $\mathcal{N}_{\mathcal{A}} - \text{consistent}$
3: **Input :** A step-length η
4: **Input :** A starting point of $\mathbf{w}_1 \in \mathbb{R}^r$
5: **for** $t = 1, \ldots$ **do**
6:     The oracle replies with $f_{\mathbf{w}_t}(\mathbf{x}_i)$ when queried with $\mathbf{w}_t$ and $\mathbf{x}_i$
7:     $\mathbf{g}_t := \mathbf{M}\left(\frac{1}{S} \sum_{i=1}^{S} \underbrace{\left(y_i - f_{\mathbf{w}_t}(\mathbf{x}_i) \right)}_{\text{Current prediction error on the } i^{th} \text{ data}} \mathbf{x}_i \right)$
8:     $\mathbf{w}_{t+1} := \mathbf{w}_t + \eta \mathbf{g}_t$
9: **end for**

In the next Section IV-A1 we shall show how one can easily sample sets $\mathcal{A}$, the training data $\mathcal{T}$ and $\mathbf{M}$ so that w.h.p they satisfy the consistency condition as required to invoke the above algorithm.

Secondly, we note that when $w = 1$, A_1 is full-rank and $\tilde{\Sigma}$ is P.D then $\mathbf{M} = \mathrm{A}_1$ is a valid choice. Then one can interpret the Neuro-Tron algorithm as a (stochastic) G.D on the following unconventional surrogate risk,

$$\mathcal{R}^{\text{Surrogate}}(\mathbf{w}) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{in}} \left[\int_0^{(\mathrm{A}_1^\top \mathbf{w})^\top \mathbf{x}} \left(-y + \cdot \max\{0, z\} \right) \mathrm{d}z \right]$$

1) It is easy to sample valid examples on which Algorithm 3 can be invoked: Firstly we note that the no constraints are placed on the labels $\{y_i\}$ of $\mathcal{T}$ in defining the Algorithm 3. Using the set of inputs in $\mathcal{T}_{\mathbf{x}} = \{\mathbf{x}_i \in \mathbb{R}^n \mid \mathbf{1} = 1, \ldots, S\}$ we have $\tilde{\Sigma} := \frac{1}{S} \sum_{i=1}^{S} \mathbf{x}_i \mathbf{x}_i^\top$ positive definite with high probability (due to law of large number) as long as the true covariance matrix of $\mathbf{x}_i$ is so. We can as well make the set $\mathcal{T}_{\mathbf{x}}$ symmetric by including $-\mathbf{x}_i$ for every vector $\mathbf{x}_i \in \mathcal{T}_{\mathbf{x}}$. This inclusion keeps the sample covariance matrix intact since $\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top \right] = \mathbb{E}\left[(-\mathbf{x}_i)(-\mathbf{x}_i)^\top \right]$.

Secondly, for any $\mathbf{M} \in \mathbb{R}^{r \times n}$ which is full rank, any matrix $\mathcal{C} \in \mathbb{R}^{r \times n}$, any width w as an even number say $2k$ one can choose $\mathcal{A} := \{\mathbf{M} - k\mathcal{C}, \mathbf{M} - (k-1)\mathcal{C}, \ldots, \mathbf{M} - \mathcal{C}, \mathbf{M} + \mathcal{C}, \ldots, \mathbf{M} + k\mathcal{C}\}$. Then $\bar{\mathrm{A}} = \mathbf{M} = $ a full rank matrix and the pair $(\tilde{\Sigma}, \mathbf{M})$ would be $\mathcal{N}_{\mathcal{A}} - \text{consistent}$ as given in Definition 3 and thus one can invoke Algorithm 3 on the training data $\{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, S\}$ with $\mathcal{N}_{\mathcal{A}}$ and this $\mathbf{M}$.

Also one can easily construct a sampling scheme to generate such a matrix $\mathbf{M} \in \mathbb{R}^{r \times n}$, with $1 \leq r \leq n$, so that it is of full rank with high probability : Generate independent $\mathbf{g}_i = (g_i^1, \cdots, g_i^r)^\top \sim N(0, \mathrm{I}_{r \times r})$ and construct $\mathrm{G} = \sum_{i=1}^{k} \mathbf{g}_i \mathbf{g}_i^\top$. Then $\mathrm{G} \sim \mathbb{W}(\mathrm{I}, k)$ i.e. G follows a Wishart distribution with k degrees of freedom. Since I is

invertible, G has a full rank with probability 1 as long as $k \geq r$. One can use this G as a sub-matrix to complete it to a $\mathbf{M} \in \mathbb{R}^{r \times n}$ matrix which is also of full rank with probability 1.

B. Guarantees about Algorithm 3

Suppose one is given a training set $\mathcal{T}$ of S samples of the form, $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \mathbb{R}\}_{i=1,\ldots,S}$ which is s.t the set $\{\mathbf{x}_i \mid i = 1, \ldots, S\}$ is symmetric and contained within a ball of radius B about the origin. Define $\tilde{\Sigma} = \frac{1}{S}\sum_{i=1}^{S}\mathbf{x}_i\mathbf{x}_i^\top$. Suppose further that we have a choice of $\mathbf{M} \in \mathbb{R}^{r \times n}$ and $\mathcal{A} = \{A_k \in \mathbb{R}^{r \times n} \mid k = 1, \ldots, w\}$ s.t $(\tilde{\Sigma}, \mathbf{M})$ is $\mathcal{N}_\mathcal{A} -$ consistent as given in Definition 3.

Lemma 4.1: Given any fixed $\mathbf{w} \in \mathbb{R}^r$ (and hence a $f_\mathbf{w} \in \mathcal{N}_\mathcal{A}$), we define β s.t $\|\frac{1}{S}\sum_{i=1}^{S}\left(y_i - f_\mathbf{w}(\mathbf{x}_i)\right)\mathbf{M}\mathbf{x}_i\| \leq \beta$. Thus for Algorithm 3, invoked with $\mathcal{T}$, $\mathcal{N}_\mathcal{A}$ and $\mathbf{M}$ as given above, we have that at the $t^{th}-$iterate $f_{\mathbf{w}_t} \in \mathcal{N}_\mathcal{A}$ and we have the following relation between any two consecutive iterates of the Neuro-Tron,

$$
\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 - \|\mathbf{w}_t - \mathbf{w}\|^2 \\
&\leq 2\eta\beta\|\mathbf{w}_t - \mathbf{w}\| - \eta(1+\alpha)\lambda_1\|\mathbf{w}_t - \mathbf{w}\|^2 \\
&+ \eta^2\Big(\beta^2 + \beta \cdot (1+\alpha)B^2 \cdot \|\mathbf{w} - \mathbf{w}_t\| \cdot \|\mathbf{M}\| \cdot \frac{1}{w}\sum_{k=1}^{w}\|A_k\| + B^2\|\mathbf{M}\|^2\tilde{L}_S(t)\Big)
\end{aligned}
$$

where we have defined, $\tilde{L}_S(t) := \frac{1}{S}\sum_{i=1}^{S}|f_\mathbf{w}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)|^2$ and $\lambda_1 := \lambda_{\min}(\bar{A}\tilde{\Sigma}\mathbf{M}^\top)$

In above we have used the notation that for any matrix Z, let $\|Z\| := \sqrt{\lambda_{\max}(ZZ^\top)}$

We prove the above lemma in Appendix B.

Remark 6: In above there is no relationship enforced by-hand between the label associated with $\mathbf{x}_i$ and the label associated with $-\mathbf{x}_i$.

We note that compared to the single gate situation in Section II-D, the above dynamical system is significantly more challenging. Also note in the above recursion the appearance of terms linear in "$\beta\|\mathbf{w} - \mathbf{w}_t\|$". This is a very unique complication of this case that we consider here which we will see in the following theorem to be the crucial term that keeps track of the fact that our training data is agnostic.

Theorem 4.2: We continue in the same setup as in Lemma 4.1 and suppose we invoke Algorithm 3 with the choice of $\mathcal{T}$, $\mathbf{M}$ and $\mathcal{N}_\mathcal{A}$ as given in Lemma 4.1. Given *any* $f_\mathbf{w} \in \mathcal{N}_\mathcal{A}$ define the corresponding, $\theta_{\mathbf{w},\mathcal{T}} := \max_{i=1,\ldots,S}|y_i - f_\mathbf{w}(\mathbf{x}_i)|$. Then for Algorithm 3 initialized with any $\mathbf{w}_1$, the following holds,

- If $\theta_{\mathbf{w},\mathcal{T}} = 0$,

$$
\forall \epsilon > 0, \exists\, \eta \text{ and } T^* = \mathcal{O}\left(\log\frac{\sqrt{\|\mathbf{w}_1 - \mathbf{w}\|}}{\epsilon}\right) \text{ s.t we have}
$$

$$
\forall\, T > T^*,\ \|\mathbf{w}_T - \mathbf{w}\| < \epsilon \tag{12}
$$

- If $\theta_{\mathbf{w},\mathcal{T}} > 0$, we define the following constants,

$$
\overline{\|A\|} := \frac{1}{w}\sum_{k=1}^{w}\|A_k\| \text{ and } \overline{\|A^2\|} := \frac{1}{w}\sum_{k=1}^{w}\|A_k\|^2
$$

Then,

$$
\forall\mu > \sqrt{\frac{B\|\mathbf{M}\|}{(1+\alpha)\lambda_1}} \text{ and } \epsilon^2 > \frac{\theta_{\mathbf{w},\mathcal{T}}^2 \cdot \mu^2}{\frac{(1+\alpha)\lambda_1}{B\|\mathbf{M}\|} - \frac{1}{\mu^2}},\ \exists\, \gamma_* > 0 \text{ s.t } \forall\, \gamma > \gamma_*, \exists\, \eta
$$

$$
\&\ T = \mathcal{O}\left(\log\left[\frac{\|\mathbf{w}_1 - \mathbf{w}\|^2}{\epsilon^2 - \frac{\mu^2\theta_{\mathbf{w},\mathcal{T}}^2}{\gamma-1}\cdot\left(\left\{\frac{2+\mu^2(1+\alpha)\overline{\|A\|}}{2\mu^2((1+\alpha)B\|\mathbf{M}\|)^2\cdot\overline{\|A^2\|}+(1+\alpha)B\overline{\|A\|}}\right\} + \gamma\cdot\left\{\frac{\mu^2 B\|\mathbf{M}\|}{(1+\alpha)\mu^2\lambda_1 - B\|\mathbf{M}\|}\right\}\right)}\right]\right)
$$

$$
\text{s.t }\ \|\mathbf{w}_T - \mathbf{w}\| < \epsilon \tag{13}
$$

Remark 7: A precise value of γ_* can be given as,

$$\gamma_* = \max \left\{ \frac{2((1+\alpha)\mu^2\lambda_1 - \mathrm{B}\|\mathbf{M}\|)^2}{(1+\alpha)\mu^2\mathrm{B}^3\|\mathbf{M}\|^2(\overline{\|\mathbf{A}\|} + 2\mu^2(1+\alpha)\mathrm{B}\overline{\|\mathbf{A}\|^2})}, \frac{\epsilon^2 + \left(\frac{2+\mu^2(1+\alpha)\overline{\|\mathbf{A}\|}}{2\mu^2((1+\alpha)\mathrm{B}\|\mathbf{M}\|)^2 \cdot \overline{\|\mathbf{A}^2\|} + (1+\alpha)\mathrm{B}\overline{\|\mathbf{A}\|}} \right)}{\epsilon^2 - \left(\frac{\mu^2\mathrm{B}\|\mathbf{M}\|}{(1+\alpha)\mu^2\lambda_1 - \mathrm{B}\|\mathbf{M}\|} \right)} \right\}$$

The above is the first such poly-time convergence guarantee for an algorithm executing a search over finitely large neural nets when the data is not assumed to be realizable. Once the consistency condition is satisfied, the above guarantee holds for *every* $f_\mathbf{w} \in \mathcal{N}_\mathcal{A}$ with its corresponding $\theta_{\mathbf{w},\mathcal{T}}$ - the error with which $f_\mathbf{w}$ interpolates the training data $\mathcal{T}$. Hence the above theorem guarantees that there is always a way to set the step-length s.t the iterates will get $O(\theta_{\mathbf{w},\mathcal{T}})$ close in $2-$norm to the corresponding $\mathbf{w}$. From the above the following important special cases can be summarized separately,

*Theorem 4.3 (**Neuro-Tron near optimally solves the infinity-norm regression on the class $\mathcal{N}_\mathcal{A}$ for the data $\mathcal{T}$):***
 We continue with the same setup as given in Theorem 4.2. Define $\theta_{\min} := \inf_{f_\mathbf{w} \in \mathcal{N}_\mathcal{A}} \left(\max_{i=1,\ldots,S} |y_i - f_\mathbf{w}(\mathbf{x}_i)| \right)$ and $\mathcal{W}_{\min} := \operatorname{arginf}_{\mathbf{w}|f_\mathbf{w} \in \mathcal{N}_\mathcal{A}} \left(\max_{i=1,\ldots,S} |y_i - f_\mathbf{w}(\mathbf{x}_i)| \right)$. Then for any initialization $\mathbf{w}_1$ we have the following guarantees about the iterates of Algorithm 3,

- If $\theta_{\min} = 0$ then $|\mathcal{W}_{\min}| = 1$ and for $\mathbf{w}_{\min} \in \mathcal{W}_{\min}$ it holds that,

$$\forall \epsilon > 0, \exists \, \eta \text{ and } \mathrm{T}^* = \mathcal{O} \left(\log \frac{\sqrt{\|\mathbf{w}_1 - \mathbf{w}\|}}{\epsilon} \right) \text{ s.t we have}$$

$$\forall \, T > \mathrm{T}^*, \ \|\mathbf{w}_\mathrm{T} - \mathbf{w}_{\min}\| < \epsilon \tag{14}$$

- If $\theta_{\min} > 0$ then,

$$\exists \, \eta \ \& \ \mathrm{T} \text{ s.t } \|\mathbf{w}_\mathrm{T} - \mathbf{w}_{\min}\| = \mathcal{O}(\theta_{\min}), \ \forall \, \mathbf{w}_{\min} \in \mathcal{W}_{\min} \tag{15}$$

Thus Algorithm 3 has iterates which give an approximate solution in polynomial time to the infinity norm neural regression problem stated in equation 2 albeit for certain special choices of $\mathcal{T}$ $\&\{\mathrm{A}_i\}$. Thus in the valid situations the algorithm finds an approximation to the pointwise optimal interpolant of the given data in the given class of nets.

The above theorem also says that (a) the pointwise optimal interpolant is unique if there is an exact interpolant in the neural class we are searching over i.e when $\theta_{\min} = 0$. Further (b) when $\theta_{\min} > 0$, then the existence of the kind of $\mathbf{w}_\mathrm{T}$ as guaranteed in equation 15 implies that $\operatorname{diam}(\mathcal{W}_{\min}) = O(\theta_{\min})$ i.e we have alongside also proven a bound on the size of the set in which all the global minimas of the optimization problem 2 can be found for choices of $\mathcal{T}$ $\&\{\mathrm{A}_i\}$ when the above theorem applies.

The proof of Theorem 4.3 follows just redoing the calculation inside the proof of Theorem 4.2 with $\mathbf{w}$ replaced by any element of the set $\mathcal{W}_{\min}$ and $\theta_{\mathbf{w},\mathcal{T}}$ replaced by $\theta_{\min}$. The uniqueness of the global minima when $\theta_{\min} = 0$, follows from observing that if there were two distinct elements in the corresponding $\mathcal{W}_{\min}$ then the first case of Theorem 4.2 applied separately on each of them as the "$\mathbf{w}$" there would imply that the algorithm is simultaneously getting arbitrarily close to each of them and that is a contradiction.

As an interesting example of the above consider the situation where the training data is 2 points $\mathcal{T} = \{(1,0), (-1,0)\}$. When Algorithm 3 is run on this training data for $\mathbf{M} = 1$ $\& \, \mathcal{N} = \{\mathbb{R} \ni z \mapsto \max\{0, w \cdot z\} \in \mathbb{R}\}$ then the algorithm is found to be converging to the solution $w = 0$ which is an exact interpolant of the training data. Its easy to see that the above theorem predicts the same.

Remark 8: We note that the architecture or the generalized patch matrices of nets in $\mathcal{N}_\mathcal{A}$ or the choice of $\mathbf{w}$ (and the value of $\theta_{\mathbf{w},\mathcal{T}}$) only enter the above theorem's guarantees and are not a part of Algorithm 3's specification or its inputs like the training data $\mathcal{T}$

C. Proof of Theorem 4.2

Proof: For brevity we shall denote $\theta_{\mathbf{w},\mathcal{T}}$ as just θ. Given the definition of $\mathbf{w}$ and θ it follows that $\exists \xi_i \in \mathbb{R}, \forall i = 1, \ldots, S$ s.t, $\|\frac{1}{S}\sum_{i=1}^{S}\left(y_i - f_{\mathbf{w}}(\mathbf{x}_i)\right)\mathbf{M}\mathbf{x}_i\| = \frac{1}{S}\|\sum_{i=1}^{S}\xi_i \mathbf{M}\mathbf{x}_i\| \leq \frac{1}{S}\sum_{i=1}^{S}|\xi_i|\|\mathbf{M}\mathbf{x}_i\| \leq \theta B\|\mathbf{M}\|$

Further for the $t^{th}-$iterate we also have,

$$
\begin{aligned}
\frac{1}{S}\sum_{i=1}^{S}|f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)|^2 &= \frac{1}{S}\sum_{i=1}^{S}\left|\frac{1}{w}\sum_{k=1}^{w}(\sigma(\mathbf{w}^\top A_k \mathbf{x}_i) - \sigma(\mathbf{w}_t^\top A_k \mathbf{x}_i))\right|^2 \\
&\leq \frac{1}{Sw^2}\sum_{i=1}^{S}\left|\sum_{k=1}^{w}\sigma(\mathbf{w}^\top A_k \mathbf{x}_i) - \sigma(\mathbf{w}_t^\top A_k \mathbf{x}_i)\right|^2
\end{aligned}
$$

Now recall that by Cauchy-Schwarz inequality we have for any set of numbers $\{z_i \mid i = 1, \ldots, w\}$,

$$
\left(\sum_{i=1}^{w}z_i\right)^2 = \left(\sum_{i=1}^{w}1 \cdot z_i\right)^2 \leq \|(1,\ldots,1)\|^2\|(z_1,\ldots,z_w)\|^2 \leq w\sum_{i=1}^{w}z_i^2
$$

Thus the above implies,

$$
\begin{aligned}
\frac{1}{S}\sum_{i=1}^{S}|f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)|^2 &\leq \frac{(1+\alpha)^2}{Sw}\sum_{i=1}^{S}\sum_{k=1}^{w}\|\mathbf{w} - \mathbf{w}_t\|^2\|A_k\|^2\|\mathbf{x}_i\|^2 \\
&\leq (1+\alpha)^2 \cdot B^2 \cdot \|\mathbf{w} - \mathbf{w}_t\|^2 \cdot \left(\frac{1}{w}\sum_{k=1}^{w}\|A_k\|^2\right)
\end{aligned}
$$

So we can invoke Lemma 4.1 between consecutive iterates of Neuro-Tron with $\mathbf{w}$ as above, $\beta = \theta B\|\mathbf{M}\|$ and $\tilde{L}_S(t) \leq (1+\alpha)^2 \cdot B^2 \cdot \|\mathbf{w} - \mathbf{w}_t\|^2 \cdot \left(\frac{1}{w}\sum_{k=1}^{w}\|A_k\|^2\right)$ and we get,

$$
\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 - \|\mathbf{w}_t - \mathbf{w}\|^2 \\
&\leq 2\eta\beta\|\mathbf{w}_t - \mathbf{w}\| - \eta(1+\alpha)\lambda_1\|\mathbf{w}_t - \mathbf{w}\|^2 \\
&\quad + \eta^2\left(\beta^2 + \beta \cdot (1+\alpha)B^2 \cdot \|\mathbf{w} - \mathbf{w}_t\| \cdot \|\mathbf{M}\| \cdot \frac{1}{w}\sum_{k=1}^{w}\|A_k\| + B^2\|\mathbf{M}\|^2\tilde{L}_S(t)\right) \\
&\leq \theta B\|\mathbf{M}\| \cdot 2\eta\|\mathbf{w}_t - \mathbf{w}\| - \eta(1+\alpha)\lambda_1\|\mathbf{w}_t - \mathbf{w}\|^2 \\
&\quad + \eta^2\left((\theta B\|\mathbf{M}\|)^2 + \theta B\|\mathbf{M}\| \cdot (1+\alpha)B^2 \cdot \|\mathbf{M}\| \cdot \left(\frac{1}{w}\sum_{k=1}^{w}\|A_k\|\right) \cdot \|\mathbf{w}_t - \mathbf{w}\|\right) \\
&\quad + \eta^2 B^2\|\mathbf{M}\|^2 \cdot (1+\alpha)^2 \cdot B^2 \cdot \|\mathbf{w}_t - \mathbf{w}\|^2 \cdot \left(\frac{1}{w}\sum_{k=1}^{w}\|A_k\|^2\right)
\end{aligned}
$$

Now we define the following variables,

$$
\Delta_t = \|\mathbf{w}_t - \mathbf{w}\|^2
$$

$$
a_1 = (1+\alpha)\lambda_1, \ a_2 = B^4\|\mathbf{M}\|^2(1+\alpha)^2 \cdot \left(\frac{1}{w}\sum_{k=1}^{w}\|A_k\|^2\right)
$$

$$
a_3 = B^3\|\mathbf{M}\|^2(1+\alpha) \cdot \left(\frac{1}{w}\sum_{k=1}^{w}\|A_k\|\right), \ a_5 = \frac{a_4^2}{4} = B^2\|\mathbf{M}\|^2
$$

Hence we can relate the updates of the algorithm as,

$$\Delta_{t+1} \le (1 - \eta a_1 + \eta^2 a_2)\Delta_t + (\eta^2 a_3 + \eta a_4)\sqrt{\Delta_t}\theta + \eta^2(a_5\theta^2) \tag{16}$$

If $\theta = 0$ then we choose $\eta = \frac{a_1}{\gamma \cdot a_2}$ and observe that for any $\gamma > \max\{1, \frac{a_1^2}{a_2}\}$, $(1 - \eta a_1 + \eta^2 a_2) \in (0, 1)$ and thus we get from above the geometric convergence guarantee given in equation 12.

For $\theta > 0$, using the A.M-G.M inequality with a parameter μ, on the middle term of equation 16 above we get,

$$\Delta_{t+1} \le \left(1 - \eta\left(a_1 - \frac{a_4}{2\mu^2}\right) + \eta^2\left(a_2 + \frac{a_3}{2\mu^2}\right)\right)\Delta_t + \eta^2\left(\frac{a_3\mu^2\theta^2}{2} + a_5\theta^2\right) + \eta\left(\frac{a_4\theta^2\mu^2}{2}\right)$$

The above is of the same form as the discrete dynamical system that has been analyzed in Lemma C.2 in Appendix C and the claimed guarantee in equation 13 immediately follows on substituting the values.

∎

V. Conclusion

In this work we have initiated a number of directions of investigation towards understanding the trainability of finite sized nets while making minimal assumptions about the distribution of the data. A lot of open questions emanate from here which await answers. Of them we would like to particularly emphasize the issue of seeking a generalization of the results of Section III to (a) the algorithm being standard S.G.D and (b) the nets being single filter depth 2 nets as given in Definition 1. The next more general class of nets to which one could consider generalizing the results of both sections III and IV would be the *multi-filter nets of depth* 2 which would have the same structure as in Definition 1 but there could be different weights say $\mathbf{w}_i$ at the $i^{th}-$gate and all of whose ground truth values would have to be discovered in tandem by the algorithm. Such results if obtained could possibly constitute significantly towards the outstanding open question of finding regimes of poly-times training over nets with constant number of gates and with the data being non-realizable.

References

[1] Allen-Zhu, Z., Li, Y.: What can resnet learn efficiently, going beyond kernels? In: Advances in Neural Information Processing Systems, pp. 9015–9025 (2019)

[2] Allen-Zhu, Z., Li, Y., Liang, Y.: Learning and generalization in overparameterized neural networks, going beyond two layers. In: Advances in neural information processing systems, pp. 6155–6166 (2019)

[3] Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. In: International Conference on Machine Learning, pp. 242–252 (2019)

[4] Allen-Zhu, Z., Li, Y., Song, Z.: On the convergence rate of training recurrent neural networks. In: Advances in Neural Information Processing Systems, pp. 6673–6685 (2019)

[5] Arora, R., Basu, A., Mianjy, P., Mukherjee, A.: Understanding deep neural networks with rectified linear units. arXiv preprint arXiv:1611.01491 (2016)

[6] Arora, S., Du, S., Hu, W., Li, Z., Wang, R.: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In: International Conference on Machine Learning, pp. 322–332 (2019)

[7] Arora, S., Du, S.S., Hu, W., Li, Z., Salakhutdinov, R.R., Wang, R.: On exact computation with an infinitely wide neural net. In: Advances in Neural Information Processing Systems, pp. 8139–8148 (2019)

[8] Arora, S., Du, S.S., Li, Z., Salakhutdinov, R., Wang, R., Yu, D.: Harnessing the power of infinitely wide deep nets on small-data tasks. arXiv preprint arXiv:1910.01663 (2019)

[9] Chizat, L., Bach, F.: On the global convergence of gradient descent for over-parameterized models using optimal transport. In: Advances in neural information processing systems, pp. 3036–3046 (2018)

[10] Diakonikolas, I., Goel, S., Karmalkar, S., Klivans, A.R., Soltanolkotabi, M.: Approximation schemes for relu regression. In: Conference on Learning Theory (2020)

[11] Diakonikolas, I., Kontonis, V., Tzamos, C., Zarifis, N.: Learning halfspaces with massart noise under structured distributions. arXiv preprint arXiv:2002.05632 (2020)

[12] Du, S., Lee, J.: On the power of over-parametrization in neural networks with quadratic activation. In: International Conference on Machine Learning, pp. 1329–1338 (2018)

[13] Du, S.S., Lee, J.D., Li, H., Wang, L., Zhai, X.: Gradient descent finds global minima of deep neural networks (2018)

[14] Frei, S., Cao, Y., Gu, Q.: Agnostic learning of a single neuron with gradient descent. arXiv preprint arXiv:2005.14426 (2020)

[15] Freund, Y., Schapire, R.E.: Large margin classification using the perceptron algorithm. Machine learning **37**(3), 277–296 (1999)

[16] Goel, S., Kanade, V., Klivans, A., Thaler, J.: Reliably learning the relu in polynomial time. arXiv preprint arXiv:1611.10258 (2016)

[17] Goel, S., Klivans, A.: Learning depth-three neural networks in polynomial time. arXiv preprint arXiv:1709.06010 (2017)

[18] Goel, S., Klivans, A., Meka, R.: Learning one convolutional layer with overlapping patches. arXiv preprint arXiv:1802.02547 (2018)

[19] Huang, J., Yau, H.T.: Dynamics of deep neural networks and neural tangent hierarchy. arXiv preprint arXiv:1909.08156 (2019)

[20] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. In: Advances in neural information processing systems, pp. 8571–8580 (2018)

[21] 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)

[22] Kalan, S.M.M., Soltanolkotabi, M., Avestimehr, A.S.: Fitting relus via sgd and quantized sgd. In: 2019 IEEE International Symposium on Information Theory (ISIT), pp. 2469–2473. IEEE (2019)

[23] Kawaguchi, K., Huang, J.: Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In: 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 92–99. IEEE (2019)

[24] 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)

[25] Lee, J., Bahri, Y., Novak, R., Schoenholz, S.S., Pennington, J., Sohl-Dickstein, J.: Deep neural networks as gaussian processes (2017)

[26] Li, Z., Wang, R., Yu, D., Du, S.S., Hu, W., Salakhutdinov, R., Arora, S.: Enhanced convolutional neural tangent kernels. arXiv preprint arXiv:1911.00809 (2019)

[27] Mukherjee, A., Muthukumar, R.: Guarantees on learning depth-2 neural networks under a data poisoning attack. arXiv preprint arXiv:2005.01699 (2020)

[28] Neal, R.M.: Priors for infinite networks. In: Bayesian Learning for Neural Networks, pp. 29–53. Springer (1996)

[29] Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. IEEE transactions on neural networks **3 5**, 683–97 (1992)

[30] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)

[31] Soltanolkotabi, M.: Learning relus via gradient descent. In: Advances in neural information processing systems, pp. 2007–2017 (2017)

[32] Su, L., Yang, P.: On learning over-parameterized neural networks: A functional approximation perspective. In: Advances in Neural Information Processing Systems, pp. 2637–2646 (2019)

[33] Wei, C., Lee, J.D., Liu, Q., Ma, T.: Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In: Advances in Neural Information Processing Systems, pp. 9709–9721 (2019)

[34] Wu, X., Du, S.S., Ward, R.: Global convergence of adaptive gradient methods for an over-parameterized neural network. arXiv preprint arXiv:1902.07111 (2019)

[35] Zou, D., Cao, Y., Zhou, D., Gu, Q.: Stochastic gradient descent optimizes over-parameterized deep relu networks. arXiv preprint arXiv:1811.08888 (2018)

[36] Zou, D., Gu, Q.: An improved analysis of training over-parameterized deep neural networks. In: Advances in Neural Information Processing Systems, pp. 2053–2062 (2019)

APPENDIX A
PROOFS OF SECTION II-D

A. Proof of Lemma 2.1

Proof: We observe that,

$$\|\mathbf{w}_t - \mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|^2 = \|\mathbf{w}_t - \mathbf{w}\|^2 - \|\Big(\mathbf{w}_t + \frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\mathbf{x}_i\Big) - \mathbf{w}\|^2$$

$$= -\frac{2}{S}\sum_{i=1}^{S}\Big\langle\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\mathbf{x}_i, \mathbf{w}_t - \mathbf{w}\Big\rangle - \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2$$

$$= \frac{2}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{w},\mathbf{x}_i\rangle - \langle\mathbf{w}_t,\mathbf{x}_i\rangle\Big) - \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \qquad (17)$$

Analyzing the first term in the RHS above we get,

$$\frac{2}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{w},\mathbf{x}_i\rangle - \langle\mathbf{w}_t,\mathbf{x}_i\rangle\Big)$$

$$= \frac{2}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w},\mathbf{x}_i\rangle) + \sigma(\langle\mathbf{w},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{w},\mathbf{x}_i\rangle - \langle\mathbf{w}_t,\mathbf{x}_i\rangle\Big)$$

$$= \frac{2}{S}\sum_{i=1}^{S}\Big\langle\Big(y_i - \sigma(\langle\mathbf{w},\mathbf{x}_i\rangle)\Big)\mathbf{x}_i, \mathbf{w} - \mathbf{w}_t\Big\rangle + \frac{2}{S}\sum_{i=1}^{S}\Big(\sigma(\langle\mathbf{w},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{x}_i,\mathbf{w}\rangle - \langle\mathbf{x}_i,\mathbf{w}_t\rangle\Big)$$

$$\geq -2\eta W + \frac{2}{S}\sum_{i=1}^{S}\Big(\sigma(\langle\mathbf{w},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{x}_i,\mathbf{w}\rangle - \langle\mathbf{x}_i,\mathbf{w}_t\rangle\Big)$$

In the first term above we have invoked the definition of η and W given in the lemma. Further since we are given that σ is non-decreasing and $L-$Lipschitz, we have for the second term in the RHS above,

$$\frac{2}{S}\sum_{i=1}^{S}\Big(\sigma(\langle\mathbf{w},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{x}_i,\mathbf{w}\rangle - \langle\mathbf{x}_i,\mathbf{w}_t\rangle\Big)$$

$$\geq \frac{2}{SL}\sum_{i=1}^{S}\Big(\sigma(\langle\mathbf{w},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)^2 =: \frac{2}{L}\tilde{L}_S(h_t)$$

Thus together we have,

$$\frac{2}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle\mathbf{w}_t,\mathbf{x}_i\rangle)\Big)\Big(\langle\mathbf{w},\mathbf{x}_i\rangle - \langle\mathbf{w}_t,\mathbf{x}_i\rangle\Big) \geq -2\eta W + \frac{2}{L}\tilde{L}_S(h_t) \qquad (18)$$

Now we look at the second term in the RHS of equation 21 and that gives us,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 = \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) + \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2$$

$$\leq \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 + 2\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \times \|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|$$

$$+ \|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2$$

$$\leq \eta^2 + 2\eta\|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| + \|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \quad (19)$$

Now by Jensen's inequality we have,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \leq \frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)^2 = \tilde{L}_S(h_t)$$

And we have from the definition of L and W,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \leq \frac{L}{S}\sum_{i=1}^{S}\|\mathbf{w} - \mathbf{w}_t\| \leq L \times W$$

Substituting the above two into the RHS of equation 19 we have,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \leq \eta^2 + 2\eta LW + \tilde{L}_S(h_t) \quad (20)$$

Now we substitute equations 18 and 20 into equation 21 to get,

$$\|\mathbf{w}_t - \mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|^2 \geq \Big(-2\eta W + \frac{2}{L}\tilde{L}_S(h_t)\Big) - (\eta^2 + 2\eta LW + \tilde{L}_S(h_t))$$

The above simplifies to the inequality we claimed in the lemma i.e,

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 \leq \|\mathbf{w}_t - \mathbf{w}\|^2 - \Big(\frac{2}{L} - 1\Big)\tilde{L}_S(h_t) + \Big(\eta^2 + 2\eta W(L+1)\Big)$$

∎

B. Proof of Theorem 2.2

Proof: The equation defining the labels in the data-set i.e $y_i = \sigma(\langle \mathbf{w}_*, \mathbf{x}_i\rangle) + \xi_i$ with $|\xi_i| \leq \theta$ along with our assumption that, $\|\mathbf{x}_i\| \leq 1$ implies that , $\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \leq \theta$. Thus we can invoke the above Lemma 2.1 between the tth and the $t+1^{th}$ iterate with $\eta = \theta$ and W as defined there to get,

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - \Big[\Big(\frac{2}{L} - 1\Big)\tilde{L}_S(h_t) - (\theta^2 + 2\theta W(L+1))\Big]$$

If $\tilde{L}_S(h_t) \geq \frac{L}{2-L}\Big(\epsilon + (\theta^2 + 2\theta W(L+1))\Big)$ then, $\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 \leq \|\mathbf{w}_t - \mathbf{w}\|^2 - \epsilon$. Thus if the above lowerbound on $\tilde{L}_s(h_t)$ holds in the tth step then at the start of the $(t+1)^{th}$ step we still satisfy, $\|\mathbf{w}_{t+1} - \mathbf{w}\| \leq W$. Since the iterations start with $\mathbf{w}_1 = 0$, in the first step we can choose $W = \|\mathbf{w}_*\|$. Thus in at most $\frac{\|\mathbf{w}\|}{\epsilon}$ steps of the above kind we can have a decrease in distance of the iterate to $\mathbf{w}$.

Thus in at most $T = \frac{\|\mathbf{w}\|}{\epsilon}$ steps we have attained,

$$\tilde{L}_S(h_T) = \frac{1}{S} \sum_{i=1}^{S} \Big(\sigma(\langle \mathbf{w}_T, x_i \rangle) - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) \Big)^2 < \frac{L}{2-L} \Big(\epsilon + (\theta^2 + 2\theta W(L+1)) \Big)$$

And that proves the theorem we wanted. ∎

C. Proof of Theorem 2.3

Proof: Let the true empirical risk at the $T^{th}-$iterate be defined as,

$$L_S(h_T) = \frac{1}{S} \sum_{i=1}^{S} \Big(\sigma(\langle \mathbf{w}_T, x_i \rangle) - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) - \xi_i \Big)^2$$

Then it follows that,

$$\tilde{L}_S(h_T) - L_S(h_T) = \frac{1}{S} \sum_{i=1}^{S} \Big(\sigma(\langle \mathbf{w}_T, x_i \rangle) - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) \Big)^2 - \frac{1}{S} \sum_{i=1}^{S} \Big(\sigma(\langle \mathbf{w}_T, x_i \rangle) - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) - \xi_i \Big)^2$$

$$= \frac{1}{S} \sum_{i=1}^{S} \xi_i \Big(-\xi_i + 2\sigma(\langle \mathbf{w}_T, x_i \rangle) - 2\sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) \Big) = -\frac{1}{S} \sum_{i=1}^{S} \xi_i^2 + \frac{2}{S} \sum_{i=1}^{S} \xi_i \Big(\sigma(\langle \mathbf{w}_T, x_i \rangle) - \sigma(\langle \mathbf{w}_*, \mathbf{x}_i \rangle) \Big)$$

By the assumption of ξ_i being an unbiased noise the second term vanishes when we compute, $\mathbb{E}_{\{(\mathbf{x}_i, \xi_i)\}_{i=1,\dots S}} \Big[\tilde{L}_S(h_T) - L_S(h_T) \Big]$ Thus we are led to,

$$\mathbb{E}_{\{(\mathbf{x}_i, \xi_i)\}_{i=1,\dots S}} \Big[\tilde{L}_S(h_T) - L_S(h_T) \Big] = -\frac{1}{m} \mathbb{E}_{\{\xi_i\}_{i=1,\dots S}} \Big[\sum_{i=1}^{m} \xi_i^2 \Big] = -\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\{\xi_i\}} \Big[\xi_i^2 \Big] = -\mathbb{E}_\xi[\xi^2]$$

For $T = \frac{\|\mathbf{w}\|}{\epsilon}$, we invoke the upperbound on $\tilde{L}_S(h_T)$ from the previous theorem and we can combine it with the above to say,

$$\mathbb{E}_{\{(\mathbf{x}_i, \xi_i)\}_{i=1,\dots S}} \Big[L_S(h_T) \Big] \leq \mathbb{E}_\xi[\xi^2] + \frac{L}{2-L} \Big(\epsilon + (\theta^2 + 2\theta W(L+1)) \Big)$$

This proves the theorem we wanted. ∎

APPENDIX B

PROOF OF LEMMA 4.1 FROM SECTION IV

Proof: We have between consecutive iterates of Algorithm 3

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 = \|\mathbf{w}_t + \eta\mathbf{g}_t - \mathbf{w}\|^2 = \|\mathbf{w}_t - \mathbf{w}\|^2 + \underbrace{2\eta\langle\mathbf{w}_t - \mathbf{w}, \mathbf{g}_t\rangle}_{\text{Term1}} + \underbrace{\eta^2\|\mathbf{g}_t\|^2}_{\text{Term2}} \tag{21}$$

a) **Upperbounding** Term 1:

$$\text{Term1} = 2\eta\langle\mathbf{w}_t - \mathbf{w}, \mathbf{g}_t\rangle = 2\eta\Big\langle\mathbf{w}_t - \mathbf{w}, \mathbf{M}\Big(\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{x}_i\Big)\Big\rangle$$

$$= \frac{2\eta}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\langle\mathbf{w}_t - \mathbf{w}, \mathbf{M}\mathbf{x}_i\rangle$$

$$= \frac{2\eta}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}}(\mathbf{x}_i)\Big)\langle\mathbf{w}_t - \mathbf{w}, \mathbf{M}\mathbf{x}_i\rangle$$

$$+ \frac{2\eta}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\langle\mathbf{w}_t - \mathbf{w}, \mathbf{M}\mathbf{x}_i\rangle$$

$$\leq 2\eta\|\mathbf{w}_t - \mathbf{w}\|\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|$$

$$+ \frac{2\eta}{wS}\sum_{k=1}^{w}\Big[\sum_{i=1}^{S}\Big(\sigma(\langle\mathbf{A}_k^\top\mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle\mathbf{A}_k^\top\mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\langle\mathbf{w}_t - \mathbf{w}, \mathbf{M}\mathbf{x}_i\rangle\Big]$$

We invoke the definition of β for the first term in the RHS above and Lemma B.1 on the second term to get,

$$\text{Term1} = 2\eta\langle\mathbf{w}_t - \mathbf{w}, \mathbf{g}_t\rangle$$

$$\leq 2\eta\beta\|\mathbf{w}_t - \mathbf{w}\| + \frac{2\eta}{wS}\cdot\frac{S(1+\alpha)}{2}\sum_{k=1}^{w}\Big[(\mathbf{w} - \mathbf{w}_t)^\top(\mathbf{A}_k\tilde{\Sigma}\mathbf{M}^\top)(\mathbf{w}_t - \mathbf{w})\Big]$$

$$\leq 2\eta\beta\|\mathbf{w}_t - \mathbf{w}\| - \eta(1+\alpha)\lambda_1\|\mathbf{w}_t - \mathbf{w}\|^2 \tag{22}$$

In the last step above we have invoked the definition of λ_1.

b) **Upperbounding** Term 2:

$$\text{Term2} = \eta^2\|\mathbf{g}_t\|^2 = \eta^2\|\mathbf{M}\Big(\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{x}_i\Big)\|^2$$

$$= \eta^2\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}}(\mathbf{x}_i) + f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}\Big)(\mathbf{M}\mathbf{x}_i)\|^2$$

$$\leq \eta^2\Big[\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|^2 + \|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|^2\Big]$$

$$+ 2\eta^2\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - f_{\mathbf{w}}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|\cdot\|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|$$

$$\leq \eta^2\Big[\beta^2 + 2\beta\|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|\Big]$$

$$+ \eta^2\|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|^2 \tag{23}$$

Consider 2 vectors $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^S$ s.t $\mathbf{p}_{1,i} = |f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)|$ and $\mathbf{p}_{2,i} = \|\mathbf{M}\mathbf{x}_i\|$. Further this implies that $\|\mathbf{p}_2\|^2 = \sum_{i=1}^{S}\|\mathbf{M}\mathbf{x}_i\|^2 \leq SB^2\|\mathbf{M}\|^2$. Then we have,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\| \leq \frac{1}{S}\sum_{i=1}^{S}|\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)|\|\mathbf{M}\mathbf{x}_i\| \leq \frac{1}{S}\mathbf{p}_1^\top\mathbf{p}_2$$

$$\implies \|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\|^2 \leq \frac{1}{S^2}\cdot\|\mathbf{p}_1\|^2\cdot\|\mathbf{p}_2\|^2 \leq \frac{B^2\|\mathbf{M}\|^2}{S}\sum_{i=1}^{S}|f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)|^2 \quad (24)$$

Also,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\| = \|\frac{1}{wS}\sum_{k=1}^{w}\sum_{i=1}^{S}\Big(\sigma(\mathbf{w}^\top A_k\mathbf{x}_i) - \sigma(\mathbf{w}_t^\top A_k\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\| \quad (25)$$

Because of the parity symmetry of the set, $\{\mathbf{x}_i\}$ we have,

$$2\sum_{i=1}^{S}\Big(\sigma(\mathbf{w}^\top A_k\mathbf{x}_i) - \sigma(\mathbf{w}_t^\top A_k\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i$$

$$= \sum_{i=1}^{S}\Big(\sigma(\mathbf{w}^\top A_k\mathbf{x}_i) - \sigma(\mathbf{w}_t^\top A_k\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i - \Big(\sigma(-\mathbf{w}^\top A_k\mathbf{x}_i) - \sigma(-\mathbf{w}_t^\top A_k\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i$$

$$= \sum_{i=1}^{S}\Big(\sigma(\mathbf{w}^\top A_k\mathbf{x}_i) - \sigma(-\mathbf{w}^\top A_k\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i + \sum_{i=1}^{S}\Big(\sigma(-\mathbf{w}_t^\top A_k\mathbf{x}_i) - \sigma(\mathbf{w}_t^\top A_k\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i$$

$$= (1+\alpha)\sum_{i=1}^{S}(\mathbf{w}^\top A_k\mathbf{x}_i)\mathbf{M}\mathbf{x}_i - (\mathbf{w}_t^\top A_k\mathbf{x}_i)\mathbf{M}\mathbf{x}_i$$

Substituting the above into equation 25 we have,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)\Big)\mathbf{M}\mathbf{x}_i\| = \frac{1+\alpha}{2wS}\|\sum_{k=1}^{w}\sum_{i=1}^{S}(\mathbf{w} - \mathbf{w}_t)^\top A_k\mathbf{x}_i(\mathbf{M}\mathbf{x}_i)\|$$

$$\leq \frac{1+\alpha}{2wS}\cdot SB^2\cdot\|\mathbf{w} - \mathbf{w}_t\|\cdot\|\mathbf{M}\|\cdot\sum_{k=1}^{w}\|A_k\| \quad (26)$$

Substituting the above and equation 24 into equation 23 we get,

$$\text{Term2} \leq \eta^2\Big(\beta^2 + 2\beta\cdot\frac{1+\alpha}{2w}\cdot B^2\cdot\|\mathbf{w} - \mathbf{w}_t\|\cdot\|\mathbf{M}\|\cdot\sum_{k=1}^{w}\|A_k\| + \frac{B^2\|\mathbf{M}\|^2}{S}\sum_{i=1}^{S}|f_{\mathbf{w}}(\mathbf{x}_i) - f_{\mathbf{w}_t}(\mathbf{x}_i)|^2\Big) \quad (27)$$

We recall the definition of $\tilde{L}_S(t)$ and substitute the above and equation 22 into equation 21 to get,

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|^2 - \|\mathbf{w}_t - \mathbf{w}\|^2 \leq 2\eta\beta\|\mathbf{w}_t - \mathbf{w}\| - \eta(1+\alpha)\lambda_1\|\mathbf{w}_t - \mathbf{w}\|^2$$

$$+ \eta^2\Big(\beta^2 + \beta\cdot(1+\alpha)B^2\cdot\|\mathbf{w} - \mathbf{w}_t\|\cdot\|\mathbf{M}\|\cdot\frac{1}{w}\sum_{k=1}^{w}\|A_k\| + B^2\|\mathbf{M}\|^2\tilde{L}_S(t)\Big)$$

Rearranging the above we get the lemma we set out to prove. ∎

Lemma B.1: Suppose $\mathcal{T} := \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, S\}$ is a set if tuples such that the set $\mathcal{T}_\mathbf{x} := (\mathbf{x}_i)_{i=1,\ldots,S}$ is invariant under replacing each $\mathbf{x}_i$ with $-\mathbf{x}_i$.

Then for any $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{r \times n}$ and $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^r$ and $\forall i, \mathbf{x}_i \in \mathbb{R}^n$ and $\sigma : \mathbb{R} \to \mathbb{R}$ mapping, $x \mapsto x\mathbf{1}_{x \geq 0} + \alpha x \mathbf{1}_{x \leq 0}$ for some $\alpha > 0$ we have,

$$\sum_{i=1}^{S} \sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle = S \frac{1+\alpha}{2} \mathbf{z}_1^\top \Big(\mathbf{A}\tilde{\Sigma}\mathbf{M}^\top\Big)\mathbf{z}_2$$

where we have defined, $\tilde{\Sigma} := \frac{1}{S}\Big[\sum_{i=1}^{S} \mathbf{x}_i \mathbf{x}_i^\top\Big]$

Note that in above the label associated with $\mathbf{x}_i$ is not forced to have any relation to the label associated to $-\mathbf{x}_i$.

Proof: By the assumed symmetry of the set $\mathcal{T}_\mathbf{x}$ we have,

$$\sum_{i=1}^{S} \sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle = \sum_{i=1}^{S} \sigma\Big(-\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle -\mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle$$

$$\implies 2\sum_{i=1}^{S} \sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle = \sum_{i=1}^{S} \Big[\sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle + \sigma\Big(-\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle -\mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle\Big]$$

$$\implies 2\sum_{i=1}^{S} \sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle = \sum_{i=1}^{S} \Big[\Big\{\sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big) - \sigma\Big(-\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\}\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle\Big]$$

Note that $\forall p \geq 0 \implies \sigma(p) = p$ and $\forall p < 0 \implies \sigma(p) = \alpha p$. Thus it follows that $\forall p \geq 0 \implies \sigma(p) - \sigma(-p) = p - \alpha(-p) = (1+\alpha)p$ and $\forall p < 0 \implies \sigma(p) - \sigma(-p) = \alpha p - (-p) = (1+\alpha)p$. Substituting this above we have,

$$2\sum_{i=1}^{S} \sigma\Big(\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \rangle\Big)\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle = (1+\alpha)\sum_{i=1}^{S} \Big[\Big\{\Big\langle \mathbf{A}^\top \mathbf{z}_1, \mathbf{x}_i \Big\rangle\Big\}\Big\langle \mathbf{M}\mathbf{x}_i, \mathbf{z}_2 \Big\rangle\Big]$$

$$= (1+\alpha)(\mathbf{A}^\top \mathbf{z}_1)^\top \Big[\sum_{i=1}^{S} \mathbf{x}_i \mathbf{x}_i^\top\Big]\mathbf{M}^\top \mathbf{z}_2$$

$$= S(1+\alpha)\mathbf{z}_1^\top\Big(\mathbf{A}\tilde{\Sigma}\mathbf{M}^\top\Big)\mathbf{z}_2$$

$\blacksquare$

APPENDIX C
TWO RECURSION ESTIMATES

Lemma C.1: Given constants $\eta', b, c_1, c_2 > 0$ suppose one has a sequence of real numbers $\Delta_1 = C, \Delta_2, ..$ s.t,

$$\Delta_{t+1} \leq (1 - \eta' b_1 + \eta'^2 c_1)\Delta_t + \eta'^2 c_2$$

Given any $\epsilon' > 0$ in the following two cases we have, $\Delta_T \leq \epsilon'^2$

- If $c_2 = 0, C > 0$ and for some $\delta_0 > 0$ we have, $c_1 > b_1^2 \frac{\delta_0}{(1+\delta_0)^2}$,
  $\eta' = \frac{b}{(1+\delta_0))c_1}$ and $T = O\left(\log \frac{C}{\epsilon'^2}\right)$
- If $0 < c_2 \leq c_1, \epsilon'^2 \leq C, \frac{b^2}{c_1} \leq \left(\sqrt{\epsilon'} + \frac{1}{\sqrt{\epsilon'}}\right)^2$,
  $\eta' = \frac{b}{c_1} \cdot \frac{\epsilon'^2}{(1+\epsilon'^2)}$ and $T = O\left(\frac{\log\left(\frac{\epsilon'^2(c_1-c_2)}{Cc_1-c_2\epsilon'^2}\right)}{\log\left(1 - \frac{b^2}{c_1} \cdot \frac{\epsilon'^2}{(1+\epsilon'^2)^2}\right)}\right)$.

Proof: Suppose we define $\alpha = 1 - \eta' b + \eta'^2 c_1$ and $\beta = \eta'^2 c_2$. Then we have by unrolling the recursion,

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

We recall that $\Delta_1 = C$ to realize that our lemma gets proven if we can find T s.t,

$$\alpha^{T-1}C + \beta\frac{1 - \alpha^{T-1}}{1 - \alpha} = \epsilon'^2$$

Thus we need to solve the following for T s.t, $\alpha^{T-1} = \frac{\epsilon'^2(1-\alpha)-\beta}{C(1-\alpha)-\beta}$

Case 1 : $\beta = 0$ In this case we see that if $\eta > 0$ is s.t $\alpha \in (0, 1)$ then,

$$\alpha^{T-1} = \frac{\epsilon'^2}{C} \implies T = 1 + \frac{\log \frac{C}{\epsilon'^2}}{\log \frac{1}{\alpha}}$$

But $\alpha = \eta'^2 c_1 - \eta' b + 1 = \left(\eta'\sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 + \left(1 - \frac{b^2}{4c_1}\right)$ Thus $\alpha \in (0, 1)$ is easily ensured by choosing $\eta' = \frac{b_1}{(1+\delta_0)c_1}$ for some $\delta_0 > 0$ and $c_1 > b_1^2 \frac{\delta_0}{(1+\delta_0)^2}$

This gives us the first part of the theorem.

Case $2 :$ $\beta > 0$

This time we are solving,

$$\alpha^{T-1} = \frac{\epsilon'^2(1 - \alpha) - \beta}{C(1 - \alpha) - \beta} \tag{28}$$

Towards showing convergence, we want to set η' such that $\alpha^{t-1} \in (0, 1)$ for all t. Since $\epsilon'^2 < C$, it is sufficient to require,

$$\beta < \epsilon'^2(1 - \alpha) \implies \alpha < 1 - \frac{\beta}{\epsilon'^2} \Leftrightarrow 1 - \frac{b^2}{4c_1} + \left(\eta'\sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 \leq 1 - \frac{\beta}{\epsilon'^2}$$

$$\Leftrightarrow \frac{\eta'^2 c_2}{\epsilon'^2} \leq \frac{b^2}{4c_1} - \left(\eta'\sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 \Leftrightarrow \frac{c_2}{\epsilon'^2} \leq \frac{b^2}{4c_1\eta'^2} - \left(\sqrt{c_1} - \frac{b}{2\sqrt{c_1}\eta'}\right)^2$$

Set $\eta' = \frac{b}{\gamma c_1}$ for some constant $\gamma > 0$ to be chosen such that,

$$\frac{c_2}{\epsilon'^2} \leq \frac{b^2}{4c_1 \cdot \frac{b^2}{\gamma^2 c_1^2}} - \left(\sqrt{c_1} - \frac{b}{2\sqrt{c_1} \cdot \frac{b}{\gamma c_1}}\right)^2 \implies \frac{c_2}{\epsilon'^2} \leq c_1 \frac{\gamma^2}{4} - c_1 \cdot \left(\frac{\gamma}{2} - 1\right)^2 \implies c_2 \leq \epsilon'^2 \cdot c_1(\gamma - 1)$$

Since $c_2 \leq c_1$ we can choose, $\gamma = 1 + \frac{1}{\epsilon'^2}$ and we have $\alpha^{t-1} < 1$. Also note that,

$$\alpha = 1 + \eta'^2 c_1 - \eta' b = 1 + \frac{b^2}{\gamma^2 c_1^2} - \frac{b^2}{\gamma c_1} = 1 - \frac{b^2}{c_1} \cdot \left(\frac{1}{\gamma} - \frac{1}{\gamma^2}\right).$$

$$= 1 - \frac{b^2}{c_1} \cdot \frac{\epsilon'^2}{(1 + \epsilon'^2)^2} = 1 - \frac{b^2}{c_1} \cdot \frac{1}{\left(\epsilon' + \frac{1}{\epsilon'}\right)^2}$$

And here we recall that the condition that the lemma specifies on the ratio $\frac{b^2}{c_1}$ which ensures that the above equation leads to $\alpha > 0$

Now in this case we get the given bound on T in the lemma by solving equation 28. To see this, note that,

$$\alpha = 1 - \frac{b^2}{c_1} \cdot \frac{\epsilon'^2}{(1 + \epsilon'^2)^2} \text{ and } \beta = \eta'^2 c_2 = \frac{b^2}{\gamma^2 c_1} \cdot c_2 = \frac{b^2 c_2}{c_1} \cdot \frac{(\epsilon'^2)^2}{(1 + \epsilon'^2)^2}.$$

Plugging the above into equation 28 we get, $\alpha^{T-1} = \frac{\epsilon'^2 \Delta (c_1 - c_2)}{C c_1 - c_2 \epsilon'^2} \implies T = 1 + \frac{\log\left(\frac{\epsilon'^2 (c_1 - c_2)}{C c_1 - c_2 \epsilon'^2}\right)}{\log\left(1 - \frac{b^2}{c_1} \cdot \frac{\epsilon'^2}{(1 + \epsilon'^2)^2}\right)}$ ∎

Lemma C.2: Suppose we have a sequence of real numbers $\Delta_1, \Delta_2, \ldots$ s.t

$$\Delta_{t+1} \leq (1 - \eta' b_1 + \eta'^2 c_1)\Delta_t + \eta'^2 c_2 + \eta' c_3$$

for some fixed parameters $b_1, c_1, c_2, c_3 > 0$ s.t $\Delta_1 > \frac{c_3}{b_1}$ and free parameter $\eta' > 0$. Then for,

$$\epsilon'^2 \in \left(\frac{c_3}{b_1}, \Delta_1\right), \ \eta' = \frac{b_1}{\gamma c_1}, \quad \gamma > \max\left\{\frac{b_1^2}{c_1}, \left(\frac{\epsilon'^2 + \frac{c_2}{c_1}}{\epsilon'^2 - \frac{c_3}{b_1}}\right)\right\} > 1$$

it follows that $\Delta_T \leq \epsilon'^2$ for,

$$T = \mathcal{O}\left(\log\left[\frac{\Delta_1}{\epsilon'^2 - \left(\frac{\frac{c_2}{c_1} + \gamma \cdot \frac{c_3}{b_1}}{\gamma - 1}\right)}\right]\right)$$

Proof: Let us define $\alpha = 1 - \eta' b_1 + \eta'^2 c_1$ and $\beta = \eta'^2 c_2 + \eta' c_3$. Then by unrolling the recursion we get,

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

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,

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

then $\alpha^{T-1} = \frac{\epsilon'^2(1 - \alpha) - \beta}{\Delta_1(1 - \alpha) - \beta}$. 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,

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

In the second equality above we have estimated the expression for T after substituting $\eta' = \frac{b_1}{\gamma c_1}$ in the expressions for α and β.

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

We recall that we have set $\eta' = \frac{b_1}{\gamma c_1}$. This implies that, $\alpha = 1 - \frac{b_1^2}{c_1}\cdot\left(\frac{1}{\gamma}-\frac{1}{\gamma^2}\right)$. Hence $\alpha > 0$ is ensured by the assumption that $\gamma > \frac{b_1^2}{c_1}$. 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,

$$-\frac{1}{\epsilon'^2}\cdot\left(\epsilon'^2(1-\alpha)-\beta\right) = \alpha - \left(1-\frac{\beta}{\epsilon'^2}\right)$$

$$= 1 - \frac{b_1^2}{4c_1} + \left(\eta'\sqrt{c_1}-\frac{b_1}{2\sqrt{c_1}}\right)^2 - \left(1-\frac{\beta}{\epsilon'^2}\right)$$

$$= \frac{\eta'^2 c_2 + \eta' c_3}{\epsilon'^2} + \left(\eta'\sqrt{c_1}-\frac{b_1}{2\sqrt{c_1}}\right)^2 - \frac{b_1^2}{4c_1}$$

$$= \frac{\left(\eta'\sqrt{c_2}+\frac{c_3}{2\sqrt{c_2}}\right)^2 - \frac{c_3^2}{4c_2}}{\epsilon'^2} + \left(\eta'\sqrt{c_1}-\frac{b_1}{2\sqrt{c_1}}\right)^2 - \frac{b_1^2}{4c_1}$$

$$= \eta'^2\left(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{c_3}{2\eta'\sqrt{c_2}}\right)^2 + \left(\sqrt{c_1}-\frac{b_1}{2\eta'\sqrt{c_1}}\right)^2 - \frac{1}{\eta'^2}\left[\frac{b_1^2}{4c_1}+\frac{1}{\epsilon'^2}\left(\frac{c_3^2}{4c_2}\right)\right]\right)$$

Now we substitute $\eta' = \frac{b_1}{\gamma c_1}$ for the quantities in the expressions inside the parantheses to get,

$$-\frac{1}{\epsilon'^2}\cdot\left(\epsilon'^2(1-\alpha)-\beta\right) = \alpha - \left(1-\frac{\beta}{\epsilon'^2}\right)$$

$$= \eta'^2\left(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{\gamma c_1 c_3}{2b_1\sqrt{c_2}}\right)^2 + c_1\cdot\left(\frac{\gamma}{2}-1\right)^2 - c_1\frac{\gamma^2}{4} - \frac{1}{\epsilon'^2}\cdot\frac{\gamma^2 c_1^2 c_3^2}{4b_1^2 c_2}\right)$$

$$= \eta'^2\left(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{\gamma c_1 c_3}{2b_1\sqrt{c_2}}\right)^2 + c_1(1-\gamma) - \frac{1}{\epsilon'^2}\cdot\frac{\gamma^2 c_1^2 c_3^2}{4b_1^2 c_2}\right)$$

$$= \frac{\eta'^2}{\epsilon'^2}\left(c_2 + \frac{\gamma c_1 c_3}{b_1} - \epsilon'^2 c_1(\gamma-1)\right)$$

$$= \frac{\eta'^2 c_1}{\epsilon'^2}\left((\epsilon'^2 + \frac{c_2}{c_1}) - \gamma\cdot\left(\epsilon'^2 - \frac{c_3}{b_1}\right)\right)$$

Therefore, $-\frac{1}{\epsilon'^2}\left(\epsilon'^2(1-\alpha)-\beta\right) < 0$ since by assumption $\epsilon'^2 > \frac{c_3}{b_1}$, and $\gamma > \frac{\left(\epsilon'^2+\frac{c_2}{c_1}\right)}{\epsilon'^2-\frac{c_3}{b_1}}$ ∎