Neural Networks 151 (2022) 264–275

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Provable training of a ReLU gate with an iterative non-gradient algorithm☆

Sayar Karmakar [a,*], Anirbit Mukherjee [b]

[a] Department of Statistics, University of Florida, 230 Newell Drive, Gainesville, 32611, FL, U.S.A.
[b] Department of Computer Science, The University of Manchester, Kilburn Building, Manchester, M13 9PL, U.K.

ARTICLE INFO

Article history:
Received 22 July 2021
Received in revised form 27 December 2021
Accepted 29 March 2022
Available online 4 April 2022

Keywords:
Neural nets
Non-gradient iterative algorithms
Stochastic algorithms
Non-smooth non-convex optimization

ABSTRACT

In this work, we demonstrate provable guarantees on the training of a single ReLU gate in hitherto unexplored regimes. We give a simple iterative 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 such results.

Leveraging certain additional moment assumptions, we also show a first-of-its-kind approximate recovery of the true label generating parameters under an (online) data-poisoning attack on the true labels, while training a ReLU gate by the same algorithm. Our guarantee is shown to be nearly optimal in the worst case and its accuracy of recovering the true weight degrades gracefully with increasing probability of attack and its magnitude.

For both the realizable and the non-realizable cases as outlined above, our analysis allows for mini-batching and computes how the convergence time scales with the mini-batch size. We corroborate our theorems with simulation results which also bring to light a striking similarity in trajectories between our algorithm and the popular S.G.D. algorithm — for which similar guarantees as here are still unknown.

1. Introduction

Over the last few years, there has been a surge of activity in using neural networks for complex artificial intelligence tasks. Human world champions of classic hard board games have famously been defeated by neural net-based approaches, Schrittwieser et al. (2020), Silver et al. (2016, 2018, 2017). At the core of many of these successes lie the ability of various heuristics to be able to solve the *learning theory* question of function optimization/*risk minimization*,

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

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 a uniform distribution on a finite set of points. But as of today, we have little or no mathematical guarantees about these heuristics which seemingly very efficiently solve the many useful instances of these optimization problems.

To the best of our knowledge about the state-of-the-art in deep-learning theory, any of these two optimization problems is typically *provably* solvable in poly-time for nets with more than 1 neuron in either of the following two mutually exclusive scenarios : **(a)** the nets in the class $\mathcal{N}$ are of constant size and the data comes as tuples $\mathbf{z}=(\mathbf{x},\mathbf{y})$ with $\mathbf{y}$ being the noise corrupted output at input $\mathbf{x}$ for a net (of a known architecture that which would be common to the class $\mathcal{N}$). And **(b)** the nets in $\mathcal{N}$ would be asymptotically large and the data comes as tuples $\mathbf{z}=(\mathbf{x},\mathbf{y})$ with no explicit functional relationship between $\mathbf{x}$ and $\mathbf{y}$ (but there could be geometric or statistical assumptions about the $\mathbf{x}$ and $\mathbf{y}$).

The simplifications that happen for infinitely large networks have been discussed since Neal (1996) and this theme has had a recent resurgence in works like Chizat and Bach (2018) and Jacot, Gabriel, and Hongler (2018). Eventually this led to an explosion of literature in getting linear time training of various kinds of neural nets when their width is a high degree polynomial in training set size and inverse accuracy (a somewhat *unrealistic* regime), Allen-Zhu, Li and Liang (2019), Allen-Zhu, Li, and Song (2019a, 2019b), Arora et al. (2019), Arora, Du, Hu, Li and Wang (2019), Arora et al. (2019), Du and Lee (2018), Du, Lee, Li, Wang, and Zhai (2018), Huang and Yau (2019), Kawaguchi and Huang (2019), Lee et al. (2017), Li et al. (2019), Su and Yang (2019), Wu, Du, and Ward

☆ Most of this work was done when Anirbit was at Wharton, the Department of Statistics at UPenn and at the Department of Applied Mathematics and Statistics, J.H.U.

* Corresponding author.
E-mail address: sayarkarmakar@ufl.edu (S. Karmakar).

https://doi.org/10.1016/j.neunet.2022.03.040
0893-6080/

(2019), Zou, Cao, Zhou, and Gu (2018) and Zou and Gu (2019). The essential proximity of this regime to kernel methods have been thought of separately in works like Allen-Zhu and Li (2019) and Wei, Lee, Liu, and Ma (2019). On the other hand we note that in the fully agnostic setting training even a single ReLU gate can be SPN-hard as shown in Goel, Kanade, Klivans, and Thaler (2016). Hence its an interesting mathematical question to isolate general conditions when the convergence speed can be fast for a single ReLU gate.

To the best of our knowledge, for training a single neuron to $\epsilon-$accuracy by (Stochastic) Gradient Descent ((S.)G.D.) existing results are restricted to a sample complexity of $\mathcal{O}(\text{poly}(1/\epsilon))$ even with realizable data. And any improvements to this have been known to happen only for the case of the marginal distribution on the input being Gaussian or for modifications of S.G.D. running on symmetric input distributions. We refer the interested readers to Frei, Cao, and Gu (2020) for a comprehensive summary of these results — against many of which we will compare our results too. In this paper, we break this barrier and improve the sample-complexity of training a single ReLU gate to $\mathcal{O}(\log(1/\epsilon))$ for realizable data and *without* tying ourselves to any specific symmetry in the distribution. We emphasize that not only are we able to achieve this only by slightly tweaking the popular S.G.D. algorithm itself but also that our algorithm has guarantees in cases where we make the data non-realizable by allowing for a data-poisoning attack. Our distributional assumptions are mild and reminiscent of the subspace eigenvalue conditions from Du, Lee, and Tian (2017). Moreover, through thorough experiments, we will show that our modified S.G.D. has strikingly similar convergence features as the traditional S.G.D. We summarize the technical details of our results in the following subsection.

1.1. A summary of our results

To make progress with provable training of a single gate we draw inspiration from the different avatars of iterative stochastic non-gradient algorithms analyzed in the past, Freund and Schapire (1999), Goel and Klivans (2017), Goel, Klivans, and Meka (2018), Kakade, Kanade, Shamir, and Kalai (2011), Klivans and Meka (2017), Pal and Mitra (1992) and Rosenblatt (1958). We shall organize our contributions in this paper under four groups as follows.

Firstly, in the short Section 2 we start with a quick re-analysis of a known algorithm called the GLM-Tron (Kakade et al., 2011) but under more general conditions than the previous proofs about it. We show how well it can do (empirical) risk minimization on any Lipschitz gate with Lipschitz constant < 2 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 certain benign assumptions on the noise.

Secondly, in Section 3, we exclusively focus on training the ReLU gate, $\mathbb{R}^n \ni \mathbf{x} \mapsto \max\{0, \mathbf{w}^\top \mathbf{x}\} \in \mathbb{R}$ for $\mathbf{w} \in \mathbb{R}^n$ being its weight. We note that for this gate, the corresponding empirical or the population risk is neither convex nor smooth w.r.t. how it depends on the weights. And yet we show a very simple iterative stochastic algorithm which can provably recover in linear time the underlying parameter $\mathbf{w}_*$ of the ReLU gate when the data being sampled is exactly realizable of the form $(\mathbf{x}, \max\{0, \mathbf{w}_*^\top \mathbf{x}\})$. That is, with high probability, in $\log\left(\frac{1}{\epsilon}\right)$ iterations we get ϵ close to $\mathbf{w}_*$ while starting from any arbitrary initial point. (We recall that for stochastic algorithms, linear time convergence i.e. getting ϵ close to the global minima in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ time is a hallmark of specialized optimization methods adapted for smooth strongly convex objectives like Johnson and Zhang (2013)). To achieve this we use a mild distributional condition which essentially 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.

Note that, in Section 3 we are using a stochastic algorithm while solving a regression problem specific to a ReLU gate and are exploiting the structure of the ReLU gate (and mild distributional assumptions) to directly achieve parameter recovery. The results in Section 2 also apply to a ReLU gate as a special case but in contrast, therein we used full-batch iterative updates to gain other advantages, namely of being able to handle more general gates while having essentially no distributional assumptions on the training data.

Thirdly, by making a slightly stronger distributional assumption, in Case (II) of Theorem 3.1 in Section 3 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 of the adversary's coin 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 of the adversary's attack or the magnitude of the distortion increases.

To the best of our knowledge, this is the first guarantee on training a ReLU gate while under any kind of an 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.

Lastly, in Section 3.1 we give an experimental demonstration of the performance of our algorithm. We do a side-by-side comparison on a ReLU gate between S.G.D. and our modified-S.G.D., under various setting which fall under the ambit of Theorem 3.1. In particular we track how the distance to the original optima $(\mathbf{w}_*)$ changes with time for the various settings that we consider. Seen from this perspective we emphasize that while guarantees like Case (II) of Theorem 3.1 still remain unknown for S.G.D., our algorithm's behavior in experiments closely resembles that of S.G.D. under similar settings. Thus our experiments encourage the conjecture that maybe our modification only very slightly changes the stochastic process induced by S.G.D. on a ReLU gate. We leave it for future work to investigate this possibility and to try generalizing this for larger nets.

1.2. Comparison to concurrent literature

Firstly, we note that the result in Goel et al. (2018) includes learning a ReLU gate under realizable settings as a special case of their result but only under the assumption of the distribution being symmetric. Specific to the marginal distribution on the data being Gaussian, works like Soltanolkotabi (2017) and Kalan, Soltanolkotabi, and Avestimehr (2019) had solved the same problem using gradient-based methods.

A notable recent progress with understanding the behavior of (stochastic) gradient descent on a ReLU gate was achieved in Frei et al. (2020). 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 what is guaranteed in Frei et al. (2020).

Also significantly in contrast to these previous results cited above, our Theorem 3.1 Case (II) encompasses the situation of a probabilistic adversary causing distortions to the true labels. 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.

Lastly, unlike any of these previous results, we keep track of the subtleties of using mini-batches and how the mini-batch size affects the convergence time.

In Diakonikolas, Goel, Karmalkar, Klivans and Soltanolkotab (2020), 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 the generating weights (the $\mathbf{w}_*$ therein) as closely as possible under adversarial corruption of the training labels. But this result of ours can be seen as a natural regression analogue of the recent result in Diakonikolas, Kontonis, Tzamos and Zarifi (2020) about learning half-space indicators under a Massart noise.

2. Re-analyzing the GLM-Tron

In this section we shall take a relook at the GLM-Tron algorithm (given below) from Kakade et al. (2011) and show that it converges on certain Lipschitz gates with no distributional assumption on the data.

Algorithm 1 GLM-Tron

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

First, we state the following crucial lemma,

Lemma 2.1. *Assume that for all $i = 1, \dots, S$ $\|\mathbf{x}_i\| \le 1$ and in Algorithm 1, σ is a $L-$Lipschitz non-decreasing function. Suppose the vector $\mathbf{w}$ and the scalar W are s.t at iteration t, we have $\|\mathbf{w}_t - \mathbf{w}\| \le W$ and we define $\eta > 0$ s.t $\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \le \eta$. Then it follows that,*

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

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

We give the proof of the above lemma in Appendix A.1. The above Algorithm 1 was introduced in Kakade et al. (2011) 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 will see in the following theorem as to how the above lemma leads to convergence of the *effective-E.R.M.*, $\tilde{L}_S$ by GLM-Tron on a single gate.

Theorem 2.2 (*GLM-Tron (Algorithm 1) Solves the Effective-E.R.M. on a ReLU Gate Up to Noise Bound with Minimal Distributional Assumptions*)**.** *Assume that for all $i = 1, \dots, S$ $\|\mathbf{x}_i\| \le 1$ 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 $\forall i, |\xi_i| \le \theta$ for some $\theta \ge 0$ 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,*

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

The proof of the above theorem is deferred to Appendix A.2.

Remark. *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 have been chosen *adversarially* up to the constraint on its norm.

If we make some assumptions on the noise being benign then we can get the following.

Theorem 2.3 (*Performance Guarantees on the GLM-Tron (Algorithm 1) When Solving E.R.M.*)**.** *Assume that the noise random variables $\xi_i, i = 1, \dots, S$ are identically distributed as a centered random variable say ξ. Then for $T = \frac{\|\mathbf{w}_*\|}{\epsilon}$, we have the following guarantee for GLM-Tron on the empirical risk after T iterations (say $L_S(h_T)$),*

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

The proof for the above has been given in Appendix A.3. Here we note a slight generalization of the above that can be easily read off from the above.

Corollary 2.4. *Suppose that the joint distribution of $\{\xi_i\}_{i=1,\dots,S}$ is s.t $\mathbb{P}\Big[|\xi_i| \le \theta \ \forall i \in \{1, \dots, S\}\Big] \ge 1 - \delta$ Then the guarantee of the above Theorem 2.3 still holds but now with probability at least $1-\delta$ over the noise distribution.*

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.

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

In this section we consider an adversary executing a data-poisoning attack on an iterative stochastic learning algorithm (Algorithm 2). Given a marginal distribution $\mathcal{D}$ on the inputs $\mathbf{x}$, suppose the corresponding true labels are generated as $y = \mathrm{ReLU}(\mathbf{w}_*^\top \mathbf{x})$ for some unknown $\mathbf{w}_* \in \mathbb{R}^n$. We assume sampling access to $\mathcal{D}$ and an adversarial label oracle that on the $t^{th}-$iterate gets queried with b inputs $\{\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_b}\}$ drawn uncorrelatedly from $\mathcal{D}$. The oracle then flips a coin for each minibatch data point 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 flip gives 1 only then does the adversary do a bounded (by a constant θ_*) additive distortion to the true label of the corresponding data.

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

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 $y = \mathrm{ReLU}(\mathbf{w}_*^\top \mathbf{x}) + \xi$ 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 Section 3.2 that our recovery guarantee is optimal in a certain sense. Additionally in the realizable case ($\theta_* = 0$ or $\beta = 0$ identically), 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.

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

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

  $\mathbf{w}_{t+1} := \mathbf{w}_t - \eta \mathbf{g}_t$
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 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 our theorems we define the following notation.

Definition 1. 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 Appendix 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 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 $\lambda_1 := \lambda_1(0)$ 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

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

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

Suppose $\mathbf{w}_*$ and θ_* are such that (a) $a_2, a_4, \beta_1 (> 0), \beta_2, \beta_3$ exist and (b) $\lambda_1(\theta_*) > 0$. Then there exist 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 $\mathrm{T} = O\left(\log \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{\epsilon^2 \delta - \theta_*^2 \cdot \left(\frac{c_2' + \gamma \cdot \frac{c_3'}{b_1'}}{\frac{c_1'}{\gamma - 1}} \right)} \right)$ iterations we have

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

where $\epsilon > 0$ and $\delta \in (0, 1)$ are s.t.

$$\epsilon^2 \delta = \beta_1^2 \cdot \frac{K \cdot \theta_*^2}{(2\lambda_1(\theta_*) - \frac{1}{K})} \tag{2}$$

and $K > 0$ large enough s.t $2\lambda_1(\theta_*) > \frac{1}{K}$, and

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

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

$$\text{and } \gamma > \max \left(\frac{b_1'^2}{c_1'}, \frac{\epsilon^2 \delta + \theta_*^2 \cdot \frac{c_2'}{c_1}}{\epsilon^2 \delta - \theta_*^2 \cdot \frac{c_3'}{b_1}} \right). \tag{3}$$

Remark 1. We collate 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) $\|\mathbf{w}_\mathrm{T} - \mathbf{w}_*\|^2 \le \epsilon^2 \implies \mathbb{E}_{\mathbf{x}}\left[\left(\mathrm{ReLU}(\mathbf{w}_\mathrm{T}^\top \mathbf{x}) - \mathrm{ReLU}(\mathbf{w}_*^\top \mathbf{x})\right)^2\right] \le \epsilon^2 \mathbb{E}\left[\|\mathbf{x}\|^2\right]$ 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 Section 3.2 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 upon choosing $\eta = b_1/(c_1(1+\delta_0))$ we 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 to achieve a pre-specified level of precision, it takes lesser time 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 p.d.f. This implies $\lambda_1(\theta_*)$ is finite and positive and thus convergence is ensured.

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.

3.1. Experimental demonstration of Algorithm 2

For experiments we sample the data $\mathbf{x}_{t_i}$ (Algorithm 2) in i.i.d fashion from a standard normal distribution in $n = 500$ dimensions. We instantiate a data-poisoning attack consistent with the assumptions in Theorem 3.1 in the following way : at the tth iterate we choose $\xi_{t_i} = \theta_* \mathbb{1}_{\{i \mod 2=0\}} - \theta_* \mathbb{1}_{\{i \mod 2\neq 0\}}$ and α_{t_i} is 0/1 w.p $\beta \in [0, 1]$ for $i = 1, \ldots, b$.

Then for a chosen value of $\mathbf{w}_*$ and $\eta = 0.01$, we plot how the parameter recovery error $\|\mathbf{w}_t - \mathbf{w}_*\|$ (averaged over multiple runs of the algorithm) varies with t,

- for different values of b, at fixed $\theta_* = 2$ and $\beta = 0.5$ in Fig. 1. Here we can see that larger values of mini-batch help attain lower errors faster.
- for different values of β, at fixed $\theta_* = 2$ and $b = 16$ in Fig. 2. Here we can see that there is a graceful degradation of the best achieved error with increasing probability of attack.
- for different values of θ_*, at fixed $\beta = 0.5$ and $b = 16$ in Fig. 3. Here we can see that there is a graceful degradation of the best achieved error with increasing magnitude of the attack.

We note that all the three observations above are consistent with what we would have expected from Theorem 3.1.

We recall that in Algorithm 2 if we redefined $\mathbf{g}_t$ to, $-\frac{1}{b}\sum_{i=1}^{b} \mathbb{1}_{\left\{\mathbf{w}_t^\top \mathbf{x}_{t_i} > 0\right\}}(y_{t_i} - \mathbf{w}_t^\top \mathbf{x}_{t_i})\mathbf{x}_{t_i}$ then it would be standard S.G.D. For comparison, we repeat the last two experiments with this S.G.D. and give the corresponding plots in Figs. 4 and 5.

We notice the striking similarity between the plots in Figs. 2 & 4 and Figs. 3 & 5 respectively. This motivates that our algorithm very closely mimics the behavior of S.G.D. while similar guarantees as in Theorem 3.1 yet remain elusive for S.G.D.

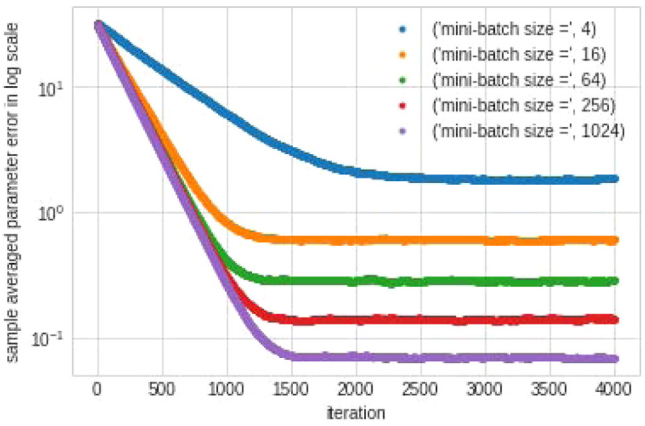

Fig. 1. Performance of Algorithm 2 with changing mini-batch size for $n = 500$, $\beta = 0.5$ and $\theta_* = 2$.

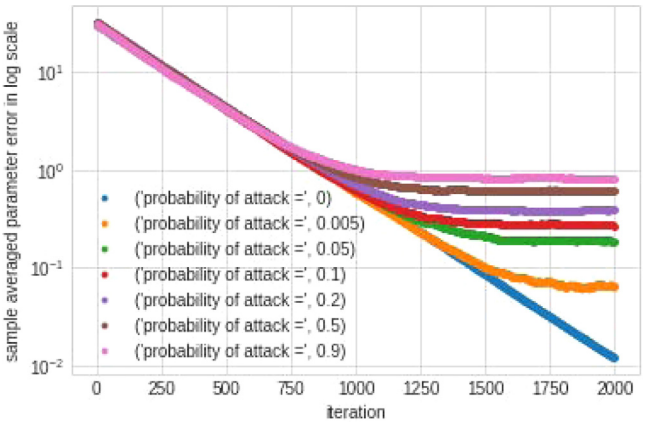

Fig. 2. Performance of Algorithm 2 with changing probability of attack for $n = 500$, $\theta_* = 2$ and $b = 16$.

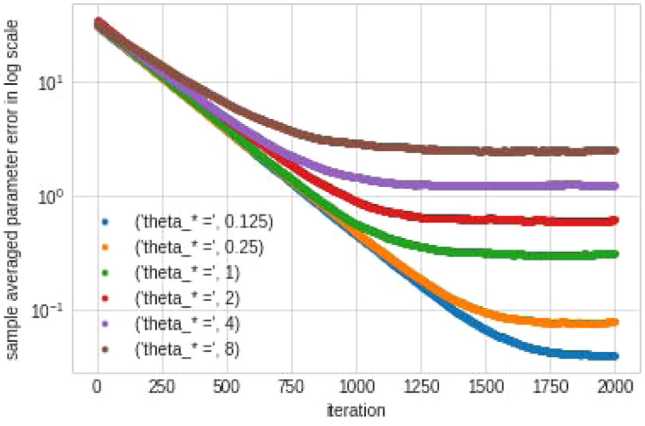

Fig. 3. Performance of Algorithm 2 with changing θ_* for $n = 500$, $\beta = 0.5$ and $b = 16$.

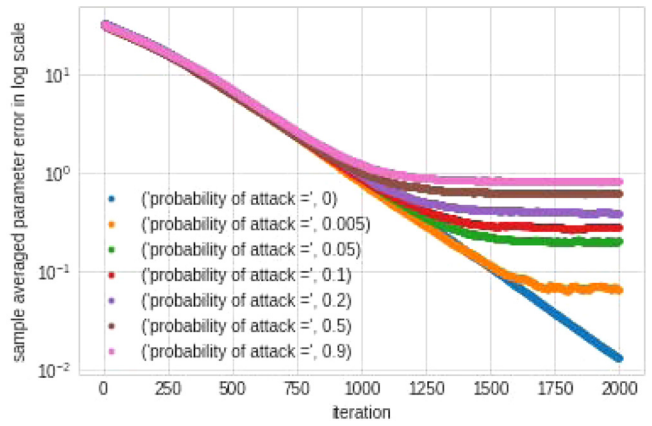

Fig. 4. Performance of S.G.D. with changing probability of attack for $n = 500, \theta_* = 2$ and $b = 16$.

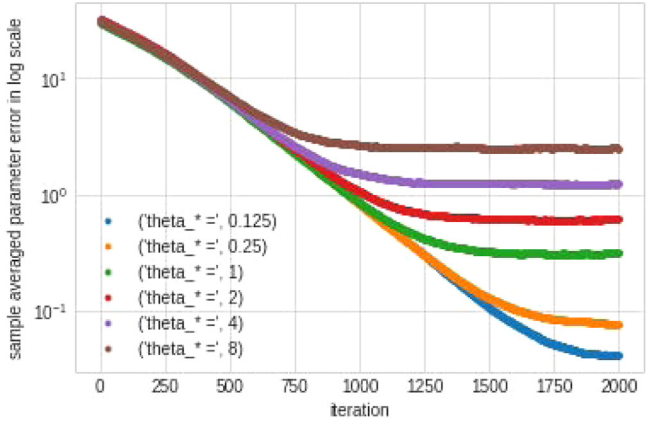

Fig. 5. Performance of S.G.D. with changing θ_* for $n = 500$, $\beta = 0.5$ and $b = 16$.

3.2. 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

$$\theta_{\text{adv}} \geq \sup_{\mathbf{x} \in \text{supp}(\mathcal{D})} |\text{ReLU}(\mathbf{w}_{\text{adv}}^\top \mathbf{x}) - \text{ReLU}(\mathbf{w}_*^\top \mathbf{x})| \tag{4}$$

It is easy to imagine cases where the supremum on 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,

$$\epsilon^2 = \frac{\theta^{\star 2}}{\delta c_{\text{bound}}} \implies \epsilon^2 \geq \frac{\theta_{\text{adv}}^2}{\mathbf{c}_{\text{bound}}} \tag{5}$$

Hence proving the 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 Eq. (5).

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 Eq. (4),

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

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 $\epsilon^2 \geq \frac{\theta_{\text{adv}}^2}{c_{\text{bound}}}$ is optimal up to a constant $\frac{r^2}{c_{\text{bound}}}$. We note that this argument also implies the worst-case near optimality of guarantees like Eq. (5) for *any* algorithm defending against this attack which also has the property of recovering the parameters correctly when the labels are exactly realizable.

4. Conclusion

In this work we have shown provable training of a ReLU gate under mild distributional conditions and pointed out cases where this happens in linear time while assuming only certain mild non-degeneracy conditions on the distribution. Also our results have probed how closely we can recover the original generating weights when the true training labels are subject to an (online) data-poisoning attack. And in this particular regime, in Section 3.1, we have given careful experimental evidence as to how our provably convergent modification of S.G.D. on a ReLU gate (Algorithm 2) seems to have very similar time dynamics as S.G.D. - while for the later such guarantees remain unknown.

We believe this raises the interesting question as to whether indeed one can rigorously show that the stochastic process induced by Algorithm 2, is a close approximant of true S.G.D. on a ReLU gate. We posit that this is a fruitful direction for future investigations and might lead to insights about the dynamics of S.G.D. for nets with a constant number of gates, which has so far mostly remained out of current mathematical reach.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are thankful to the Co-Editor in Chief, Action Editor, and referees for their constructive comments which have significantly helped towards this final form of the paper. Sayar Karmakar's research is partially supported by NSF DMS, USA 2124222. Anirbit Mukherjee would like to thank the inaugural MINDS Data Science Fellowship at J.H.U., Wharton Dean's Fund for Postdoctoral Research and Weijie Su's NSF CAREER DMS-1847415 for funding this research at various stages.

We would like to thank Daniel Dadush for his critical insights which led to the initial version of the Algorithm 2 (which first appeared in Mukherjee et al. (2021)). Multiple discussions with Amitabh Basu and Anup Rao (during Anirbit's internship at Adobe, San Jose) helped shape the core questions that were pursued in this paper. We would also like to acknowledge the collaboration with Ramchandran Muthukumar during the initial stages of the project.

Appendix A. Proofs of Section 2

A.1. Proof of Lemma 2.1

Proof of Lemma 2.1. We observe that,

$$\begin{aligned}
&\|\mathbf{w}_t - \mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}\|^2 - \|\Big(\mathbf{w}_t + \frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\Big) - \mathbf{w}\|^2 \\
&= -\frac{2}{S}\sum_{i=1}^{S}\Big\langle\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i, \mathbf{w}_t - \mathbf{w}\Big\rangle \\
&\quad - \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \\
&= \frac{2}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\Big(\langle \mathbf{w}, \mathbf{x}_i\rangle - \langle \mathbf{w}_t, \mathbf{x}_i\rangle\Big) \\
&\quad - \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \qquad \text{(A.1)}
\end{aligned}$$

Analyzing the first term on the RHS above we get,

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

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 on the RHS above,

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

Thus together we have,

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

Now we look at the second term on the RHS of Eq. (A.1) and that gives us,

$$\begin{aligned}
&\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \\
&= \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) + \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \\
&\leq \|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \\
&\quad + 2\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \\
&\quad \times \|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \\
&\quad + \|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \\
&\leq \eta^2 + 2\eta\|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\| \\
&\quad + \|\frac{1}{S}\sum_{i=1}^{S}\Big(\sigma(\langle \mathbf{w}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \qquad \text{(A.3)}
\end{aligned}$$

Now by Jensen's inequality we have,

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

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 Eq. (A.3) we have,

$$\|\frac{1}{S}\sum_{i=1}^{S}\Big(y_i - \sigma(\langle \mathbf{w}_t, \mathbf{x}_i\rangle)\Big)\mathbf{x}_i\|^2 \leq \eta^2 + 2\eta LW + \tilde{L}_S(h_t) \qquad \text{(A.4)}$$

Now we substitute Eqs. (A.2) and (A.4) into Eq. (A.1) to get,

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

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

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

A.2. Proof of Theorem 2.2

Proof of Theorem 2.2. 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 $\mathbf{w} = \mathbf{w}_*$, $\eta = \theta$ and $W = W_t$ s.t $W_t \geq \|\mathbf{w}_t - \mathbf{w}\| = \|\mathbf{w}_t - \mathbf{w}_*\|$ to get,

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

Thus, if $\tilde{L}_S(h_t) \geq \frac{L}{2-L}\Big(\epsilon + (\theta^2 + 2\theta \cdot W_t \cdot (L+1))\Big)$ 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, $\|\mathbf{w}_{t+1} - \mathbf{w}\| < \|\mathbf{w}_t - \mathbf{w}\|$. Since the iterations

start with $\mathbf{w}_1 = 0$, in the first step we can choose $W_1 = \|\mathbf{w}_*\|$. Now we proceed via induction : from what was argued earlier it follows that if till step t we can keep choosing $W_t = \|\mathbf{w}_*\|$, then till step t we have reduced the distance to $\mathbf{w}_*$ by $\mathcal{O}(t \cdot \epsilon)$ and either $\tilde{L}_S(h_t) < \frac{L}{2-L}\Big(\epsilon + (\theta^2 + 2\theta \cdot \|\mathbf{w}_*\| \cdot (L+1))\Big)$ or in the next step we would have $\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - \epsilon$ and hence the distance to $\mathbf{w}_*$ would decrease further by ϵ.

But the distance to $\mathbf{w}_*$ is lowerbounded by 0 and hence in at most $\frac{\|\mathbf{w}_*\|}{\epsilon}$ steps of the above kind we would have to have attained,

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

And that proves the theorem we wanted. □

A.3. Proof of Theorem 2.3

Proof of Theorem 2.3. Let the true empirical risk at the T^{th}−iterate be defined as,

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

Then it follows that,

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

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

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

For $T = \frac{\|\mathbf{w}_*\|}{\epsilon}$, we invoke the upperbound on $\tilde{L}_S(h_T)$ from Theorem 2.2 and we can combine it with the above to say,

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

And this proves the theorem we wanted. □

Appendix B. Proofs of Section 3

B.1. Proof of Theorem 3.1

Proof of Theorem 3.1. Here we analyze the dynamics of the Algorithm 2.

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

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

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

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

$$\begin{aligned}\text{Term 2} &= \eta^2\cdot\mathbb{E}\Big[\|\mathbf{g}_t\|^2 \mid S_{t-1}\Big] \\ &= \frac{\eta^2}{b^2}\sum_{i,j=1}^{b}\mathbb{E}\Big[\mathbf{1}_{y_{t_i}>\theta_*}\mathbf{1}_{y_{t_j}>\theta_*}\cdot(y_{t_i} - \mathbf{w}_t^\top\mathbf{x}_{t_i})\cdot(y_{t_j} - \mathbf{w}_t^\top\mathbf{x}_{t_j}) \\ &\cdot\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j}\rangle\Big|S_{t-1}\Big] \\ &= \frac{\eta^2}{b^2}\sum_{i,j=1}^{b}\mathbb{E}\Big[\mathbf{1}_{y_{t_i}>\theta_*}\mathbf{1}_{y_{t_j}>\theta_*}\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j}\rangle\cdot\Big[\alpha_{t_i}\alpha_{t_j}\xi_{t_i}\xi_{t_j} \\ &+ \big(\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_i}) - \mathbf{w}_t^\top\mathbf{x}_{t_i}\big)\big(\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_j}) - \mathbf{w}_t^\top\mathbf{x}_{t_j}\big) \\ &+ \alpha_{t_i}\xi_{t_i}\big(\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_j}) - \mathbf{w}_t^\top\mathbf{x}_{t_j}\big) \\ &+ \alpha_{t_j}\xi_{t_j}\big(\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_i}) - \mathbf{w}_t^\top\mathbf{x}_{t_i}\big)\Big]\Big|S_{t-1}\Big] \\ &\leq \frac{\eta^2}{b^2}\sum_{i,j=1}^{b}\Bigg(\mathbb{E}\Big[\mathbf{1}_{y_{t_i}>\theta_*}\mathbf{1}_{y_{t_j}>\theta_*}|\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j}\rangle| \\ &\times\Big[\alpha_{t_i}\alpha_{t_j}\theta_*^2 + |\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_i}) - \mathbf{w}_t^\top\mathbf{x}_{t_i}|\cdot|\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_j}) - \mathbf{w}_t^\top\mathbf{x}_{t_j}| \\ &+ \theta_*\big(\alpha_{t_i}\big|\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_j}) - \mathbf{w}_t^\top\mathbf{x}_{t_j}\big| \\ &+ \alpha_{t_j}\big|\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_i}) - \mathbf{w}_t^\top\mathbf{x}_{t_i}\big|\big)\Big]\Big|S_{t-1}\Big]\Bigg)\end{aligned}$$

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

$$\begin{aligned}&\text{Term 2} \\ &\leq \frac{\eta^2}{b^2}\sum_{i,j=1}^{b}\Bigg\{\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_j}>0}|\langle \mathbf{x}_{t_i}, \mathbf{x}_{t_j}\rangle| \\ &\cdot\Big[\alpha_{t_i}\alpha_{t_j}\theta_*^2 + |\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_i}) - \mathbf{w}_t^\top\mathbf{x}_{t_i}|\cdot|\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_j}) - \mathbf{w}_t^\top\mathbf{x}_{t_j}| \\ &+ \theta_*\big(\alpha_{t_i}\big|\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_j}) - \mathbf{w}_t^\top\mathbf{x}_{t_j}\big| \\ &+\alpha_{t_j}\big|\text{ReLU}(\mathbf{w}_*^\top\mathbf{x}_{t_i}) - \mathbf{w}_t^\top\mathbf{x}_{t_i}\big|\big)\Big]\Big|S_{t-1}\Big]\Bigg\}\end{aligned}$$

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

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 on the RHS of Eq. (B.2) as,

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

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 Eq. (B.2),

$$\mathbb{E}\,[\text{Term 2}]\leq\frac{\eta^2}{b^2}\cdot\theta_*^2(b\cdot\beta_2+(b^2-b)\cdot\beta_1^2)$$
$$+\frac{\eta^2}{b^2}\sum_{i=1}^{b}\Bigg\{\mathbb{E}\Bigg[\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\|\mathbf{x}_{t_i}\|^2\cdot|\mathbf{w}_*^\top\mathbf{x}_{t_i}-\mathbf{w}_t^\top\mathbf{x}_{t_i}|^2\Big|S_{t-1}\Big]\Bigg]$$
$$+2\theta_*\cdot\Bigg(\mathbb{E}\Bigg[\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\cdot\beta(\mathbf{x}_{t_i})\cdot\|\mathbf{x}_{t_i}\|^2|\mathbf{w}_*^\top\mathbf{x}_{t_i}-\mathbf{w}_t^\top\mathbf{x}_{t_i}|\Big|S_{t-1}\Big]\Bigg]\Bigg)\Bigg\}$$
$$+\frac{\eta^2}{b^2}\sum_{i,j=1,i\neq j}^{b}\Bigg\{\mathbb{E}\Bigg[\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\|\mathbf{x}_{t_i}\|\cdot|\mathbf{w}_*^\top\mathbf{x}_{t_i}-\mathbf{w}_t^\top\mathbf{x}_{t_i}|\Big|S_{t-1}\Big]$$
$$\times\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_j}>0}\|\mathbf{x}_{t_j}\|\cdot|\mathbf{w}_*^\top\mathbf{x}_{t_j}-\mathbf{w}_t^\top\mathbf{x}_{t_j}|\Big|S_{t-1}\Big]\Bigg]$$
$$+\theta_*\cdot\Bigg(\mathbb{E}\Bigg[\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\cdot\beta(\mathbf{x}_{t_i})\cdot\|\mathbf{x}_{t_i}\||\mathbf{w}_*^\top\mathbf{x}_{t_i}-\mathbf{w}_t^\top\mathbf{x}_{t_i}|\Big|S_{t-1}\Big]\Bigg]$$
$$\cdot\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_j}>0}\|\mathbf{x}_{t_j}\|\Big]+(i\leftrightarrow j)\Bigg)\Bigg\}$$

In the last term on 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,

$$\mathbb{E}\,[\text{Term 2}]\leq\frac{\eta^2}{b^2}\cdot\theta_*^2(b\cdot\beta_2+(b^2-b)\cdot\beta_1^2)$$
$$+\frac{\eta^2}{b^2}\sum_{i=1}^{b}\Bigg\{\mathbb{E}\left[\|\mathbf{w}_*-\mathbf{w}_t\|^2\right]\cdot\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\|\mathbf{x}_{t_i}\|^4\Big]$$
$$+2\theta_*\cdot\Bigg(\mathbb{E}\left[\|\mathbf{w}_*-\mathbf{w}_t\|\right]\cdot\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\cdot\beta(\mathbf{x}_{t_i})\cdot\|\mathbf{x}_{t_i}\|^3\Big]\Bigg)\Bigg\}$$
$$+\frac{\eta^2}{b^2}\sum_{i,j=1,i\neq j}^{b}\Bigg\{\mathbb{E}\left[\|\mathbf{w}_*-\mathbf{w}_t\|^2\right]\cdot\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\|\mathbf{x}_{t_i}\|^2\Big]$$
$$\times\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_j}>0}\|\mathbf{x}_{t_j}\|^2\Big]$$
$$+\theta_*\cdot\Bigg(\mathbb{E}\left[\|\mathbf{w}_*-\mathbf{w}_t\|\right]\cdot\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_i}>0}\cdot\beta(\mathbf{x}_{t_i})\cdot\|\mathbf{x}_{t_i}\|^2\Big]$$
$$\cdot\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top\mathbf{x}_{t_j}>0}\|\mathbf{x}_{t_j}\|\Big]+(i\leftrightarrow j)\Bigg)\Bigg\}$$
$$\leq\frac{\eta^2}{b}\cdot\Big\{a_4\cdot X_t+2\theta_*\cdot\mathbb{E}\left[\beta_3\cdot\|\mathbf{w}_*-\mathbf{w}_t\|\right]\Big\}$$
$$+\frac{\eta^2}{b^2}\cdot(b^2-b)\cdot\Big\{a_2^2\cdot X_t+2\theta_*\cdot\mathbb{E}\left[\beta_2a_1\cdot\|\mathbf{w}_*-\mathbf{w}_t\|\right]\Big\}$$
$$+\frac{\eta^2}{b^2}\cdot\theta_*^2(b\cdot\beta_2+(b^2-b)\cdot\beta_1^2)\tag{B.3}$$

In the last line above we have recalled that $\mathbf{x}_{t_i}$ and $\mathbf{x}_{t_j}$ are identically distributed and the definitions of a_1,a_2,a_4,β_2 &β_3 and have defined $X_t:=\mathbb{E}\left[\|\mathbf{w}_*-\mathbf{w}_t\|^2\right]$. In the second and the fourth terms on the RHS above we invoke the inequalities,

$$2\theta_*\cdot\mathbb{E}\left[\beta_3\cdot\|\mathbf{w}_*-\mathbf{w}_t\|\right]\leq(\theta_*\cdot\beta_3)^2+X_t$$
$$2\theta_*\cdot\mathbb{E}\left[\beta_2a_1\cdot\|\mathbf{w}_*-\mathbf{w}_t\|\right]\leq(\theta_*\cdot\beta_2\cdot a_1)^2+X_t$$

Thus we have,

$$\mathbb{E}\,[\text{Term 2}]\leq\left(\frac{a_4+1}{b}+\frac{(a_2^2+1)(b^2-b)}{b^2}\right)\cdot\eta^2\cdot X_t$$
$$+\left(\frac{(\theta_*\cdot\beta_3)^2}{b}+\frac{(\theta_*\cdot\beta_2\cdot a_1)^2(b^2-b)}{b^2}\right.$$
$$\left.+\frac{\theta_*^2(b\cdot\beta_2+(b^2-b)\cdot\beta_1^2)}{b^2}\right)\cdot\eta^2\tag{B.4}$$

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

$$\times (\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{x}_{t_i} \Big| S_{t-1} \Big]$$

Since $|\xi_{t_i}| \le \theta_*$ it follows that $y_{t_i} > \theta_*$

$\implies \mathbf{w}_*^\top \mathbf{x}_{t_i} > 0$. Hence,

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

$$\times (\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{x}_{t_i} \Big| S_{t-1} \Big]$$

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

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

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

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

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

$$\le -2\eta\lambda_1(\theta_*) \cdot X_t + \eta\Big(K(\theta_* \cdot \beta_1)^2 + \frac{1}{K}X_t\Big)\mathbf{1}_{\theta_* > 0} \tag{B.5}$$

In the last line above we used the following argument to write the upperbound in terms of $\lambda_1(\theta_*)$ as given in Definition 1. We observe that for any i, $\mathbb{E}\Big[\mathbf{1}_{y_{t_i} > \theta_*} \cdot \|\mathbf{x}_{t_i}\| \Big| S_{t-1}\Big] \le \mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 0} \cdot \|\mathbf{x}_{t_i}\| \Big| S_{t-1}\Big]$. 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,

$\mathbf{v}^\top \left(\mathbb{E}\Big[\left(\mathbf{1}_{y_{t_i} > \theta_*} - \mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 2\theta_*} \right) \mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top \Big| S_{t-1} \Big] \right) \mathbf{v} \ge 0$ and that in turn implies,

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

$$= \lambda_{\min}\Big(\mathbb{E}\Big[\mathbf{1}_{\mathbf{w}_*^\top \mathbf{x}_{t_i} > 2\theta_*} \mathbf{x}_{t_i}\mathbf{x}_{t_i}^\top \Big] \Big)$$

Case 1 : $\theta_* = 0$. Taking total expectations on both sides of Eq. (B.1) and setting $\theta_* = 0$ in the RHS of Eqs. (B.3) and (B.5) we have,

$$X_{t+1} \le \Big(1 - 2\eta\lambda_1 + \frac{\eta^2}{b} \cdot (a_4 + a_2^2(b-1))\Big)X_t \tag{B.6}$$

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 $\epsilon > 0, \delta \in (0,1)$, $X_T \le \epsilon^2 \cdot \delta$ for,

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

$$\eta = \frac{2b\lambda_1}{(a_4 + a_2^2(b-1))(1+\delta_0)}$$

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

Case 2 : $\theta_* > 0$. Taking total expectations on both sides of Eq. (B.1) and invoking the RHS of Eqs. (B.4) and (B.5) we have,

$$X_{t+1} \le \Big(1 - 2\eta\lambda_1(\theta_*) + \frac{\eta}{K} + \frac{\eta^2}{b} \cdot ((1+a_4) + (1+a_2^2)(b-1))\Big)X_t$$

$$+ K\theta_*^2 \cdot \eta \cdot \beta_1^2 + \theta_*^2 \cdot \frac{\eta^2}{b}$$

$$\cdot \Big(\beta_3^2 + (\beta_2 \cdot a_1)^2 \cdot (b-1) + (\beta_2 + (b-1) \cdot \beta_1^2)\Big) \tag{B.7}$$

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

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

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

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 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

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

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,

$\mathbb{P}\Big[\|\mathbf{w}_T - \mathbf{w}_*\|^2 \le \epsilon^2\Big] \ge 1 - \delta$. □

Appendix C. Proofs of 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,*

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

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

- *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 $T = O\Big(\log \frac{C}{\epsilon'^2}\Big)$*
- *If $0 < c_2 \le c_1, \epsilon'^2 \le C, \frac{b^2}{c_1} \le \Big(\sqrt{\epsilon'} + \frac{1}{\sqrt{\epsilon'}}\Big)^2$,*

$$\eta' = \frac{b}{c_1} \cdot \frac{\epsilon'^2}{(1+\epsilon'^2)} \text{ and } T = O\Big(\frac{\log\left(\frac{\epsilon'^2(c_1 - c_2)}{C \cdot c_1 - c_2\epsilon'^2}\right)}{\log\left(1 - \frac{b^2}{c_1} \cdot \frac{\epsilon'^2}{(1+\epsilon'^2)^2}\right)}\Big).$$

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

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

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,

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

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,

$$\begin{aligned}
\beta < \epsilon'^2(1-\alpha) &\implies \alpha < 1 - \frac{\beta}{\epsilon'^2} \\
&\Leftrightarrow 1 - \frac{b^2}{4c_1} + \left(\eta'\sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 \le 1 - \frac{\beta}{\epsilon'^2} \\
&\Leftrightarrow \frac{\eta'^2 c_2}{\epsilon'^2} \le \frac{b^2}{4c_1} - \left(\eta'\sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 \\
&\Leftrightarrow \frac{c_2}{\epsilon'^2} \le \frac{b^2}{4c_1\eta'^2} - \left(\sqrt{c_1} - \frac{b}{2\sqrt{c_1}\eta'}\right)^2
\end{aligned}$$

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

$$\begin{aligned}
&\frac{c_2}{\epsilon'^2} \le \frac{b^2}{4c_1 \cdot \frac{b^2}{\gamma^2 c_1^2}} - \left(\sqrt{c_1} - \frac{b}{2\sqrt{c_1} \cdot \frac{b}{\gamma c_1}}\right)^2 \\
&\implies \frac{c_2}{\epsilon'^2} \le c_1 \frac{\gamma^2}{4} - c_1 \cdot \left(\frac{\gamma}{2} - 1\right)^2 \implies c_2 \le \epsilon'^2 \cdot c_1(\gamma - 1)
\end{aligned}$$

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

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

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 Eq. (C.1). 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 Eq. (C.1) we get,

$$\alpha^{T-1} = \frac{\epsilon'^2 \Delta(c_1 - c_2)}{Cc_1 - c_2\epsilon'^2} \implies T = 1 + \frac{\log\left(\frac{\epsilon'^2(c_1-c_2)}{Cc_1 - c_2\epsilon'^2}\right)}{\log\left(1 - \frac{b^2}{c_1} \cdot \frac{\epsilon'^2}{(1+\epsilon'^2)^2}\right)}. \quad \square$$

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

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

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,

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

it follows that $\Delta_T \le \epsilon'^2$ 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 of Lemma C.2. 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,

$$\begin{aligned}
\Delta_t &\le \alpha\Delta_{t-1} + \beta \le \alpha(\alpha\Delta_{t-2} + \beta) + \beta \le \cdots \le \alpha^{t-1}\Delta_1 \\
&\quad + \beta(1 + \alpha + \cdots + \alpha^{t-2}).
\end{aligned}$$

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,

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

then $\alpha^{T-1} = \frac{\epsilon'^2(1-\alpha)-\beta}{\Delta_1(1-\alpha)-\beta}$. Note that **Claim 2** along with 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,

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

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. $\square$

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 $\gamma > 1$ $\square$

Proof of Claim 2. $0 < \epsilon'^2(1-\alpha) - \beta$. We note the following,

$$\begin{aligned}
&-\frac{1}{\epsilon'^2} \cdot \left(\epsilon'^2(1-\alpha) - \beta\right) \\
&= \alpha - \left(1 - \frac{\beta}{\epsilon'^2}\right) \\
&= 1 - \frac{b_1^2}{4c_1} + \left(\eta'\sqrt{c_1} - \frac{b_1}{2\sqrt{c_1}}\right)^2 - \left(1 - \frac{\beta}{\epsilon'^2}\right) \\
&= \frac{\eta'^2 c_2 + \eta' c_3}{\epsilon'^2} + \left(\eta'\sqrt{c_1} - \frac{b_1}{2\sqrt{c_1}}\right)^2 - \frac{b_1^2}{4c_1} \\
&= \frac{\left(\eta'\sqrt{c_2} + \frac{c_3}{2\sqrt{c_2}}\right)^2 - \frac{c_3^2}{4c_2}}{\epsilon'^2} + \left(\eta'\sqrt{c_1} - \frac{b_1}{2\sqrt{c_1}}\right)^2 - \frac{b_1^2}{4c_1} \\
&= \eta'^2 \left(\frac{1}{\epsilon'^2} \cdot \left(\sqrt{c_2} + \frac{c_3}{2\eta'\sqrt{c_2}}\right)^2 + \left(\sqrt{c_1} - \frac{b_1}{2\eta'\sqrt{c_1}}\right)^2\right.
\end{aligned}$$

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

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

$$\begin{aligned}
&-\frac{1}{\epsilon'^2}\cdot\left(\epsilon'^2(1-\alpha)-\beta\right)\\
&=\alpha-\left(1-\frac{\beta}{\epsilon'^2}\right)\\
&=\eta'^2\left(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{\gamma c_1 c_3}{2b_1\sqrt{c_2}}\right)^2+c_1\cdot\left(\frac{\gamma}{2}-1\right)^2\right.\\
&\quad\left.-c_1\frac{\gamma^2}{4}-\frac{1}{\epsilon'^2}\cdot\frac{\gamma^2c_1^2c_3^2}{4b_1^2c_2}\right)\\
&=\eta'^2\left(\frac{1}{\epsilon'^2}\cdot\left(\sqrt{c_2}+\frac{\gamma c_1 c_3}{2b_1\sqrt{c_2}}\right)^2+c_1(1-\gamma)-\frac{1}{\epsilon'^2}\cdot\frac{\gamma^2c_1^2c_3^2}{4b_1^2c_2}\right)\\
&=\frac{\eta'^2}{\epsilon'^2}\left(c_2+\frac{\gamma c_1c_3}{b_1}-\epsilon'^2c_1(\gamma-1)\right)\\
&=\frac{\eta'^2c_1}{\epsilon'^2}\left((\epsilon'^2+\frac{c_2}{c_1})-\gamma\cdot\left(\epsilon'^2-\frac{c_3}{b_1}\right)\right)
\end{aligned}$$

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

References

Allen-Zhu, Z., & Li, Y. (2019). What can ResNet learn efficiently, going beyond kernels? In *Advances in neural information processing systems* (pp. 9015–9025).

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

Allen-Zhu, Z., Li, Y., & Song, Z. (2019a). A convergence theory for deep learning via over-parameterization. In *International conference on machine learning* (pp. 242–252).

Allen-Zhu, Z., Li, Y., & Song, Z. (2019b). On the convergence rate of training recurrent neural networks. In *Advances in neural information processing systems* (pp. 6673–6685).

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., & Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in neural information processing systems* (pp. 8139–8148).

Arora, S., Du, S., Hu, W., Li, Z., & Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International conference on machine learning* (pp. 322–332).

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

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

Diakonikolas, I., Goel, S., Karmalkar, S., Klivans, A. R., & Soltanolkotabi, M. (2020). Approximation schemes for ReLU regression. In *Conference on learning theory*.

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

Du, S., & Lee, J. (2018). On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning* (pp. 1329–1338).

Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, X. (2018). Gradient descent finds global minima of deep neural networks. arXiv:1811.03804.

Du, S. S., Lee, J. D., & Tian, Y. (2017). When is a convolutional filter easy to learn? arXiv preprint arXiv:1709.06129.

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

Freund, Y., & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, *37*(3), 277–296.

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

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

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

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

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

Johnson, R., & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, *26*, 315–323.

Kakade, S. M., Kanade, V., Shamir, O., & Kalai, A. (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in neural information processing systems* (pp. 927–935).

Kalan, S. M. M., Soltanolkotabi, M., & Avestimehr, A. S. (2019). Fitting relus via sgd and quantized sgd. In *2019 IEEE international symposium on information theory (ISIT)* (pp. 2469–2473). IEEE.

Kawaguchi, K., & Huang, J. (2019). 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.

Klivans, A., & Meka, R. (2017). Learning graphical models using multiplicative weights. In *2017 IEEE 58th annual symposium on foundations of computer science (FOCS)* (pp. 343–354). IEEE.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., & Sohl-Dickstein, J. (2017). Deep neural networks as Gaussian processes. arXiv:1711.00165.

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

Mukherjee, A. (2021). A study of the mathematics of deep learning. arXiv:2104.14033.

Neal, R. M. (1996). Priors for infinite networks. In *Bayesian learning for neural networks* (pp. 29–53). Springer.

Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks*, *3 5*, 683–697.

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

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

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

Su, L., & Yang, P. (2019). On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in neural information processing systems* (pp. 2637–2646).

Wei, C., Lee, J. D., Liu, Q., & Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in neural information processing systems* (pp. 9709–9721).

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

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

Zou, D., & Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in neural information processing systems* (pp. 2053–2062).