PROVABLE UNLEARNING WITH GRADIENT ASCENT ON TWO-LAYER RELU NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Unlearning aims to remove specific data from trained models, addressing growing privacy and ethical concerns. We provide a theoretical analysis of a simple and widely used method—gradient ascent— used to reverse the influence of a specific data point without retraining from scratch. Leveraging the implicit bias of gradient descent towards solutions that satisfy the Karush-Kuhn-Tucker (KKT) conditions of a margin maximization problem, we quantify the quality of the unlearned model by evaluating how well it satisfies these conditions w.r.t. the retained data. To formalize this idea, we propose a new success criterion, termed (ϵ, δ, τ)**-successful** unlearning, and show that, for both linear models and two-layer neural networks with high dimensional data, a properly scaled gradient-ascent step satisfies this criterion and yields a model that closely approximates the retrained solution on the retained data. We also show that gradient ascent performs successful unlearning while still preserving generalization in a synthetic Gaussian-mixture setting.

1 INTRODUCTION

Machine Unlearning is an emerging field motivated by growing societal and legal demands—specifically, the need for machine learning models to "forget" specific data upon request. This concern has intensified following discoveries that private training data can be extracted from model outputs or weights (Carlini et al., 2019; Haim et al., 2022; Fredrikson et al., 2015). The demand is further reinforced by regulations such as the EU GDPR's Right to be Forgotten, as well as concerns about security and ethical AI. Machine unlearning addresses this challenge by aiming to undo the effect of particular samples without incurring the cost of full retraining.

The concept of unlearning was first formalized by Cao & Yang (2015) in the context of statistical query learning and has since been extended to deep neural networks. Broadly, two main approaches have emerged: retraining-based unlearning, which ensures complete data removal but is computationally expensive, and approximate unlearning, which aims for efficiency at the cost of weaker guarantees. Due to the stochastic and incremental nature of modern training procedures, which entangle data contributions, it is nontrivial to reverse the effect of the data to be forgotten while minimizing disruption to the retained data.

There is a large body of research on adapting given networks, namely, manipulating the weights post-training. For a training set S, a set of points $S_{\text{forget}} \subseteq S$ to unlearn, and its complement $S_{\text{retain}} = S \setminus S_{\text{forget}}$, a direct approach is to increase the training loss for samples in S_{forget} using gradient steps. This direct method was first implemented in *NegGrad* (Golatkar et al., 2020), simply taking multiple negative gradient steps for S_{forget} with respect to the training loss. Other gradient-related post-training methods use other losses and second order information for improved results (Guo et al., 2019; Golatkar et al., 2020; Warnecke et al., 2021; Triantafillou et al., 2024; Graves et al., 2021). There are also additional variants of *NegGrad*, such as *NegGrad+* (Kurmanji et al., 2023), and *Advanced NegGrad* (Choi & Na, 2023) which add a recovery phase, performing additional training steps on the retained set. In this work, we study the important building block of this foundational and widely-used method, a single gradient ascent step on the training loss w.r.t. S_{forget}.

One central question in the regime of approximate unlearning is how to measure unlearning performance. A common criterion, inspired by differential privacy (Dwork et al., 2014), evaluates success by comparing the output distributions of a model retrained from scratch with those of the unlearned model. This approach allows for approximate guarantees, where the distance between the two distributions is bounded by small parameters (Triantafillou et al., 2024; Ginart et al., 2019), providing a formal framework for quantifying the effectiveness of unlearning algorithms, albeit it is often too stringent.

To provide a rigorous framework for analyzing unlearning, we turn to recent results on the implicit bias of neural networks under gradient descent (Lyu & Li, 2019; Ji & Telgarsky, 2020). These works show that training tends toward solutions that satisfy the Karush-Kuhn-Tucker (KKT) conditions of the maximum-margin problem. We use these conditions to formulate an unlearning criterion: A successful unlearning procedure should modify the model from satisfying the KKT conditions w.r.t. S to approximately satisfying them w.r.t. S_{retain}. This property is necessary for successful unlearning. That is, since a network retrained only on S_{retain} converges to a KKT point w.r.t. S_{retain}, then a successful unlearning algorithm also needs to obtain such a KKT point, at least approximately. Note that the approximation relaxation here is analogous to the relaxation for the distribution distance, allowing bounds on the deviation from exact solution attained by retraining.

In our work, we analyze the unlearning performance of one gradient ascent step of a carefully chosen size. We define a new unlearning criterion for an unlearning algorithm $\mathcal{A}$, called (ϵ, δ, τ)**-successful** unlearning, using the KKT conditions as discussed above. Next, in both linear models and two-layer neural networks trained with high dimensional (or nearly-orthogonal) data, we prove that a gradient ascent step of an appropriate size is a successful unlearning algorithm. In addition, we show a setting where unlearning using gradient ascent is both successful and does not hurt the model's generalization performance.

In a bit more detail, our main contributions are:

- For linear predictors, where the margin-maximizing solution is unique, we prove that gradient ascent with an appropriate step size is a (ϵ, δ, τ)**-successful** unlearning algorithm. Specifically, it yields an approximately max-margin predictor for S_{retain}. Moreover, due to the uniqueness of the solution, the unlearned predictor aligns closely—measured via cosine similarity—with the exact model retrained on S_{retain}.
- We extend these findings to a two-layer neural network setting. Despite the added complexity and nonlinearity, we prove that a single gradient ascent step is a (ϵ, δ, τ)**-successful** unlearning algorithm for some small ϵ, δ, τ.
- We show that unlearning does not compromise out-of-sample prediction, using a synthetic mixture-of-Gaussians dataset. We show that models unlearned via gradient ascent maintain generalization performance comparable to the original.

Related Work

Machine unlearning was initially proposed in the statistical query setting by Cao & Yang (2015) and later extended to deep neural networks. The strongest unlearning guarantees are often formalized via *differential privacy* (Dwork et al., 2014), requiring indistinguishability between unlearned and retrained model outputs. This was relaxed using KL-divergence (Golatkar et al., 2020), while other lines of work evaluate unlearning effectiveness through privacy attacks, such as membership inference or data reconstruction (Niu et al., 2024; Haim et al., 2022).

To achieve these goals, many methods aim to avoid full retraining. For example, SISA (Bourtoule et al., 2021) partitions the training data into multiple shards to enable a faster future forgetting. Graves et al. (2021) proposed saving intermediate gradients during training with respect to different training data points, enabling faster simulation of retraining using these intermediate gradients without the forget set. Post-training approaches include fine-tuning for S_{retain} only (hoping for catastrophic forgetting of the rest of data) or with wrong labels for data in S_{forget} (Golatkar et al. (2020); Triantafillou et al. (2024); Graves et al. (2021); Kurmanji et al. (2023)), or using different losses (Golatkar et al., 2020). These techniques often rely on gradient-based updates, with loss functions adjusted for unlearning objectives. Several methods also incorporate second-order information for better precision (Guo et al., 2019; Golatkar et al., 2020; Warnecke et al., 2021).

The gradient-ascent method was first introduced by Golatkar et al. (2020) as *NegGrad*, applying negative gradient steps to increase loss on the forget set. Its extensions, *NegGrad+* (Kurmanji et al., 2023) and *advanced NegGrad* (Choi & Na, 2023), add a recovery phase by performing fine-tuning on the retained set. In this work, we isolate the basic component—gradient ascent—and study its behavior analytically.

On the theoretical side, Guo et al. (2019) analyzed linear models and proposed a certified unlearning framework. Leveraging the existence of a unique optimal solution, they argue that inspecting the training gradients on the retained dataset can reveal residual influence from the deleted point—particularly when the model incurs non-zero loss, which may indicate incomplete unlearning. Sekhari et al. (2021) analyze unlearning capacity based on test loss degradation. Our approach defines unlearning through the lens of KKT conditions, building on a line of work showing that training converges to a KKT point of the margin maximization problem for the dataset.

implicit bias and margin maximization A great body of research has studied the implicit bias of training neural networks with gradient methods toward solutions that generalize well (Neyshabur et al., 2017; Zhang et al., 2021). Our analysis is based on the characterization of the implicit bias of gradient flow on homogeneous models towards KKT solutions of the max margin problem, a result due to Lyu & Li (2019) and Ji & Telgarsky (2020). Implicit bias towards margin maximization was previously studied also for linear predictors (Soudry et al., 2018), deep linear networks and linear convolutional networks (Gunasekar et al., 2018). For a survey on implicit bias of neural networks see Vardi (2023).

2 SETTINGS

Notations. For $m \in \mathbb{N}$, we denote $[m] = \{1, 2, \ldots, m\}$, and for $l \in [m]$, we denote $[m]_{-l} = [m] \setminus \{\ell\}$. We use bold-face letters to denote vectors, e.g., $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. We use $\|\mathbf{x}\|$ to denote the Euclidean norm of a vector $\mathbf{x}$. We denote by $\mathbb{1}_{x \geq 0}$ the indicator function such that $\mathbb{1}_{x \geq 0} = 1$ if $x \geq 0$ and 0 otherwise. We denote by $\mathrm{sign}(x)$ the sign function, $\mathrm{sign}(x) = 1$ if $x \geq 0$ and -1 otherwise. We denote by $\mathcal{U}(A)$ the uniform distribution over a set A. For a distribution $\mathcal{D}$, we denote by $\mathbf{x} \sim \mathcal{D}^m$ a vector $\mathbf{x}$ that consists of m i.i.d. samples from $\mathcal{D}$. We denote by $\mathrm{cossim}(\mathbf{x}_1, \mathbf{x}_2)$ the cosine similarity of vectors $\mathbf{x}_1, \mathbf{x}_2$, defined by $\mathrm{cossim}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|}$.

2.1 ARCHITECTURES AND TRAINING

In this paper, we discuss unlearning in two fundamental models: a linear predictor and a two-layer fully connected network. For an input $\mathbf{x} \in \mathbb{R}^d$ and a vector $\mathbf{w} \in \mathbb{R}^d$, we will denote a linear predictor by $N(\mathbf{w}, \mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. Our two-layer network is defined by

$$N(\boldsymbol{\theta}, \mathbf{x}) = \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}) \ , \tag{1}$$

where $\sigma(z) = \max(z, 0)$ is the ReLU activation function. For all $j \in [n]$, we initialize $u_j \sim \mathcal{U}\left(\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}\right)$ and fix them throughout training. The parameters $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are trained. We denote by $\boldsymbol{\theta}$ a vectorization of all the trained parameters.

Given a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, we train our models using gradient descent over the empirical loss

$$L(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i N(\boldsymbol{\theta}, \mathbf{x}_i)) \ ,$$

where ℓ is either the logistic loss $\ell(q) = \log(1 + e^{-q})$ or the exponential loss $\ell(q) = e^{-q}$. That is, we have $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \beta \nabla L(\boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_t$ are the weights after the t-th training epoch, and β is the step size. We consider the limit where β is infinitesimally small, called *gradient flow*. More formally, in gradient flow the trajectory $\boldsymbol{\theta}_t$ is defined for all $t \geq 0$ and satisfies the differential equation $\frac{d\boldsymbol{\theta}_t}{dt} = -\nabla L(\boldsymbol{\theta}_t)$.

For a model $N(\boldsymbol{\theta}, \mathbf{x})$, where $\boldsymbol{\theta}$ are the parameters and $\mathbf{x}$ is the input, we say that N is homogeneous if there exists $C > 0$ such that for every $\alpha > 0$, and $\boldsymbol{\theta}, \mathbf{x}$, we have $N(\alpha \boldsymbol{\theta}, \mathbf{x}) = \alpha^C N(\boldsymbol{\theta}, \mathbf{x})$. We note that both a linear predictor and a two-layer network, as defined above, are homogeneous with $C = 1$.

For both linear and two-layer ReLU networks, there is an implicit bias towards margin maximization, as implied by the following theorem:

Theorem 2.1 (Lyu & Li (2019), Ji & Telgarsky (2020))**.** *Let $N(\mathbf{x}, \boldsymbol{\theta})$ be a homogeneous linear or ReLU neural network. Consider minimizing the logistic or exponential loss using gradient flow over a binary classification set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{-1, 1\}$. Assume that there is a time t_0 where $L(\boldsymbol{\theta}_{t_0}) < \frac{1}{m}$. Then, gradient flow converges in direction[1] to a first-order stationary point (i.e., Karush–Kuhn–Tucker point, or KKT point for short) of the margin-maximization problem:*

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta}\|^2 \quad s.t. \ \ \forall i \in [m], \ y_i N(\boldsymbol{\theta}, \mathbf{x}_i) \geq 1 \ . \tag{2}$$

[1] we say that gradient flow *converges in direction* to some $\tilde{\boldsymbol{\theta}}$ if $\lim_{t \to \infty} \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|} = \frac{\tilde{\boldsymbol{\theta}}}{\|\tilde{\boldsymbol{\theta}}\|}$.

Note that in the case of linear predictors a KKT point is always a global optimum,[2] but in the case of non-linear networks this is not necessarily the case. Thus, in non-linear homogeneous models gradient flow might converge to a KKT point which is not necessarily a global optimum of Problem 2.

While the above theorem captures the asymptotic behavior of gradient flow, namely as the time $t \to \infty$ it converges to a KKT point, the behavior of gradient flow after a finite time can be characterized by *approximate KKT points*.

Definition 2.1. *We say that $\boldsymbol{\theta}$ is a (ϵ, δ)-approximate KKT point for Problem 2, if there exist $\lambda_1, ..., \lambda_m$ such that*

1. *Dual Feasibility:* $\lambda_1, ..., \lambda_m \geq 0$.
2. *Stationarity:* $\|\boldsymbol{\theta} - \sum_{i=1}^m \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\mathbf{x}_i, \boldsymbol{\theta})\| \leq \epsilon$.
3. *Complementary Slackness:* $\forall i \in [m]$, $\lambda_i \left(y_i N(\mathbf{x}_i, \boldsymbol{\theta}) - 1\right) \leq \delta$.
4. *Primal Feasibility:* $\forall i \in [m]$, $y_i N(\mathbf{x}_i, \boldsymbol{\theta}) \geq 1$.

We note that a $(0, 0)$-approximate KKT point is a KKT point. When training with gradient flow, the parameters after finite time satisfy the following:

Theorem 2.2 (Lyu & Li (2019), Ji & Telgarsky (2020))**.** *Under the conditions of Theorem 2.1, the parameters $\boldsymbol{\theta}_t$ at time t point at the direction of an (ϵ_t, δ_t)-approximate KKT point for Problem 2, and $(\epsilon_t, \delta_t) \to (0, 0)$ as $t \to \infty$.*

Hence, when training a model it is reasonable to expect that the trained model is an (ϵ, δ)-approximate KKT point of Problem 2, for some small ϵ, δ.

2.2 An objective for unlearning

Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{-1, 1\}$ be a dataset, and let $(\mathbf{x}_r, y_r)$ be the example that we wish to unlearn. We call the dataset S the *original dataset*, and $S_{\text{retain}} = S \setminus \{(\mathbf{x}_r, y_r)\}$ the *retain dataset*. Note that we focus on unlearning a single data point. In Section 5 we will consider unlearning a subset.

Following Theorem 2.2, we assume that we start from a trained model that is an (ϵ, δ)-approximate KKT point w.r.t. the original dataset. We also note that for the same reason, retraining for S_{retain} will results in an (ϵ^*, δ^*)-approximate KKT point w.r.t. S_{retain}. Our objective can be stated as follows:

> *In the unlearning process, we wish to obtain a model that is close to an (ϵ^*, δ^*)-approximate KKT point w.r.t. the retain dataset, for some small ϵ^*, δ^*.*

Indeed, in unlearning, we wish to find a model that is "similar" to a model that we could have learned if we had trained on the retain dataset in the first place, and by Theorem 2.2 such a model must be an (ϵ^*, δ^*)-approximate KKT point w.r.t. the retain dataset. Hence, our objective can be viewed as a necessary condition for successful unlearning. That is, a successful unlearning algorithm needs to obtain a network which is close to an approximate KKT point, since otherwise the network cannot be similar to a model which is retrained with the retained dataset.

More formally, we have the following definition:

Definition 2.2 (successful unlearning)**.** *For a dataset S, and a homogeneous model with parameters $\boldsymbol{\theta}$, we say that $\mathcal{A}$ is an (ϵ, δ, τ)-**successful** unlearning algorithm w.r.t. $\boldsymbol{\theta}$ and S, if for every point $(\mathbf{x}_l, y_l) \in S$ there exists an (ϵ, δ)-approximate KKT point $\widetilde{\boldsymbol{\theta}}$ w.r.t. $S \setminus (\mathbf{x}_l, y_l)$, such that*

$$\text{cossim}(\mathcal{A}(\boldsymbol{\theta}, S, l), \widetilde{\boldsymbol{\theta}}) \geq 1 - \tau \ .$$

We note that from Theorem 2.2, retraining for time t is a $(\epsilon_t, \delta_t, \tau)$**-successful** unlearning algorithm with $\tau = 0$ and $(\epsilon_t, \delta_t) \to (0, 0)$. Our objective is to perform (ϵ, δ, τ)**-successful** unlearning for small (ϵ, δ, τ) but in an efficient manner that avoids retraining from scratch.

Definition 2.2 requires that the unlearned network $\mathcal{A}(\boldsymbol{\theta}, S, l)$ and the approximate KKT point $\widetilde{\boldsymbol{\theta}}$ have high cosine similarity. Indeed, note that since we consider homogeneous networks, the scale of the parameters only affects the

[2] For linear predictors, the theorem was obtained by Soudry et al. (2018).

scale of the output, and thus to show that $\mathcal{A}(\boldsymbol{\theta}, S, l)$ behaves similarly to $\widetilde{\boldsymbol{\theta}}$ it suffices to consider their corresponding normalized parameters. Moreover, for the normalized parameters, high cosine similarity implies small ℓ_2 distance, and since the the model is Lipschitz w.r.t. the parameters, it implies a similar behavior.

2.3 Unlearning with gradient ascent

Consider a network $N(\mathbf{x}, \boldsymbol{\theta})$ trained on a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{-1, 1\}$. In this paper, we consider the widely used *Gradient Ascent* method for unlearning. In this method, to unlearn a training point $(\mathbf{x}_r, y_r)$, we take a gradient step towards increasing the training loss for this particular point. Namely, for a step size β, the algorithm $\mathcal{A}_{\text{GA}}$ given $\boldsymbol{\theta}$, S and r, performs the following

$$\mathcal{A}_{\text{GA}}(\boldsymbol{\theta}, S, r) = \boldsymbol{\theta} + \beta \nabla_{\boldsymbol{\theta}} \ell(y_r N(\mathbf{x}_r, \boldsymbol{\theta})) \ . \tag{3}$$

Intuitively, training examples are often memorized in the sense that their training loss is too small, and gradient ascent allows us to undo it, that is, reduce the level of overfitting for these examples.

The gradient ascent method is a significant building block in the widely used unlearning method *NegGrad*, that consists of multiple such steps, and is the unlearning approach also for other variants of it (such as *NegGrad+* (Kurmanji et al., 2023) and *advanced NegGrad* (Choi & Na, 2023)) that additionally perform fine-tuning for the retained data.

In section 3 and section 4, we demonstrate that in both linear predictors and two-layer ReLU networks, respectively, unlearning with a single step of gradient ascent ($\mathcal{A}_{\text{GA}}$) is (ϵ, δ, τ)**-successful**, under certain assumptions.

2.4 Data

We consider a size-m training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{-1, 1\}$. We make the following assumption on S, for some parameters $\psi, \phi > 0$.

Assumption 2.3. *The training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ satisfies*

1. *For all $(\mathbf{x}, y) \in S$ we have $\|\mathbf{x}\|^2 \in [1 - \psi, 1 + \psi]$.*
2. *For all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$ with $i \neq j$ we have $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq \phi$.*

The data normalization assumption (Item 1 above) is very common, as data points with significantly different norms might cause biases during training, toward higher norm data points. The latter assumption can be phrased as near orthogonality of the data points, which is also quite common in the literature for high dimensional data (Frei et al., 2022; Vardi et al., 2022), and holds with high probability for popular distributions. A profound example of a distribution that satisfies both conditions with high probability is the Gaussian distribution $\mathcal{N}(0, \frac{1}{d} I_d)$, where d is the vector dimension. Another example is the uniform distribution over the unit sphere $\mathbb{S}^{d-1}$.

Example. *For a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where the $\mathbf{x}_i$'s are drawn i.i.d. from $\mathcal{N}(0, \frac{1}{d} I_d)$, Assumption 2.3 holds with probability at least $1 - (2me^{-d/500} + m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}})$, for $\psi = 0.1$ and $\phi = 1.1 \frac{\log(d)}{\sqrt{d}}$ (see Theorem A.1). Moreover, in Section 6 we will show that Assumption 2.3 holds with high probability for a mixture of Gaussians.*

3 Linear Predictors

In this section, we consider a linear predictor $N(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ trained on a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. Recall that when training a linear predictor, gradient flow converges in direction to the max-margin solution (i.e., global optimum of Problem 2), and after time t it reaches an (ϵ_t, δ_t)-approximate KKT point of Problem 2 where $(\epsilon_t, \delta_t) \to (0, 0)$ as $t \to \infty$. Moreover, recall that for linear predictors, Problem 2 has a unique global optimum.

The following theorem shows that unlearning using gradient ascent (denoted by $\mathcal{A}_{\text{GA}}$) is (ϵ, δ, τ)**-successful** w.r.t. S that satisfies Assumption 2.3 and $\mathbf{w}$ which is an approximate KKT point according to Definition 2.1, in two distinct aspects. In the first part (item 1 below), we show it for $\tau = 0$, that is, $\mathcal{A}_{\text{GA}}(\mathbf{w}, S, l)$ is a linear predictor which is an approximate KKT point of the max-margin problem w.r.t. $S \setminus (\mathbf{x}_l, y_l)$. Then, we show it for $\epsilon = \delta = 0$, namely, the cosine similarity of $\mathcal{A}_{\text{GA}}(\mathbf{w}, S, l)$ and the max-margin predictor w.r.t. $S \setminus (\mathbf{x}_l, y_l)$ is large.

Theorem 3.1. *Let $0 < \epsilon_1, \delta_1 \leq 0.5$, $\epsilon_d < 0.1$. Let $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ be a linear predictor trained on dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where S satisfies Assumption 2.3 for $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4m}$. Assume that $\mathbf{w}$ is an (ϵ_1, δ_1)-approximate KKT point for Problem 2 w.r.t. S according to Definition 2.1. Then, the gradient ascent algorithm $\mathcal{A}_{GA}$, with an appropriate step size, is a (ϵ, δ, τ)-**successful** unlearning algorithm w.r.t. $\mathbf{w}$ and S for:*

1. ***The case of*** $\epsilon = \epsilon_1 + \frac{\epsilon_1 \epsilon_d}{m - \epsilon_d}$, $\delta = \delta_1 + \frac{\delta_1 \epsilon_d}{m - \epsilon_d} + \frac{7.2\epsilon_d}{m}$, $\tau = 0$***:***
   The predictor $\mathcal{A}_{GA}(\mathbf{w}, S, l)$ has the direction of an (ϵ, δ)-approximate KKT point for the margin maximization problem (Problem 2) w.r.t. $S \setminus (\mathbf{x}_l, y_l)$.

2. ***The case of*** $\epsilon = \delta = 0$, $\tau = C(\sqrt{\epsilon_d} + \sqrt{\epsilon_1} + \sqrt{\delta_1})$ ***for some universal constant*** $C > 0$***:***
   Let $\mathbf{w}^*$ be a max-margin linear predictor w.r.t. the remaining training set $S \setminus (\mathbf{x}_l, y_l)$, i.e. the global optimum of the Problem 2 w.r.t. $S \setminus (\mathbf{x}_l, y_l)$. Then, $\operatorname{cossim}(\mathcal{A}_{GA}(\mathbf{w}, S, l), \mathbf{w}^*) \geq 1 - \tau$.

We now briefly discuss the proof intuition. Due to the stationarity condition for $\mathbf{w}$ (Definition 2.1), we can express $\mathbf{w}$ as weighted sum of the network's gradient up to some error vector $\mathbf{v}_{\epsilon_1}$ of norm ϵ_1

$$\mathbf{w} = \sum_{i=1}^m \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon = \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i + \mathbf{v}_{\epsilon_1} \ .$$

Then, by performing gradient ascent $\mathcal{A}_{\text{GA}}$ with the appropriate step size we get

$$\mathcal{A}_{\text{GA}}(\mathbf{w}, S, l) = \sum_{i=1}^m \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon - \lambda_l y_l \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_r) = \sum_{i \in [m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \ .$$

First, one can see that the subtraction will result in a stationary condition w.r.t. $S \setminus (\mathbf{x}_l, y_l)$ and the original λ_i's. Observing the margin for a point $(\mathbf{x}_t, y_t)$ (for $t \neq l$),

$$\langle \mathbf{w}, \mathbf{x}_t \rangle = \sum_{i=1}^m \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x}_t \rangle + \langle \mathbf{v}_{\epsilon_1}, \mathbf{x}_t \rangle \ ,$$

we get that the change in the parameter vector (due to the gradient step) results in an additional term of at most $\lambda_l |\langle \mathbf{x}_l, \mathbf{x}_t \rangle|$ compared to the original predictor's margin. Due to the near-orthogonality of the data points in S (Assumption 2.3), and a constant upper bound for λ_l which we prove, we get that this difference is of order $O(\frac{\epsilon_d}{m})$. Regarding the proof for (2), we consider the representation of $\mathbf{w}^*$

$$\mathbf{w}^* = \sum_{i=1}^m \lambda_i^* y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) = \sum_{i=1}^m \lambda_i^* y_i \mathbf{x}_i \ .$$

For $i \in [m]_{-l}$ we prove a small $O(\epsilon_1 + \epsilon_d)$ upper bound for the difference $\lambda_i^* - \lambda_i$, which implies that the two predictors $\mathcal{A}_{\text{GA}}(\boldsymbol{\theta}, S, l)$ and $\mathbf{w}^*$ independently reach very similar KKT multipliers for the margin maximization problem (Definition 2.1). This yield an $1 - O(\sqrt{\epsilon_d} + \sqrt{\epsilon_1} + \sqrt{\delta_1})$ lower bound in the cosine similarity. For the full proof we refer the reader to Appendix B.1.

4 Two-Layer ReLU Networks

In this section, we extend our analysis to two-layer ReLU neural networks. We consider a neural network of the form $N(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^n u_j \sigma(\mathbf{w}_j^\top \mathbf{x})$, trained on dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. Note that unlike the linear setting, the non-smoothness of $N(\mathbf{x}, \boldsymbol{\theta})$ implies that even small perturbations in $\boldsymbol{\theta}$ can cause significant shifts in the model's gradients. This introduces new challenges and, as a result, leads to a slightly weaker guarantee.

The following theorem establishes that unlearning using gradient ascent with an appropriate step size, constitutes an (ϵ, δ, τ)-**successful** unlearning w.r.t. S that satisfies Assumption 2.3 and $\boldsymbol{\theta}$ which is an approximate KKT according to Definition 2.1, where ϵ, δ, and τ are small quantities determined by the KKT approximation parameters of $\boldsymbol{\theta}$ and the underlying data characteristics. This implies that the unlearned parameter vector $\mathcal{A}_{\text{GA}}(\boldsymbol{\theta}, S, l)$ is close—in terms of cosine similarity—to an approximate KKT point $\widetilde{\boldsymbol{\theta}}$ corresponding to the retained dataset $S \setminus (\mathbf{x}_l, y_l)$.

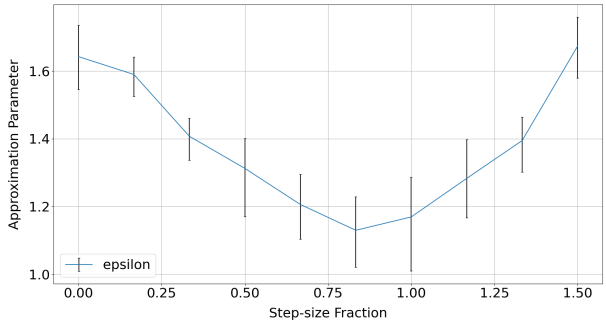

Figure 1: Effect of deviation from the correct step size on the KKT approximation parameter ϵ for a two-layer network. The x-axis shows the step size as a fraction of the step size from Theorem 4.1, and the y-axis shows the KKT approximation parameter ϵ of the unlearned model w.r.t. the retain dataset.

Theorem 4.1. *Let $0 < \epsilon_1, \delta_1 \leq 1$, $0 < \epsilon_d \leq 0.01$. Let $N(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x})$ be a two-layer ReLU network as defined in Eq. 1, such that $\boldsymbol{\theta}$ is an (ϵ_1, δ_1)-approximate KKT point for Problem 2 w.r.t. $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ according to Definition 2.1, and suppose that S satisfies Assumption 2.3 for $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4mn}$. Then, the gradient ascent algorithm $\mathcal{A}_{GA}$ with an appropriate step size is a (ϵ, δ, τ)-**successful** unlearning algorithm w.r.t. $\boldsymbol{\theta}$ and S, for $\epsilon = \epsilon_1 + \frac{9\epsilon_d\epsilon_1}{m - 9\epsilon_d} + \frac{23\epsilon_d}{\sqrt{m}}$, $\delta = \delta_1 + \frac{9\epsilon_d\delta_1}{m - 9\epsilon_d} + \frac{22.6\epsilon_d}{m}$ and $\tau = \frac{82\epsilon_d}{m}$.*

In Figure 1, we show the effect of varying the step size around the appropriate value β_l from Theorem 4.1 when unlearning a point $(\mathbf{x}_l, y_l) \in S$. The x-axis represents the step size as a fraction of β_l, and the y-axis shows the resulting KKT approximation parameter ϵ w.r.t. the retain dataset. We use a two-layer network (Eq. 1) trained on a 10-point dataset in $\mathbb{R}^{1000}$, and apply $\mathcal{A}_{\text{GA}}(\boldsymbol{\theta}, S, l)$ to a random data point. We can see that significantly deviating for β_l results in a worse approximation variable. See Appendix E for more details.

4.1 Proof sketch

We now outline the main ideas behind the proof. In this setting, unlike the linear setting, comparing the original parameter vector $\boldsymbol{\theta}$ with the unlearned parameter vector $\mathcal{A}_{\text{GA}}(\boldsymbol{\theta}, S, l)$ is nontrivial. Although the unlearning procedure introduces only a small perturbation, it may lead to significant changes in the activation map—the pattern of neuron activations across the data. Specifically, we define the activation map as the set of neurons $\mathbf{w}_j$ that are active on a data point $\mathbf{x}_i$, i.e., $\langle \mathbf{w}_j, \mathbf{x}_i \rangle \geq 0$. A key challenge arises when even small weight changes cause certain neurons to flip activation status.

To address this, we introduce an additive correction term (or "fix") for each weight vector $\mathbf{w}_j$, for $j \in [n]$, that restores the activation pattern. Using the stationarity conditions satisfied by $\boldsymbol{\theta}$ (Definition 2.1), we express each $\mathbf{w}_j$ as a weighted sum of the network's gradients, up to a small error term $\mathbf{v}_{\epsilon_1,j}$:

$$\mathbf{w}_j = \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}_j} N(\mathbf{x}_i, \boldsymbol{\theta}) + \mathbf{v}_{\epsilon,j} = u_j \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i + \mathbf{v}_{\epsilon_1,j}$$

where $\sigma'_{i,j}$ denotes the local derivative of the activation function.

After applying the gradient ascent step, the contribution of the forgotten point $(\mathbf{x}_l, y_l)$ is removed, which may alter the activation state of some neurons. To mitigate this, we construct a correction vector using a small scaling factor $c = O\left(\frac{\epsilon_d}{mn}\right)$, forming a new weight vector:

$$\widetilde{\mathbf{w}}_j = \mathbf{w}_j - u_j \lambda_l y_l \sigma'_{l,j} \mathbf{x}_l + |u_j| \lambda_l \sigma'_{l,j} c \sum_{k \in [m]_{-l}} \mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j \rangle) \,.$$

This correction reintroduces a small averaged influence from the retained points, specifically those where $\mathbf{w}_j$ was previously active. For a data point $\mathbf{x}_t$ where $\mathbf{w}_j$ was originally active, the new inner product becomes:

$$\langle \widetilde{\mathbf{w}}_j, \mathbf{x}_t \rangle = \langle \mathbf{w}_j, \mathbf{x}_t \rangle - u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_t \rangle + |u_j| \lambda_l \sigma'_{l,j} c \sum_{k \in [m]_{-l}} \langle \mathbf{x}_k, \mathbf{x}_t \rangle \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j \rangle) \ .$$

Since the data points $\mathbf{x}_l$ and $\mathbf{x}_t$ are nearly orthogonal (i.e., $\langle \mathbf{x}_l, \mathbf{x}_t \rangle = O(\frac{\epsilon_d}{mn})$, see Assumption 2.3), the middle term is of the same order as the correction, thus the correction term restores the activation. As a result, the corrected weight vector $\widetilde{\mathbf{w}}_j$ remains active on $\mathbf{x}_t$, preserving the original activation map. This activation preservation is essential: it enables us to meaningfully compare $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ in terms of margin, gradient differences, and parameter norms, facilitating the rest of the proof.

In establishing stationarity, the fixed vector introduces an additional error term beyond the original stationarity bound. In addition, because the activation map is preserved, we can upper bound the change in the margins of the remaining data points by a small factor of order $O\left(\frac{\epsilon_d}{mn}\right)$. Similar to the linear case, this margin deviation appears in both the upper and lower bounds, so we slightly rescale $\widetilde{\boldsymbol{\theta}}$ to restore feasibility and obtain an approximate KKT point for Problem 2 with respect to the reduced dataset $S \setminus \{(\mathbf{x}_l, y_l)\}$. To complete the proof, we show that $\mathcal{A}_{\text{GA}}(\boldsymbol{\theta}, S, l)$ remains close—in cosine similarity—to the rescaled $\widetilde{\boldsymbol{\theta}}$, differing only by the small fix and the minor scaling. The complete proof is provided in Appendix C.2.

5 Unlearning batches of data points

In the previous sections, we analyzed the unlearning of a single data point. We now extend these results to the case of unlearning a set of data points. Let $S_{\text{forget}} \subseteq S$ denote a subset of size k. We unlearn S_{forget} using a natural extension of the $\mathcal{A}_{\text{GA}}$ algorithm, namely by performing a step that consists of the k gradients of the points in S_{forget}, with appropriate coefficients. We denote this algorithm by $\mathcal{A}_{\text{k-GA}}$. Formally, for some real coefficients $\{\beta_r\}$, the algorithm $\mathcal{A}_{\text{k-GA}}$ performs the following

$$\mathcal{A}_{\text{k-GA}}(\boldsymbol{\theta}, S, S_{\text{forget}}) = \boldsymbol{\theta} + \sum_{(\mathbf{x}_r, y_r) \in S_{\text{forget}}} \beta_r \nabla_{\boldsymbol{\theta}} \ell(y_r N(\mathbf{x}_r, \boldsymbol{\theta})) \ .$$

In the case of linear predictors, the algorithm $\mathcal{A}_{\text{k-GA}}$ still satisfies the result from Theorem 3.1, but with slightly modified additive terms in the bounds on the KKT-approximation parameters ϵ, δ, while the bound on the cosine similarity (i.e., the parameter τ) remains unchanged. See a formal statement and proof in Appendix B.2.

For two-layer networks, we show that the result from Theorem 4.1 holds when unlearning a subset S_{forget} using the algorithm $\mathcal{A}_{\text{k-GA}}$, but with slightly modified parameters ϵ, δ, τ. See Appendix C.3 for the formal statement and proof.

6 Generalization of the Unlearned Classifier

In this section, we show that if $\boldsymbol{\theta}$ satisfies Definition 2.1 and the dataset S satisfies Assumption 2.3, then unlearning via a single gradient ascent step (i.e., $\mathcal{A}_{\text{GA}}$) may not harm generalization. As a concrete example, we consider a data distribution $\mathcal{D}_{MG}$ such that a dataset from this distribution satisfies w.h.p. Assumption 2.3 with parameters $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4mn}$. The distribution consists of two opposite Gaussian clusters, such that the cluster means have magnitude $d^{-\alpha}$ for some $\alpha \in (0, \frac{1}{4})$, and each deviation from the mean is drawn as $\boldsymbol{\zeta} \sim \mathcal{N}(0, \frac{1}{d} I_d)$. We show that both the original model and the unlearned model can generalize well, that is, classify the clusters with high probability.

Formally, our data satisfies the following. we denote the dataset by $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim \mathcal{D}_{MG}^m$, where $\forall i \in [m], (\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, and where $\mathcal{D}_{MG}$ is detailed as follows. It consists of a mixture of two Gaussians with means $\boldsymbol{\mu}_+, \boldsymbol{\mu}_- \in \mathbb{R}^d$, such that $\|\boldsymbol{\mu}_+\| = d^{-\alpha}$ for $\alpha \in (0, \frac{1}{4})$, and $\boldsymbol{\mu}_- = -\boldsymbol{\mu}_+$. For each i, we choose $\boldsymbol{\mu}_i \sim \mathcal{U}\{\boldsymbol{\mu}_+, \boldsymbol{\mu}_-\}$, then $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \frac{1}{d} I_d)$ and finally $y_i = 1$ if $\boldsymbol{\mu}_i = \boldsymbol{\mu}_+$ and -1 otherwise. Note that we can denote $\mathbf{x}_i = \boldsymbol{\mu}_i + \boldsymbol{\zeta}_i$ where $\boldsymbol{\zeta}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{d} I_d)$. We refer the reader to Lemma D.5, where we prove that for a given $\epsilon_d > 0$, m and α, S satisfies Assumption 2.3 for $\psi \leq 0.1$ and $\phi \leq \|\boldsymbol{\mu}_i\|^2 + 2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} + 1.1 \frac{\log(d)}{\sqrt{d}} \leq \frac{\epsilon_d}{4mn}$, w.p. $\geq 1 - (2me^{-\frac{d}{1700}} + m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}})$ for large enough d.

The following theorem shows that the unlearned network achieves generalization bounds comparable to those of the original classifier. Combined with the fact that it is close to an approximate KKT point of Problem 2 with respect to the

retained dataset (as established in Theorem 4.1), this demonstrates a clean setting where unlearning is successful, and it does not hurt generalization.

Theorem 6.1. *Let $0 < \epsilon_d \leq 0.01$. Let $N(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x})$ be a two-layer ReLU network as defined in Eq. 1, such that $\boldsymbol{\theta}$ is a KKT point for Problem 2 w.r.t. $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim \mathcal{D}_{MG}^m$ according to Definition 2.1. Fix $l \in [m]$ and denote by $\mathcal{A}_{GA}(\boldsymbol{\theta}, S, l)$ the parameters vector obtained by the gradient ascent algorithm $\mathcal{A}_{GA}$ for the data point $(\mathbf{x}_l, y_l) \in S$ with the appropriate step size from Theorem 4.1. Then, w.p. $\geq 1 - (2me^{-\frac{d}{1700}} + m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}})$ over the choice of the dataset S, both $N(\mathbf{x}, \boldsymbol{\theta})$ and $N(\mathbf{x}, \mathcal{A}_{GA}(\boldsymbol{\theta}, S, l))$ generalize. Namely,*

$$\Pr_{(\mathbf{x}_t, y_t) \sim \mathcal{D}_{MG}} [y_t N(\mathbf{x}_t, \boldsymbol{\theta}) > 0] \geq 1 - (2e^{-\frac{d}{1700}} + me^{-d/500} + 2md^{-\frac{\log(d)}{2}}) ,$$

$$\Pr_{(\mathbf{x}_t, y_t) \sim \mathcal{D}_{MG}} [y_t N(\mathbf{x}_t, \mathcal{A}_{GA}(\boldsymbol{\theta}, S, l)) > 0] \geq 1 - (2e^{-\frac{d}{1700}} + me^{-d/500} + 2md^{-\frac{\log(d)}{2}}) .$$

We briefly outline the intuition behind the generalization proof. Due to the small cluster means and relatively large variance, the data points in S are nearly orthogonal. Although the deviation from orthogonality is small, it is crucially structured: the inner product sign is determined by whether two points belong to the same or different clusters, namely

$$\begin{aligned} \mathbf{x}_i, \mathbf{x}_j \text{ are in the same cluster } &\Rightarrow \langle \mathbf{x}_i, \mathbf{x}_j \rangle > 0 , \\ \mathbf{x}_i, \mathbf{x}_j \text{ are in different clusters } &\Rightarrow \langle \mathbf{x}_i, \mathbf{x}_j \rangle < 0 . \end{aligned}$$

Now, using the fact that the classifier $\boldsymbol{\theta}$ satisfies the stationarity conditions with respect to S (Definition 2.1), we denote it by the weighted sum of its gradients direction, and consider its inner product with some $\mathbf{x}_t \sim \mathcal{D}_{MG}$

$$\langle \mathbf{w}_j, \mathbf{x}_t \rangle = \langle \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}_j} N(\mathbf{x}_i, \boldsymbol{\theta}), \mathbf{x}_t \rangle = u_j \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_t \rangle .$$

Since the inner product and the label align, we get that the activation map is of the same sign as u_j, hence each training point contributes positively to the classification of other points in the same cluster, and negatively to the others. This similarity of contribution implies that removing a point from S during unlearning does not significantly degrade the model's classification accuracy. The full proof is provided in Appendix D.2. Finally, we note that Theorem 6.1 can be readily extended to the case of unlearning a subset of data points using the algorithm $\mathcal{A}_{\text{k-GA}}$ discussed in Section 5.

7 Discussion and Future Work

In this work, we analyze the theoretical effectiveness of a single gradient-ascent step as a machine unlearning algorithm. Focusing on post-training unlearning methods, we propose a new criterion for unlearning success—called (ϵ, δ, τ)-**successful** unlearning—based on approximate satisfaction of KKT conditions. We prove that, in both linear models and two-layer neural networks, applying a gradient-ascent step $\mathcal{A}_{\text{GA}}$ with an appropriate step size w.r.t. the point we wish to forget is a (ϵ, δ, τ)-**successful** unlearning algorithm with some small ϵ, δ, τ, for a dataset S that satisfies Assumption 2.3 and a parameter vector $\boldsymbol{\theta}$ that is an approximate KKT point according to Definition 2.1. In the linear case, we additionally achieve near-exact recovery of the margin-maximizing predictor, implying stronger unlearning guarantees. We also demonstrate a clean distribution where unlearning is both successful and does not hurt generalization. Together, our results offer a rigorous foundation for analyzing gradient-based unlearning and confirm the practical utility of this simple yet widely used technique.

This work opens several avenues for further exploration. First, while we focus on a gradient-ascent step, it would be valuable to analyze the effect of an additional recovery phase for the retain data, including those used in NegGrad+ and related variants, under the same KKT-based framework. Second, it would be interesting to develop tighter bounds connecting approximate KKT satisfaction with practical privacy metrics, such as membership inference risk. On the applied side, evaluating unlearning methods under the new success criterion can lead to interesting comparisons between different methods. Moreover, a broader integration of our theoretical criterion with empirical privacy guarantees (e.g., differential privacy) could help bridging the gap between formal definitions and real-world deployment in safety-critical applications. Finally, extending our results to deeper architectures and additional distributions remains an important challenge.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Spencer Frei, Gal Vardi, Peter L Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35:22911–22924, 2022.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Jun Niu, Peng Liu, Xiaoyan Zhu, Kuo Shen, Yuecong Wang, Haotian Chi, Yulong Shen, Xiaohong Jiang, Jianfeng Ma, and Yuqing Zhang. A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2(5):404–454, 2024. ISSN 2949-7159. doi: https://doi.org/10.1016/j.jiixd.2024.02.001. URL https://www.sciencedirect.com/science/article/pii/S2949715924000064.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, et al. Are we making progress in unlearning? findings from the first neurips unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Advances in Neural Information Processing Systems*, 35:20921–20932, 2022.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A PROOFS OF DATA PRELIMINARIES FOR SECTION 2.1

Theorem A.1. *Let a set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ such that $\forall i$, $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$, $y_i \in \{-1, 1\}$ and $n \in \mathbb{N}$. Then, w.p. $\geq 1 - (2me^{-d/500} + m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}})$, the dataset S satisfies Assumption 2.3 for $\psi = 0.1$ and $\phi = 1.1\frac{\log(d)}{\sqrt{d}}$.*

Proof: Assumption 2.3 have 2 conditions:

1. For all $(\mathbf{x}, y) \in S$, $\|\mathbf{x}\|^2 \in [1 - \psi, 1 + \psi]$.
   Follows from Lemma A.7 w.p. $\geq 1 - 2me^{-\frac{d}{500}}$.
2. For all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$ s.t. $i \neq j$, $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq \phi$.
   From Lemma A.8 we have that w.p. $\geq 1 - (m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}})$,For all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$
   $$|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq 1.1\frac{\log(d)}{\sqrt{d}} .$$

□

Lemma A.1. *Let $w \in \mathbb{R}^n$ such that $w \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. Then:*

$$\mathbb{P}\left[\|w\|^2 \leq 0.9\sigma^2 n\right] \leq e^{-\frac{n}{400}} .$$

Proof: Note that $\left\|\frac{w}{\sigma}\right\|^2$ has the Chi-squared distribution. A concentration bound by Laurent and Massart (Laurent & Massart, 2000, Lemma 1) implies that for all $t > 0$ we have

$$\Pr\left[n - \left\|\frac{w}{\sigma}\right\|^2 \geq 2\sqrt{nt}\right] \leq e^{-t} .$$

Plugging-in $t = c \cdot n$, we get

$$\Pr\left[n - \left\|\frac{w}{\sigma}\right\|^2 \geq 2\sqrt{c}n\right] = \Pr\left[\left\|\frac{w}{\sigma}\right\|^2 \leq (1 - 2\sqrt{c})n\right] \leq e^{-c \cdot n} .$$

Thus, we have for $c = \frac{1}{400}$

$$\Pr\left[\left\|\frac{w}{\sigma}\right\|^2 \leq (1 - 2\frac{1}{\sqrt{400}})n\right] = \Pr\left[\left\|\frac{w}{\sigma}\right\|^2 \leq \frac{9}{10}n\right] \leq e^{-\frac{n}{400}} .$$

And finally,

$$\Pr\left[\|w\|^2 \leq \frac{9}{10}\sigma^2 n\right] \leq e^{-\frac{n}{400}} .$$

□

Lemma A.2. *Let $w \in \mathbb{R}^n$ with $w \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. Then:*

$$\Pr\left[\|w\|^2 \geq 1.1\sigma^2 n\right] \leq e^{-\frac{n}{500}} .$$

Proof: Note that $\left\|\frac{w}{\sigma}\right\|^2$ has the Chi-squared distribution. A concentration bound by Laurent and Massart (Laurent & Massart, 2000, Lemma 1) implies that for all $t > 0$ we have

$$\Pr\left[\left\|\frac{w}{\sigma}\right\|^2 - n \geq 2\sqrt{nt} + 2t\right] \leq e^{-t} .$$

Plugging-in $t = c \cdot n$, we get

$$\Pr\left[\left\|\frac{w}{\sigma}\right\|^2 - n \geq 2\sqrt{c}n + 2cn\right] = \Pr\left[\left\|\frac{w}{\sigma}\right\|^2 \geq (2\sqrt{c} + 2c + 1)n\right] \leq e^{c \cdot n} .$$

Thus, we have for $c = \frac{1}{500}$

$$\Pr\left[\left\|\frac{w}{\sigma}\right\|^2 \geq 1.1n\right] = \Pr\left[\left\|\frac{w}{\sigma}\right\|^2 \geq (2\frac{1}{\sqrt{500}} + \frac{2}{500} + 1)n\right] \leq e^{-\frac{n}{500}} .$$

And finally,

$$\Pr\left[\|w\|^2 \geq 1.1\sigma^2 n\right] \leq e^{-\frac{n}{500}} .$$

□

Lemma A.3. *For any $i \in [m]$, with probability $\geq 1 - (2e^{-\frac{d}{500}})$, $\|x_i\|^2 \in [0.9, 1.1]$.*

Proof: Using Lemma A.1 to lower bound $\|x_i\|^2$ for $x_i \sim \mathcal{N}(0, \frac{1}{d})$ w.p. $\geq 1 - e^{-\frac{n}{400}}$, and use Lemma A.2 to upper bound $\|x_i\|^2$ w.p. $\geq 1 - e^{-\frac{n}{500}}$. □

Lemma A.4. *Let $u \in \mathbb{R}^n$, and $v \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. Then, for every $t > 0$ we have*

$$\Pr\left[|\langle u, v\rangle| \geq \|u\|\, t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) .$$

Proof: We first consider $\langle \frac{u}{\|u\|}, v\rangle$. As the distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ is rotation invariant, one can rotate u and v to get $\tilde{u}$ and $\tilde{v}$ such that $\frac{\tilde{u}}{\|u\|} = e_1$, the first standard basis vector and $\langle \frac{u}{\|u\|}, v\rangle = \langle \frac{\tilde{u}}{\|u\|}, \tilde{v}\rangle$. Note, v and $\tilde{v}$ have the same distribution. We can see that $\langle \frac{\tilde{u}}{\|u\|}, \tilde{v}\rangle \sim \mathcal{N}(0, \sigma^2)$ since it is the first coordinate of $\tilde{v}$. By a standard tail bound, we get that for $t > 0$:

$$\Pr\left[|\langle \frac{u}{\|u\|}, v\rangle| \geq t\right] = \Pr\left[|\langle \frac{\tilde{u}}{\|u\|}, \tilde{v}\rangle| \geq t\right] = \Pr\left[|\tilde{v}_1| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) .$$

Therefore

$$\Pr\left[|\langle u, v\rangle| \geq \|u\|\, t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) .$$

□

Lemma A.5. *Let $u \sim \mathcal{N}(\mathbf{0}, \sigma_1^2 I_n)$, and $v \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 I_n)$. Then, for every $t > 0$ we have*

$$\Pr\left[|\langle u, v\rangle| \geq 1.1\sigma_1\sqrt{n}t\right] \leq e^{-\frac{n}{500}} + 2e^{-t^2/2\sigma_2^2} .$$

Proof:

Using Lemma A.2 we get that w.p. $\leq e^{-\frac{n}{500}}$ we have $\|u\| \geq 1.1\sigma_1\sqrt{n}$. Moreover, by Lemma A.4, w.p. $\leq 2\exp\left(-\frac{t^2}{2\sigma_2^2}\right)$ we have $|\langle u, v\rangle| \geq \|u\|\, t$. By the union bound, we get

$$\Pr\left[|\langle u, v\rangle| \geq 1.1\sigma_1\sqrt{n}t\right] \leq \Pr\left[\|u\| \geq 1.1\sigma_1\sqrt{n}\right] + \Pr\left[|\langle u, v\rangle| \geq \|u\|\, t\right] \leq e^{-\frac{n}{500}} + 2\exp\left(-\frac{t^2}{2\sigma_2^2}\right) .$$

□

Lemma A.6. *Let $u, v \sim \mathcal{N}(\mathbf{0}, \frac{1}{d} I_d)$. Then,*

$$\Pr\left[|\langle u, v\rangle| \geq 1.1\frac{\log(d)}{\sqrt{d}}\right] \leq e^{-\frac{d}{500}} + 2d^{-\frac{\log(d)}{2}} .$$

Proof: Using Lemma A.5 for $n = d$, $\sigma_1 = \sigma_2 = \frac{1}{\sqrt{d}}$ and $t = \frac{\log(d)}{\sqrt{d}}$. □

Lemma A.7. *Let a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be such that $\forall i$, $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$, for $m \leq d$. Then, w.p. $\geq 1 - 2me^{-\frac{d}{500}}$, For all $(\mathbf{x}, y) \in S$, $\|\mathbf{x}\|^2 \in [0.9, 1.1]$*

Proof: We prove both upper and lower bounds.

$$\begin{aligned}
&\Pr\left[\min_{i\in[m]} \left\{\|x_i\|^2\right\} < 0.9\right] = \\
&= \Pr\left[\exists i \in [m], \|x_i\|^2 < 0.9\right] \\
&\leq \sum_{i=1}^m \Pr\left[\|x_i\|^2 < 0.9\right] \leq me^{-\frac{d}{400}}
\end{aligned}$$

where the last inequality holds due to A.1.

$$\begin{aligned}
&\Pr\left[\max_{i\in[m]} \left\{\|x_i\|^2\right\} > 1.1\right] = \\
&= \Pr\left[\exists i \in [m], \|x_i\|^2 > 1.1\right] \\
&\leq \sum_{i=1}^m \Pr\left[\|x_i\|^2 > 1.1\right] \leq me^{-\frac{d}{500}}
\end{aligned}$$

where the last inequality holds due to A.2, and the claim follows. □

Lemma A.8. *Let a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be such that $\forall i$, $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$, for $m \leq d$. Then, w.p. $\geq 1 - (m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}})$, For all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$, $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq 1.1 \frac{\log(d)}{\sqrt{d}}$*

Proof: We prove an upper bound.

$$\begin{aligned}
&\Pr\left[\max_{i\neq j} \{|\langle x_i, x_j\rangle|\} > 1.1 \frac{\log(d)}{\sqrt{d}}\right] = \\
&= \Pr\left[\exists i.j \in [m], |\langle x_i, x_j\rangle| > 1.1 \frac{\log(d)}{\sqrt{d}}\right] \\
&\leq \sum_{i=1}^m \sum_{j=1}^m \Pr\left[|\langle x_i, x_j\rangle| > 1.1 \frac{\log(d)}{\sqrt{d}}\right] \leq m^2 e^{-d/500} + 2m^2 d^{-\frac{\log(d)}{2}}
\end{aligned}$$

where the last inequality holds due to Lemma A.6. □

B PROOFS FOR SECTION 3

Lemma B.1. *Let $\epsilon_d, \epsilon, \delta \leq 0.5$ and let $N(\mathbf{w}, \mathbf{x})$ be a linear classifier trained on a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, and assume that $\mathbf{w}$ is an (ϵ, δ)-approximate KKT point satisfying Definition 2.1, and S satisfies Assumption 2.3 for $\psi \leq 0.1, \phi \leq \frac{\epsilon_d}{4m}$. Note, for readability of the proof we denote ϵ_1 by ϵ and δ_1 by δ. Then,*

$$\max_i \lambda_i \leq 2.4$$

Proof: We look at $\lambda_r = \max_i \lambda_i$. If $\lambda_r = 0$ we are done, since the r.h.s is non-negative. Otherwise, we define $\mathbf{v}_\epsilon = \mathbf{w} - \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i$, and by item (2) from Definition 2.1 we have that $\|\mathbf{v}_\epsilon\| \leq \epsilon$. Hence, we have

$$\mathbf{w} = \sum_{i=0}^m \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \ ,$$

and from item (3) of Definition 2.1 and $\lambda_r > 0$, we have $1 + \frac{\delta}{\lambda_r} \geq y_r N(\mathbf{w}, \mathbf{x}_r) \geq 1$. Therefore,

$$
\begin{aligned}
1+\frac{\delta}{\lambda_r} \geq y_r N(\mathbf{w},\mathbf{x}_r) = y_r \sum_{i=0}^{m} \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x}_r\rangle + y_r\langle \mathbf{x}_r, \mathbf{v}_\epsilon\rangle =& \lambda_r \left\|\mathbf{x}_r\right\|^2 + y_r \sum_{i\neq r\in[m]} \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x}_r\rangle + y_r\langle \mathbf{x}_r, \mathbf{v}_\epsilon\rangle \\
\geq& \lambda_r(1-\psi) - \sum_{i\neq r\in[m]} \lambda_i |\langle \mathbf{x}_i, \mathbf{x}_r\rangle| - \left\|\mathbf{x}_r\right\| \left\|\mathbf{v}_\epsilon\right\| \\
\geq& \lambda_r(1-\psi) - \lambda_r\cdot\phi(m-1) - \epsilon\sqrt{1+\psi}
\end{aligned}
$$

where the last two inequalities holds due to Assumption 2.3 and Cauchy-Schwartz inequality.

Solving for λ_r leads to to

$$
\lambda_r^2\left((1-\psi)-\phi(m-1)\right) - (1+\epsilon\sqrt{1+\psi})\lambda_r - \delta \leq 0\,.
$$

Since $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4m}$ we get

$$
(1-\psi)-\phi(m-1) \geq 0.9-(m-1)\frac{\epsilon_d}{4m} \geq 0.9-\frac{\epsilon_d}{4} > 0\,,
$$

and we get that

$$
\lambda_r \leq \frac{(1+\epsilon\sqrt{1+\psi}) + \sqrt{(1+\epsilon\sqrt{1+\psi})^2 + 4((1-\psi)-\phi(m-1))\delta}}{2((1-\psi)-\phi(m-1))} \tag{4}
$$

Plugging in $\epsilon, \delta \leq 0.5$, $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4m}$, we get

$$
\begin{aligned}
\lambda_r \leq& \frac{(1+\epsilon\sqrt{1+\psi}) + \sqrt{(1+\epsilon\sqrt{1+\psi})^2 + 4((1-\psi)-\phi(m-1))\delta}}{2((1-\psi)-\phi(m-1))} \leq \\
\leq& \frac{(1+0.5\sqrt{1.1}) + \sqrt{(1+0.5\sqrt{1.1})^2+2}}{2(0.9-\frac{\epsilon_d}{4m}(m-1))} \\
\leq& \frac{(1+0.5\sqrt{1.1}) + \sqrt{(1+0.5\sqrt{1.1})^2+2}}{2(0.9-\frac{1}{8})} \leq \frac{3.61}{1.55} \leq 2.4\,.
\end{aligned}
$$

□

Lemma B.2. *Let $\epsilon_d, \epsilon, \delta \leq 0.5$ and let $N(\mathbf{w},\mathbf{x})$ be a linear classifier trained on a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, and assume that $\mathbf{w}$ is an (ϵ,δ)-approximate KKT point satisfying Definition 2.1, and S satisfies Assumption 2.3 for $\psi \leq 0.1, \phi \leq \frac{\epsilon_d}{4m}$. Let $t \in [m]$.Then,*

$$
\frac{1}{\left\|\mathbf{x}_t\right\|^2} - \frac{0.6\epsilon_d + 1.1\epsilon}{\left\|\mathbf{x}_t\right\|^2} \leq \lambda_t \leq \frac{1}{\left\|\mathbf{x}_t\right\|^2} + \frac{1.2\epsilon_d + 2.15\epsilon + 2.2\delta}{\left\|\mathbf{x}_t\right\|^2}\,.
$$

Proof: We begin showing the result for the more general case of $\epsilon, \delta \leq 0.5$. Let $t \in [m]$. Looking at an upper bound of the margin, we have

$$
\begin{aligned}
1 \leq y_t N(\mathbf{w},\mathbf{x}_t) = y_t \sum_{i=1}^{m} \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle + y_t\langle \mathbf{v}_\epsilon, \mathbf{x}_t\rangle \leq& \lambda_t \left\|\mathbf{x}_t\right\|^2 + \sum_{i\neq t\in[m]} \lambda_i |\langle \mathbf{x}_i, \mathbf{x}_t\rangle| + \langle \mathbf{v}_\epsilon, \mathbf{x}_t\rangle \\
\leq& \lambda_t \left\|\mathbf{x}_t\right\|^2 + \phi(m-1)\max_p \lambda_p + \langle \mathbf{v}_\epsilon, \mathbf{x}_t\rangle \\
\leq& \lambda_t \left\|\mathbf{x}_t\right\|^2 + 2.4\phi(m-1) + \epsilon\left\|\mathbf{x}_t\right\|\,,
\end{aligned}
$$

where the last inequality hold due to Lemma B.1 and Cauchy-Schwartz inequality. We solve it for λ_t with plugging in $\phi \leq \frac{\epsilon_d}{4m}$ getting a lower bound for it

$$\lambda_t \geq \frac{1}{\|\mathbf{x}_t\|^2} - \frac{2.4\phi(m-1)}{\|\mathbf{x}_t\|^2} - \frac{\epsilon}{\|\mathbf{x}_t\|} \geq \frac{1}{\|\mathbf{x}_t\|^2} - \frac{0.6\epsilon_d + 1.1\epsilon}{\|\mathbf{x}_t\|^2} .$$

We note that $1 - 0.6\epsilon_d - 1.1\epsilon \geq 0.15 > 0$, the therefore $\lambda_t > 0$. Next, to find an upper bound for λ_t, we look at a lower bound of the margin

$$\begin{aligned}
1 + \frac{\delta}{\lambda_t} \geq y_t N(\mathbf{w}, \mathbf{x}_t) = y_t \sum_{i=1}^{m} \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x}_t \rangle + y_t \langle \mathbf{v}_\epsilon, \mathbf{x}_t \rangle \geq & \lambda_t \|\mathbf{x}_t\|^2 - \sum_{i \neq t \in [m]} \lambda_i |\langle \mathbf{x}_i, \mathbf{x}_t \rangle| - \langle \mathbf{v}_\epsilon, \mathbf{x}_t \rangle \\
\geq & \lambda_t \|\mathbf{x}_t\|^2 - \phi(m-1) \max_p \lambda_p - \langle \mathbf{v}_\epsilon, \mathbf{x}_t \rangle \\
\geq & \lambda_t \|\mathbf{x}_t\|^2 - 2.4\phi(m-1) - \epsilon \|\mathbf{x}_t\| ,
\end{aligned}$$

where again the last inequalities holds due to Lemma B.1 Cauchy-Schwartz inequality. We get

$$\lambda_t^2 \|\mathbf{x}_t\|^2 - \lambda_t(1 + 2.4\phi(m-1) + \epsilon \|\mathbf{x}_t\|) - \delta \leq 0$$

and solve for λ_t with plugging in $\phi \leq \frac{\epsilon_d}{4m}, \|\mathbf{x}_t\|^2 \leq (1-\psi), \psi \leq 0.1$ we get an upper bound for λ_t

$$\begin{aligned}
\lambda_t &\leq \frac{(1 + 2.4\phi(m-1) + \epsilon \|\mathbf{x}_t\|) + \sqrt{(1 + 2.4\phi(m-1) + \epsilon \|\mathbf{x}_t\|)^2 + 4 \|\mathbf{x}_t\|^2 \delta}}{2 \|\mathbf{x}_t\|^2} \\
&\leq \frac{1 + 2.4\frac{\epsilon_d}{4m}(m-1) + \epsilon\sqrt{1+\psi} + 1 + 2.4\frac{\epsilon_d}{4m}(m-1) + \epsilon(1+\psi) + 4\delta(1+\psi)}{2 \|\mathbf{x}_t\|^2} \\
&\leq \frac{1}{\|\mathbf{x}_t\|^2} + \frac{2.4\frac{\epsilon_d}{4m}(m-1) + \epsilon\sqrt{1+\psi} + 2.4\frac{\epsilon_d}{4m}(m-1) + \epsilon(1+\psi) + 4\delta(1+\psi)}{2 \|\mathbf{x}_t\|^2} \\
&\leq \frac{1}{\|\mathbf{x}_t\|^2} + \frac{2.4\frac{\epsilon_d}{4} + \epsilon\sqrt{1.1} + 2.4\frac{\epsilon_d}{4} + \epsilon(1.1) + 4\delta(1.1)}{2 \|\mathbf{x}_t\|^2} \\
&\leq \frac{1}{\|\mathbf{x}_t\|^2} + \frac{1.2\epsilon_d + 2.15\epsilon + 2.2\delta}{\|\mathbf{x}_t\|^2} .
\end{aligned}$$

which finishes the proof. □

We next define an $(\epsilon, \delta, \gamma)$-approximate KKT. It is very similar to the (ϵ, δ)-approximate KKT definition given in Definition 2.1, with an extra γ relaxation of the margin.

Definition B.1. *A $(\epsilon, \delta, \gamma)$-approximate KKT for $\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta}\|^2$ s.t.$\forall i \in [m], y_i N(\boldsymbol{\theta}, \mathbf{x}_i) \geq 1$: $\exists \lambda_1, ..., \lambda_m$ such that*

1. $\lambda_1, ..., \lambda_m \geq 0$
2. $\left\| \boldsymbol{\theta} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) \right\|_2 \leq \epsilon$
3. $\forall i \in [m]$, $\lambda_i \left(y_i N(\boldsymbol{\theta}, \mathbf{x}_i) - 1 \right) \leq \delta$
4. $\forall i \in [m]$, $y_i N(\boldsymbol{\theta}, \mathbf{x}_i) \geq 1 - \gamma$

Now, we show that scaling an $(\epsilon, \delta, \gamma)$-approximate KKT can result in an (ϵ', δ')-approximate KKT, and determine the scaling effect on the approximation parameters.

Lemma B.3. *Let a network $N(\boldsymbol{\theta}, \mathbf{x})$ be such that $N(\boldsymbol{\theta}, \mathbf{x})$ is a 1-homogeneous function with respect to the weights. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a dataset. Then, if $\boldsymbol{\theta}$ is a $(\epsilon, \delta, \gamma)$-approximate KKT (according to the above Definition B.1) w.r.t S with corresponding $\{\lambda_i\}_{i=1}^m$, then $\frac{1}{1-\gamma}\boldsymbol{\theta}$ is a $(\frac{1}{1-\gamma}\epsilon, \max_p \lambda_p \frac{\gamma}{1-\gamma} + \frac{1}{1-\gamma}\delta)$-approximate KKT (according to Definition 2.1) w.r.t S with with the corresponding $\lambda_i' = C\lambda_i$.*

Proof: Let $N(\boldsymbol{\theta}, \mathbf{x})$ a 1-homogeneous function with respect to the weights, and $\boldsymbol{\theta}$ be a $(\epsilon, \delta, \gamma)$-approximate KKT. From 1-homogeneity, for all $C > 0$

$$N(C\boldsymbol{\theta}, \mathbf{x}) = CN(\boldsymbol{\theta}, \mathbf{x})$$

and the gradient is 0-homogeneous, meaning

$$\nabla_{\boldsymbol{\theta}} N(C\boldsymbol{\theta}, \mathbf{x}) = \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}) \ .$$

We denote $C = \frac{1}{1-\gamma}$, and show that $C\boldsymbol{\theta}$ satisfies the conditions in Definition 2.1.

1. $\left\| C\boldsymbol{\theta} - \sum_{i=2}^{m} C\lambda_i y_i \nabla_{\boldsymbol{\theta}} N(C\boldsymbol{\theta}, \mathbf{x}_i) \right\| = C \left\| \boldsymbol{\theta} - \sum_{i=2}^{m} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) \right\| \leq C\epsilon.$
2. Let $i \in [m]$. Then, $y_i N(C\boldsymbol{\theta}, \mathbf{x}_i) = C y_i N(\boldsymbol{\theta}, \mathbf{x}_i) \geq C(1-\gamma) = 1$
3. Let $i \in [m]$. Assume $\lambda_i \left(y_i N(\boldsymbol{\theta}, \mathbf{x}_i) - 1\right) \leq \delta$. If $\lambda_i = 0$ we are done. Else, $\lambda_i > 0$ and $y_i N(\boldsymbol{\theta}, \mathbf{x}_i) \leq 1 + \frac{\delta}{\lambda_i}$. Then,

$$\lambda_i \left(y_i N(C\boldsymbol{\theta}, \mathbf{x}_i) - 1\right) = \lambda_i \left(C y_i N(\boldsymbol{\theta}, \mathbf{x}_i) - 1\right) \leq$$
$$\leq \lambda_i \left(C(1 + \frac{\delta}{\lambda_i}) - 1 \right) = \lambda_i (C-1) + C\delta \leq \max_p \lambda_p \frac{\gamma}{1-\gamma} + \frac{1}{1-\gamma}\delta \ ,$$

which finishes the proof.

□

B.1 Proof for Theorem 3.1

Proof: Note, for readability of the proof we denote ϵ_1 by ϵ and δ_1 by δ.

Using the stationarity condition in Definition 2.1 for $\mathbf{w}$, we denote $\mathbf{v}_\epsilon = \mathbf{w} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i)$, so we get that $\|\mathbf{v}_\epsilon\| \leq \epsilon$ and

$$\mathbf{w} = \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \ .$$

Let $l \in [m]$, we wish to take a negative gradient step of size β, such that

$$\beta \nabla_{\mathbf{w}} \ell(y_l N(\mathbf{w}, \mathbf{x}_l)) = -\lambda_l y_l \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_l)$$

so we pick a step size $\beta = \frac{-\lambda_l}{\ell'(y_l N(\mathbf{w}, \mathbf{x}_l))}$. Then, when taking one gradient ascent step for $(\mathbf{x}_l, y_l)$ of size β, we get the following $\hat{\mathbf{w}}$

$$\hat{\mathbf{w}} = \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon - \lambda_l y_l \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_r) = \sum_{i \in [m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \ .$$

B.1.1 Proof of 1. $\hat{\mathbf{w}}$ has the direction of an $(\epsilon + \frac{\epsilon\epsilon_d}{m-\epsilon_d}, \delta + \frac{\delta\epsilon_d}{m-\epsilon_d} + \frac{7.2\epsilon_d}{m})$-approximate KKT point for the margin maximization problem for $S \setminus (\mathbf{x}_l, y_l)$.

For readability, we show that $\hat{\mathbf{w}}$ satisfies the conditions for $(\epsilon, \delta + \frac{1.44\epsilon_d}{m}, \frac{0.6\epsilon_d}{m})$-approximate KKT by Definition B.1, and then use Lemma B.3 to deduce that $\frac{1}{1-\frac{0.6\epsilon_d}{m}}\hat{\mathbf{w}}$ satisfies the conditions for $(\epsilon + \frac{\epsilon\epsilon_d}{m-\epsilon_d}, \delta + \frac{\delta\epsilon_d}{m-\epsilon_d} + \frac{7.2\epsilon_d}{m})$-approximate KKT according to Definition 2.1.

(1) Dual Feasibility: For all $i \in [m]_{-l}$, $\lambda_i \geq 0$. directly from dual feasibility for $\mathbf{w}$ (Definition 2.1).

(2) Stationarity: $\left\|\hat{\mathbf{w}} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\hat{\mathbf{w}}, \mathbf{x}_i)\right\| \leq \epsilon$**.** Since $\nabla_{\mathbf{w}} N(\hat{\mathbf{w}}, \mathbf{x}) = \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}) = x$, one can write

$$\hat{\mathbf{w}} = \sum_{i \in [m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon = \sum_{i \in [m]_{-l}} \lambda_i y_i \nabla_{\mathbf{w}} N(\hat{\mathbf{w}}, \mathbf{x}_i) + \mathbf{v}_\epsilon$$

and the claim follows from (2) stationarity for $\mathbf{w}$ (Definition 2.1).

Let $t \in [m]_{-l}$. Using the definitions of $\mathbf{w}$ and $\hat{\mathbf{w}}$, we can write the margin as

$$y_t N(\mathbf{w}, \mathbf{x}_t) = y_t \sum_{i=1}^{m} \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x}_t \rangle + y_t \langle \mathbf{v}_\epsilon, \mathbf{x}_t \rangle = y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) + y_t \lambda_l y_l \langle \mathbf{x}_l, \mathbf{x}_t \rangle \ . \tag{5}$$

Using this equality we prove the next two conditions:

(3) Complementarity Slackness: For all $t \in [m]_{-l}$**,** $\lambda_t \left(y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - 1\right) \leq \delta + \frac{1.44\epsilon_d}{m}$**.** If $\lambda_t = 0$ we are done. Else, $\lambda_t > 0$. From complementarity slackness of $\mathbf{w}$ being an (ϵ, δ)-approximate KKT, we know that $y_t N(\mathbf{w}, \mathbf{x}_t) \leq 1 + \frac{\delta}{\lambda_t}$. We use 5 to lower bound the margin of $y_t N(\mathbf{w}, \mathbf{x}_t)$, getting

$$\begin{aligned} 1 + \frac{\delta}{\lambda_t} \geq y_t N(\mathbf{w}, \mathbf{x}_t) =& y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) + y_t \lambda_l y_l |\langle \mathbf{x}_l, \mathbf{x}_t \rangle| \\ \geq& y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - \lambda_l |\langle \mathbf{x}_l, \mathbf{x}_t \rangle| \\ \geq& y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - \phi \max_p \lambda_p \ , \end{aligned}$$

plugging in $\phi \leq \frac{\epsilon_d}{4m}$ and the λ_p upper bound from Lemma B.1 we get

$$y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - \phi \max_p \lambda_p \geq y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - \frac{\epsilon_d}{4m} 2.4 \geq y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - \frac{0.6\epsilon_d}{m} \ .$$

We deduce an upper bound for the margin of $N(\hat{\mathbf{w}}, \mathbf{x}_t)$-

$$y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) \leq 1 + \frac{\delta}{\lambda_t} + \frac{0.6\epsilon_d}{m} = 1 + \frac{\delta + \frac{3}{5m}\lambda_t \epsilon_d}{\lambda_t} \leq 1 + \frac{\delta + \frac{3}{5m} 2.4\epsilon_d}{\lambda_t} \leq 1 + \frac{\delta + \frac{1.44\epsilon_d}{m}}{\lambda_t}$$

as desired.

(4) Primal Feasibility: For all $t \in [m]_{-l}$**,** $y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) \geq 1 - \frac{0.6\epsilon_d}{m}$**.** We use 5 to lower bound the margin of $N(\hat{\mathbf{w}}, \mathbf{x}_t)$, and use primal feasibility for $\mathbf{w}$ (Definition 2.1), getting

$$y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) = y_t N(\mathbf{w}, \mathbf{x}_t) - y_t \lambda_l y_l |\langle \mathbf{x}_l, \mathbf{x}_t \rangle| \geq 1 - \lambda_l |\langle \mathbf{x}_l, \mathbf{x}_t \rangle| \geq 1 - \phi \max_p \lambda_p \ .$$

Plugging in $\phi \leq \frac{\epsilon_d}{4m}$ and the λ_p upper bound from Lemma B.1 we get that

$$\phi \max_p \lambda_p \leq \frac{2.4\epsilon_d}{4m} \leq \frac{0.6\epsilon_d}{m} \ .$$

Hence, $y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) \geq 1 - \frac{0.6\epsilon_d}{m}$.

To conclude, we showed that $\hat{\mathbf{w}}$ is an $(\epsilon, \delta + \frac{1.44\epsilon_d}{m}, \frac{0.6\epsilon_d}{m})$-approximate KKT by Definition B.1 . Finally, we look at the scaled weights $\frac{1}{1 - \frac{0.6\epsilon_d}{m}} \hat{\mathbf{w}}$. For $\epsilon_d \leq 1$ We calculate

$$\frac{1}{1 - \frac{0.6\epsilon_d}{m}} \epsilon \leq \frac{m}{m - \epsilon_d} \epsilon = \left(1 + \frac{\epsilon_d}{m - \epsilon_d}\right) \epsilon = \epsilon + \frac{\epsilon \epsilon_d}{m - \epsilon_d} \ ,$$

and

$$\max_p \lambda_p \frac{\frac{0.6\epsilon_d}{m}}{1-\frac{0.6\epsilon_d}{m}} + \frac{\delta + \frac{1.44\epsilon_d}{m}}{1-\frac{0.6\epsilon_d}{m}} \leq \delta + \frac{\delta\epsilon_d}{m-\epsilon_d} + \frac{7.2\epsilon_d}{m}$$

and get from Lemma B.3 that $\frac{1}{1-\frac{0.6\epsilon_d}{m}}\hat{\mathbf{w}}$ is a $(\epsilon + \frac{\epsilon\epsilon_d}{m-\epsilon_d}, \delta + \frac{\delta\epsilon_d}{m-\epsilon_d} + \frac{7.2\epsilon_d}{m})$-approximate KKT by Definition 2.1 w.r.t. $S \setminus (\mathbf{x}_l, y_l)$. We note that $\hat{\mathbf{w}}$ and $\frac{1}{1-\frac{0.6\epsilon_d}{m}}\hat{\mathbf{w}}$ have the same direction, which finishes the proof.

B.1.2 Proof of 2. $Cosine - Similarity(\hat{\mathbf{w}}, \mathbf{w}^*) \geq 1 - C(\sqrt{\epsilon_d} + \sqrt{\epsilon_d} + \sqrt{\delta})$ for some $C > 0$.

Let $N(\mathbf{w}^*, \mathbf{x})$ be a max-margin linear predictor w.r.t. the remaining training set $S \setminus (\mathbf{x}_l, y_l)$. Hence, $\mathbf{w}^*$ is a KKT point of the margin maximization problem (2) w.r.t. $\{\mathbf{x}_i, y_i\}_{i\in[m]_{-l}}$, as in Definition 2.1 (with $\epsilon = \delta = 0$). From the stationarity condition we denote $\mathbf{w}^* = \sum_{i\in[m]_{-l}} \lambda_i^* y_i \mathbf{x}_i$.

Let $t \in [m]_{-l}$. We use Lemma B.2 to prove tight bounds for λ_t and λ_t^*. For a given t, λ_t and λ_t^* are close up to a small additive factor depend on ϵ_d, ϵ and δ. For λ_t we can use the results from Lemma B.2 directly, having

$$\frac{1}{\|\mathbf{x}_t\|^2} - \frac{0.6\epsilon_d + 1.1\epsilon}{\|\mathbf{x}_t\|^2} \leq \lambda_t \leq \frac{1}{\|\mathbf{x}_t\|^2} + \frac{1.2\epsilon_d + 2.15\epsilon + 2.2\delta}{\|\mathbf{x}_t\|^2} . \tag{6}$$

For λ_t^*, since $\mathbf{w}^*$ is a KKT point of 2 w.r.t. $S \setminus (\mathbf{x}_l, y_l)$, we have a dataset of size $m-1$ and $\epsilon = \delta = 0$. To accommodate the different parameter, we note that $\phi \leq \frac{\epsilon_d}{4m} \leq \frac{\epsilon_d}{4(m-1)}$, conclude that

$$\frac{1}{\|\mathbf{x}_t\|^2} - \frac{0.6\epsilon_d}{\|\mathbf{x}_t\|^2} \leq \lambda_t^* \leq \frac{1}{\|\mathbf{x}_t\|^2} + \frac{1.2\epsilon_d}{\|\mathbf{x}_t\|^2} . \tag{7}$$

And, similar note hold for B.1 resulting in $\lambda^* \leq 2.4$. We are now ready to prove the cosine similarity lower bound. For $\hat{\mathbf{w}} = \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon$ and $\mathbf{w}^* = \sum_{i\in[m]_{-l}} \lambda_i^* y_i \mathbf{x}_i$, we have

$$\frac{\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle}{\|\hat{\mathbf{w}}\| \, \|\mathbf{w}^*\|} = \frac{\langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon, \sum_{i\in[m]_{-l}} \lambda_i^* y_i \mathbf{x}_i \rangle}{\|\hat{\mathbf{w}}\| \, \|\mathbf{w}^*\|} .$$

We upper bound the norm of the predictors, when using 6 and 7 for any $i \in [m]_{-l}$ separately, bounding $\lambda_i \|\mathbf{x}_i\|^2$ and $\lambda_i^* \|\mathbf{x}_i\|^2$ respectively. Upper bounding $\|\hat{\mathbf{w}}\|^2$ we get

$$\begin{aligned}
\|\hat{\mathbf{w}}\|^2 &= \left\| \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \right\|^2 = \langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon, \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \rangle = \\
&= \langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i \rangle + 2\langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i, \mathbf{v}_\epsilon \rangle + \langle \mathbf{v}_\epsilon, \mathbf{v}_\epsilon \rangle \\
&\leq \sum_{i\in[m]_{-l}} \lambda_i^2 \|x_i\|^2 + \sum_{i\neq k\in[m]_{-l}} \lambda_i \lambda_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle + 2 \sum_{i\in[m]_{-l}} \lambda_i \langle \mathbf{x}_i, \mathbf{v}_\epsilon \rangle + \epsilon^2
\end{aligned}$$

From 6 we get that $\lambda_i \|x_i\|^2 \leq (1 + 1.2\epsilon_d + 2.15\epsilon + 2.2\delta)$, from Lemma B.1 we get that for all i, $\lambda_i \leq 2.4$ and by Assumption 2.3 we get that for all $i, k \in [m]$ $\langle \mathbf{x}_i, \mathbf{x}_k \rangle \leq \phi$. Using Cauchy–Schwarz inequality we get that for all $i \in [m]$, $\langle \mathbf{x}_i, \mathbf{v}_\epsilon \rangle \leq \|\mathbf{x}_i\| \, \|\mathbf{v}_\epsilon\| \leq \epsilon\sqrt{1+\psi}$. Plug it all in we have

$$
\begin{aligned}
\|\hat{\mathbf{w}}\|^2 &\leq \sum_{i\in[m]_{-l}} \lambda_i^2 \|x_i\|^2 + \sum_{i\neq k\in[m]_{-l}} \lambda_i\lambda_k\langle \mathbf{x}_i, \mathbf{x}_k\rangle + 2\sum_{i\in[m]_{-l}} \lambda_i\langle \mathbf{x}_i, \mathbf{v}_\epsilon\rangle + \epsilon^2 \\
&\leq (1+1.2\epsilon_d+2.15\epsilon+2.2\delta)\sum_{i\in[m]_{-l}} \lambda_i + 2.4m\phi\sum_{i\in[m]_{-l}} \lambda_i + \epsilon\sqrt{1+\psi}\sum_{i\in[m]_{-l}} \lambda_i + \epsilon^2 \\
&\leq \sum_{i\in[m]_{-l}} \lambda_i\left((1+1.2\epsilon_d+2.15\epsilon+2.2\delta)+2.4m\phi+\epsilon\sqrt{1+\psi}\right)+\epsilon^2
\end{aligned}
$$

We denote $\Lambda = \sum_{i\in[m]_{-l}} \lambda_i$ and plug in $\phi \leq \frac{\epsilon_d}{4m}$ and $\psi \leq 0.1$ and get

$$
\begin{aligned}
\|\hat{\mathbf{w}}\|^2 &\leq \sum_{i\in[m]_{-l}} \lambda_i\left((1+1.2\epsilon_d+2.15\epsilon+2.2\delta)+2.4m\phi+\epsilon\sqrt{1+\psi}\right)+\epsilon^2 \\
&\leq \Lambda\left((1+1.2\epsilon_d+2.15\epsilon+2.2\delta)+0.6\epsilon_d+1.1\epsilon\right)+\epsilon^2 \\
&\leq \Lambda\left(1+1.8\epsilon_d+3.25\epsilon+2.2\delta\right)+\epsilon^2
\end{aligned}
$$

For the upper bound of $\|\mathbf{w}^*\|^2$ we do similar calculations, using 7 and Lemma B.1 getting

$$
\begin{aligned}
\|\mathbf{w}^*\|^2 &= \left\|\sum_{i\in[m]_{-l}} \lambda_i^* y_i\mathbf{x}_i\right\|^2 = \langle \sum_{i\in[m]_{-l}} \lambda_i^* y_i\mathbf{x}_i, \sum_{i\in[m]_{-l}} \lambda_i^* y_i\mathbf{x}_i\rangle \\
&\leq \sum_{i\in[m]_{-l}} (\lambda_i^*)^2\|\mathbf{x}_i\|^2 + \sum_{i\neq k\in[m]_{-l}} \lambda_i^*\lambda_k^*\langle \mathbf{x}_i, \mathbf{x}_k\rangle \\
&\leq (1+1.2\epsilon_d+2.15\epsilon+2.2\delta)\sum_{i\in[m]_{-l}} \lambda_i^* + 2.4m\phi\sum_{i\in[m]_{-l}} \lambda_i^*
\end{aligned}
$$

W.L.O.G, we assume that $\sum_{i\in[m]_{-l}} \lambda_i \geq \sum_{i\in[m]_{-l}} \lambda_i^*$ (the other direction is proven similarly). This allow as to upper bound $\|\mathbf{w}^*\|^2$ using λ_i, with plugging in $\phi \leq \frac{\epsilon_d}{4m}$, we get

$$
\begin{aligned}
\|\mathbf{w}^*\|^2 &\leq (1+1.2\epsilon_d)\sum_{i\in[m]_{-l}} \lambda_i^* + 2.4m\phi\sum_{i\in[m]_{-l}} \lambda_i^* \\
&\leq (1+1.2\epsilon_d)\sum_{i\in[m]_{-l}} \lambda_i + 2.4m\phi\sum_{i\in[m]_{-l}} \lambda_i \\
&\leq \Lambda\left(1+1.8\epsilon_d\right)
\end{aligned}
$$

For the norm multiplication we have

$$
\begin{aligned}
\|\hat{\mathbf{w}}\|\,\|\mathbf{w}^*\| &= \sqrt{\|\hat{\mathbf{w}}\|^2\|\mathbf{w}^*\|^2} = \sqrt{\left[\Lambda\left(1+1.8\epsilon_d+3.25\epsilon+2.2\delta\right)+\epsilon^2\right]\left[\Lambda\left(1+1.8\epsilon_d\right)\right]} \\
&\leq \Lambda\sqrt{(1+C(\epsilon_d+\epsilon+\delta))+\frac{\epsilon^2}{\Lambda}(1+C\epsilon_d)} \\
&\leq \Lambda\sqrt{1+C(\epsilon_d+\epsilon+\delta)+\frac{\epsilon^2}{\Lambda}+\frac{\epsilon^2}{\Lambda}C\epsilon_d} \\
&\leq \Lambda+\Lambda\sqrt{C(\epsilon_d+\epsilon+\delta)+\frac{\epsilon^2}{\Lambda}+\frac{\epsilon^2}{\Lambda}C\epsilon_d}
\end{aligned}
$$

for some constant $C > 0$, where the last inequality hold since $1+\sqrt{x} \geq \sqrt{1+x}$ for all $x > 0$. We next lower bound the inner product of $\hat{\mathbf{w}}$ and $\mathbf{w}^*$

$$
\begin{aligned}
\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle &= \langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon, \sum_{i\in[m]_{-l}} \lambda_i^* y_i \mathbf{x}_i \rangle = \\
&= \langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \lambda_i^* y_i \mathbf{x}_i \rangle + \langle \sum_{i\in[m]_{-l}} \lambda_i^* y_i \mathbf{x}_i, \mathbf{v}_\epsilon \rangle \\
&\geq \sum_{i\in[m]_{-l}} \lambda_i^* \lambda_i \left\|\mathbf{x}_i\right\|^2 - \sum_{i\neq k\in[m]_{-l}} \lambda_i^* \lambda_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle - \sum_{i\in[m]_{-l}} \lambda_i^* \langle \mathbf{x}_i, \mathbf{v}_\epsilon \rangle
\end{aligned}
$$

Here, we use the lower bound for $\lambda_i^* \left\|\mathbf{x}_i\right\|^2 \geq (1 - 0.6\epsilon_d)$, the upper bound $\lambda_i^* \leq 2.4$ from Lemma B.1, and the Cauchy–Schwarz inequality, having

$$
\begin{aligned}
\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle &\geq \sum_{i\in[m]_{-l}} \lambda_i^* \lambda_i \left\|\mathbf{x}_i\right\|^2 - \sum_{i\neq k\in[m]_{-l}} \lambda_i^* \lambda_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle - \sum_{i\in[m]_{-l}} \lambda_i^* \langle \mathbf{x}_i, \mathbf{v}_\epsilon \rangle \\
&\geq (1 - 0.6\epsilon_d) \sum_{i\in[m]_{-l}} \lambda_i - 2.4m\phi \sum_{i\in[m]_{-l}} \lambda_i - \epsilon\sqrt{1+\psi} \sum_{i\in[m]_{-l}} \lambda_i
\end{aligned}
$$

and by plugging in $\phi \leq \frac{\epsilon_d}{4m}$, $\psi \leq 0.1$ we have

$$
\begin{aligned}
\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle &\geq (1 - 0.6\epsilon_d) \sum_{i\in[m]_{-l}} \lambda_i - 2.4m\phi \sum_{i\in[m]_{-l}} \lambda_i - \epsilon\sqrt{1+\psi} \sum_{i\in[m]_{-l}} \lambda_i \\
&\geq \Lambda \left(1 - 0.6\epsilon_d - 0.6\epsilon_d - 1.1\epsilon\right) \\
&\geq \Lambda - \Lambda \left(1.2\epsilon_d + 1.1\epsilon\right)
\end{aligned}
$$

Join all the bounds toghter, we get for the cosine similarity

$$
\begin{aligned}
\frac{\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle}{\|\hat{\mathbf{w}}\| \|\mathbf{w}^*\|} &\geq \frac{\Lambda - \Lambda \left(1.2\epsilon_d + 1.1\epsilon\right)}{\Lambda + \Lambda\sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{\Lambda} + \frac{\epsilon^2}{\Lambda} C\epsilon_d}} \\
&\geq 1 - \frac{\Lambda \left(1.2\epsilon_d + 1.1\epsilon\right) + \Lambda\sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{\Lambda} + \frac{\epsilon^2}{\Lambda} C\epsilon_d}}{\Lambda + \Lambda\sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{\Lambda} + \frac{\epsilon^2}{\Lambda} C\epsilon_d}} \\
&\geq 1 - \frac{\left(1.2\epsilon_d + 1.1\epsilon\right) + \sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{\Lambda} + \frac{\epsilon^2}{\Lambda} C\epsilon_d}}{1 + \sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{\Lambda} + \frac{\epsilon^2}{\Lambda} C\epsilon_d}} \\
&\geq 1 - \left(1.2\epsilon_d + 1.1\epsilon\right) - \sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{\Lambda} + \frac{\epsilon^2}{\Lambda} C\epsilon_d}
\end{aligned}
$$

We note that by Lemma B.2

$$
\begin{aligned}
\Lambda = \sum_{i\in[m]_{-l}} \lambda_i \geq &(m-1) \left(\frac{1}{\left\|\mathbf{x}_t\right\|^2} - \frac{0.6\epsilon_d + 1.1\epsilon}{\left\|\mathbf{x}_t\right\|^2} \right) \\
\geq &(m-1) 0.9 \left(1 - 0.6\epsilon_d - 1.1\epsilon\right) \\
\geq &0.1(m-1) \,,
\end{aligned}
$$

Concluding,

$$\frac{\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle}{\|\hat{\mathbf{w}}\| \|\mathbf{w}^*\|} \geq 1 - (1.2\epsilon_d + 1.1\epsilon) - \sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{0.1(m-1)} + \frac{\epsilon^2}{0.1(m-1)} C\epsilon_d}$$
$$\geq 1 - C_2 \left(\sqrt{\epsilon_d} + \sqrt{\epsilon} + \sqrt{\delta} \right)$$

for some constant $C_2 > 0$. □

B.2 Proof for forgetting subset of points using $\mathcal{A}_{\text{K-GA}}$ – linear predictors

We formalize and prove the statement for unlearning a subset of data points. Here, the term *successful unlearning* is the natural extension of Definition 2.2 to unlearning a subset, rather than a single point.

Theorem B.1. *In the same settings as Theorem 3.1, let $S_{forget} \subseteq S$ be a subset of size k.*

*Then, the extended algorithm $\mathcal{A}_{K\text{-}GA}$, with appropriate coefficients $\{\beta_r\}$, is an (ϵ, δ, τ)-**successful** unlearning algorithm w.r.t. $\mathbf{w}$ and S, where:*

1. ***The case of** $\epsilon = \epsilon_1 + \frac{\epsilon_1 \epsilon_d}{\frac{m}{k} - \epsilon_d}$, $\delta = \delta_1 + \frac{\delta_1 \epsilon_d}{\frac{m}{k} - \epsilon_d} + \frac{7.2\epsilon_d}{m}$, $\tau = 0$:*
   The predictor $\mathcal{A}_{k\text{-}GA}(\mathbf{w}, S, l)$ has the direction of an (ϵ, δ)-approximate KKT point for the margin maximization problem (2) w.r.t. $S \setminus (\mathbf{x}_l, y_l)$.

2. ***The case of** $\epsilon = \delta = 0$, $\tau = C(\sqrt{\epsilon_d} + \sqrt{\epsilon_1} + \sqrt{\delta_1})$ **for some universal constant** $C > 0$:*
   Let $\mathbf{w}^*$ be a max-margin linear predictor w.r.t. the remaining training set $S \setminus (\mathbf{x}_l, y_l)$, i.e. the global optimum of the 2 w.r.t. $S \setminus (\mathbf{x}_l, y_l)$. Then, $\text{cossim}(\mathcal{A}_{k\text{-}GA}(\mathbf{w}, S, l), \mathbf{w}^*) \geq 1 - \tau$.

Proof: Let a forget set $S_f \subset S$ such that $|S_f| = k$. We denote $I_f = \{i : (\mathbf{x}_i, y_i) \in S_f\}$. We denote $S_r = S \setminus S_f$ and $I_r = \{i : (\mathbf{x}_i, y_i) \in S_r\}$. The proof is highly similar to the proof for unlearning single point in B.1.

Similarly, we denote $\mathbf{v}_\epsilon = \mathbf{w} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i)$, so we get that $\|\mathbf{v}_\epsilon\| \leq \epsilon$ and

$$\mathbf{w} = \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \,.$$

According to the algorithm $\mathcal{A}_{\text{k-GA}}$, we take a step consists of the sum of k gradients w.r.t. data points in S_f with the following sizes- For any $(\mathbf{x}_l, y_l) \in S_f$, we sum a gradient of size $\beta = \frac{-\lambda_l}{\ell'(y_l N(\mathbf{w}, \mathbf{x}_l))}$. We get

$$\hat{\mathbf{w}} = \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon - \sum_{l \in I_f} \lambda_l y_l \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_r) = \sum_{i \in I_r} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon \,.$$

Proof of 1. $\hat{\mathbf{w}}$ has the direction of an $(\epsilon + \frac{\epsilon\epsilon_d}{\frac{m}{k} - \epsilon_d}, \delta + \frac{\delta\epsilon_d}{\frac{m}{k} - \epsilon_d} + \frac{7.2k\epsilon_d}{m})$-approximate KKT point for the margin maximization problem for $S \setminus (\mathbf{x}_l, y_l)$.

(1) Dual Feasibility: For all $i \in [m]_{-l}$, $\lambda_i \geq 0$. Same. directly from dual feasibility for $\mathbf{w}$ (Definition 2.1).

(2) Stationarity: $\left\| \hat{\mathbf{w}} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\hat{\mathbf{w}}, \mathbf{x}_i) \right\| \leq \epsilon$. Same as in B.1.

(3) Complementarity Slackness: For all $t \in [m]_{-l}$, $\lambda_t \left(y_t N(\hat{\mathbf{w}}, \mathbf{x}_t) - 1 \right) \leq \delta + \frac{1.44k\epsilon_d}{m}$. Using the same Equation 5 we get

$$
\begin{aligned}
1+\frac{\delta}{\lambda_t} \geq y_t N(\mathbf{w},\mathbf{x}_t) =& y_t N(\hat{\mathbf{w}},\mathbf{x}_t) + y_t \sum_{l\in I_f} \lambda_l y_l |\langle \mathbf{x}_l,\mathbf{x}_t\rangle| \\
\geq& y_t N(\hat{\mathbf{w}},\mathbf{x}_t) - \sum_{l\in I_f} \lambda_l |\langle \mathbf{x}_l,\mathbf{x}_t\rangle| \\
\geq& y_t N(\hat{\mathbf{w}},\mathbf{x}_t) - k\phi \max_p \lambda_p \ ,
\end{aligned}
$$

plugging in $\phi \leq \frac{\epsilon_d}{4m}$ and the λ_p upper bound from Lemma B.1 we get

$$
y_t N(\hat{\mathbf{w}},\mathbf{x}_t) - k\phi \max_p \lambda_p \geq y_t N(\hat{\mathbf{w}},\mathbf{x}_t) - k\frac{\epsilon_d}{4m} 2.4 \geq y_t N(\hat{\mathbf{w}},\mathbf{x}_t) - \frac{0.6k\epsilon_d}{m} \ .
$$

We deduce an upper bound for the margin of $N(\hat{\mathbf{w}},\mathbf{x}_t)$-

$$
y_t N(\hat{\mathbf{w}},\mathbf{x}_t) \leq 1+\frac{\delta}{\lambda_t}+\frac{0.6k\epsilon_d}{m} = 1+\frac{\delta+\frac{3}{5m}k\lambda_t\epsilon_d}{\lambda_t} \leq 1+\frac{\delta+\frac{3}{5m}k2.4\epsilon_d}{\lambda_t} \leq 1+\frac{\delta+\frac{1.44k\epsilon_d}{m}}{\lambda_t}
$$

as desired.

(4) Primal Feasibility: For all $t \in [m]_{-l}$, $y_t N(\hat{\mathbf{w}},\mathbf{x}_t) \geq 1-\frac{0.6k\epsilon_d}{m}$. We use 5 to lower bound the margin of $N(\hat{\mathbf{w}},\mathbf{x}_t)$, and use primal feasibility for $\mathbf{w}$ (Definition 2.1), getting

$$
y_t N(\hat{\mathbf{w}},\mathbf{x}_t) = y_t N(\mathbf{w},\mathbf{x}_t) - y_t \sum_{l\in I_f} \lambda_l y_l |\langle \mathbf{x}_l,\mathbf{x}_t\rangle| \geq 1 - k\phi \max_p \lambda_p \ .
$$

Plugging in $\phi \leq \frac{\epsilon_d}{4m}$ and the λ_p upper bound from Lemma B.1 we get that

$$
k\phi \max_p \lambda_p \leq \frac{2.4k\epsilon_d}{4m} \leq \frac{0.6k\epsilon_d}{m} \ .
$$

Hence, $y_t N(\hat{\mathbf{w}},\mathbf{x}_t) \geq 1-\frac{0.6k\epsilon_d}{m}$.

We showed that $\hat{\mathbf{w}}$ is an $(\epsilon, \delta+\frac{1.44k\epsilon_d}{m}, \frac{0.6k\epsilon_d}{m})$-approximate KKT by Definition B.1 . Finally, we look at the scaled weights $\frac{1}{1-\frac{0.6k\epsilon_d}{m}}\hat{\mathbf{w}}$. For $\epsilon_d \leq 1$ We calculate

$$
\frac{1}{1-\frac{0.6k\epsilon_d}{m}}\epsilon \leq \frac{\frac{m}{k}}{\frac{m}{k}-\epsilon_d}\epsilon = \left(1+\frac{\epsilon_d}{\frac{m}{k}-\epsilon_d}\right)\epsilon = \epsilon + \frac{\epsilon\epsilon_d}{\frac{m}{k}-\epsilon_d} \ ,
$$

and

$$
\max_p \lambda_p \frac{\frac{0.6k\epsilon_d}{m}}{1-\frac{0.6k\epsilon_d}{m}} + \frac{\delta+\frac{1.44k\epsilon_d}{m}}{1-\frac{0.6k\epsilon_d}{m}} \leq \delta + \frac{\delta\epsilon_d}{\frac{m}{k}-\epsilon_d} + \frac{7.2k\epsilon_d}{m}
$$

and get from Lemma B.3 that $\frac{1}{1-\frac{0.6k\epsilon_d}{m}}\hat{\mathbf{w}}$ is a $(\epsilon+\frac{\epsilon\epsilon_d}{\frac{m}{k}-\epsilon_d}, \delta+\frac{\delta\epsilon_d}{\frac{m}{k}-\epsilon_d}+\frac{7.2k\epsilon_d}{m})$-approximate KKT by Definition 2.1 w.r.t. $S\setminus(\mathbf{x}_l,y_l)$. We note that $\hat{\mathbf{w}}$ and $\frac{1}{1-0.6k\frac{\epsilon_d}{m}}\hat{\mathbf{w}}$ have the same direction, which finishes the proof.

PROOF OF 2. $Cosine-Similarity(\hat{\mathbf{w}},\mathbf{w}^*) \geq 1-C(\sqrt{\epsilon_d}+\sqrt{\epsilon_d}+\sqrt{\delta})$ FOR SOME $C>0$.

Let $N(\mathbf{w}^*,\mathbf{x})$ be a max-margin linear predictor w.r.t. the remaining training set $S\setminus S_f$. Hence, $\mathbf{w}^*$ is a KKT point of the margin maximization problem (2) w.r.t. $\{\mathbf{x}_i,y_i\}_{i\in I_f}$, as in Definition 2.1 (with $\epsilon=\delta=0$). From the stationarity condition we denote $\mathbf{w}^* = \sum_{i\in I_f}\lambda_i^* y_i \mathbf{x}_i$. We have same bounds for λ_i and λ_i^*, since it is independent of the unlearning.

The rest of the proof remains the same but the substitution of $\sum_{i\in[m]_{-l}} \lambda_i$ in $\sum_{i\in I_r} \lambda_i$, and the lower bound for it - by Lemma B.2

$$\begin{aligned}\Lambda = \sum_{i\in I_r} \lambda_i \geq &(m-k)\left(\frac{1}{\|\mathbf{x}_t\|^2} - \frac{0.6\epsilon_d + 1.1\epsilon}{\|\mathbf{x}_t\|^2}\right)\\ \geq &(m-k)0.9\left(1 - 0.6\epsilon_d - 1.1\epsilon\right)\\ \geq &0.1(m-k)\ ,\end{aligned}$$

That have no significant effect on the final bound

$$\begin{aligned}\frac{\langle \hat{\mathbf{w}}, \mathbf{w}^* \rangle}{\|\hat{\mathbf{w}}\|\,\|\mathbf{w}^*\|} &\geq 1 - (1.2\epsilon_d + 1.1\epsilon) - \sqrt{C(\epsilon_d + \epsilon + \delta) + \frac{\epsilon^2}{0.1(m-k)} + \frac{\epsilon^2}{0.1(m-k)} C(\epsilon_d + \epsilon + \delta)}\\ &\geq 1 - C_2\left(\sqrt{\epsilon_d} + \sqrt{\epsilon} + \sqrt{\delta}\right)\end{aligned}$$

for some constant $C_2 > 0$. □

B.3 The Identity is an Unsuccessful Unlearning Algorithm

To complement Theorem 3.1, we provide the following remark, that shows that keeping the original predictor is not a successful unlearning algorithm. Particularly, for any $\epsilon', \delta' > 0$, we show that for the predictor as defined in Theorem 3.1, its cosine similarity to any (ϵ', δ')-approximate KKT point for $S \setminus \{(\mathbf{x}_l, y_l)\}$ is relatively large.

Remark B.1. *In the same settings as 3.1, the algorithm $\mathcal{A}_I(\boldsymbol{\theta}, S, r) = \boldsymbol{\theta}$, is (ϵ, δ, τ)-successful only for $\tau \geq \frac{C}{m} - C(\epsilon_d + \epsilon)$ for some $C > 0$.*

As a short intuition for the proof, we note that the original network weight parameter, denoted as

$$\mathbf{w} = \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_{\epsilon_1}\ ,$$

consists of a sum of m summons, while any other KKT point w.r.t. $S \setminus \{(\mathbf{x}_l, y_l)\}$, $\widetilde{\mathbf{w}}$, consists of a sum of the $(m-1)$ gradients of the remaining dataset. This gap creates an inevitable angle between the two vectors.

Proof: In this section, we show that the original network $\mathbf{w}$ is not a good candidate for the unlearning tasks according to the (ϵ, δ, τ)-successful definition (Definition 2.2). Formally, we look at the simple unlearning algorithm $\mathcal{A}_I(\mathbf{w}, S, r) = \mathbf{w}$. We show that for any (ϵ', δ')-approximate KKT point $\widetilde{\mathbf{w}}$, where $\epsilon', \delta' < 0.5$ and $\epsilon_d < 0.1$, there exists $C > 0$ such that

$$\text{cossim}(\mathbf{w}, \widetilde{\mathbf{w}}) \leq 1 - \frac{C}{m} + C(\epsilon_d + \epsilon + \widetilde{\epsilon})\ ,$$

leading to

$$\tau \geq \frac{C}{m} - C(\epsilon_d + \epsilon + \widetilde{\epsilon})\ .$$

We recall that due to the stationary condition for the original network $\mathbf{w}$ w.r.t. the full dataset S we have

$$\mathbf{w} = \sum_{i\in[m]} \lambda_i y_i \nabla_{\mathbf{w}} N(\mathbf{w}, \mathbf{x}_i) + \mathbf{v}_\epsilon = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i + \mathbf{v}_\epsilon\ .$$

We denote an $(\widetilde{\epsilon}, \widetilde{\delta})$-approximate KKT point of the margin maximization problem w.r.t. the retain dataset $S \setminus (\mathbf{x}_l, y_l)$ by $\widetilde{\mathbf{w}}$. From the stationarity condition we get that

$$\widetilde{\mathbf{w}} = \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i + \mathbf{v}_{\widetilde{\epsilon}}\ .$$

Next, we show that the cosine similarity between $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ is lower bounded by $\frac{C}{m} + C(\epsilon_d + \epsilon + \widetilde{\epsilon})$. We denote $\underline{\mathbf{w}} = \mathbf{w} - \mathbf{v}_\epsilon$ and $\underline{\widetilde{\mathbf{w}}} = \mathbf{w} - \mathbf{v}_{\widetilde{\epsilon}}$. For the cosine similarity between $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ we have

$$\operatorname{cossim}(\mathbf{w}, \widetilde{\mathbf{w}}) = \frac{\langle \mathbf{w}, \widetilde{\mathbf{w}} \rangle}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} = \frac{\langle \underline{\mathbf{w}} + \mathbf{v}_\epsilon, \underline{\widetilde{\mathbf{w}}} + \mathbf{v}_{\widetilde{\epsilon}} \rangle}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|}$$

We first use Cauchy–Schwarz inequality and separate it into two expressions

$$\begin{aligned}
\operatorname{cossim}(\mathbf{w}, \widetilde{\mathbf{w}}) &= \frac{\langle \underline{\mathbf{w}} + \mathbf{v}_\epsilon, \underline{\widetilde{\mathbf{w}}} + \mathbf{v}_{\widetilde{\epsilon}} \rangle}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} \\
&\leq \frac{\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}} \rangle}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} + \frac{|\langle \mathbf{v}_\epsilon, \underline{\widetilde{\mathbf{w}}} \rangle| + |\langle \mathbf{v}_{\widetilde{\epsilon}}, \underline{\mathbf{w}} \rangle| + |\langle \mathbf{v}_\epsilon, \mathbf{v}_{\widetilde{\epsilon}} \rangle|}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} \\
&\leq \frac{\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}} \rangle}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} + \frac{\|\mathbf{v}_\epsilon\| \, \|\underline{\widetilde{\mathbf{w}}}\| + \|\mathbf{v}_{\widetilde{\epsilon}}\| \, \|\underline{\mathbf{w}}\| + \|\mathbf{v}_\epsilon\| \, \|\mathbf{v}_{\widetilde{\epsilon}}\|}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|}
\end{aligned} \tag{8}$$

We next lower bound the norm of the parameter vectors. We note that

$$\|\mathbf{w}\| = \|\underline{\mathbf{w}} + \mathbf{v}_\epsilon\| \geq \|\underline{\mathbf{w}}\| - \epsilon$$

and

$$\begin{aligned}
\|\underline{\mathbf{w}}\|^2 = \left\| \sum_{i \in [m]} \lambda_i y_i \mathbf{x}_i \right\|^2 &= \langle \sum_{i \in [m]} \lambda_i y_i \mathbf{x}_i, \sum_{i \in [m]} \lambda_i y_i \mathbf{x}_i \rangle = \\
&\geq \sum_{i \in [m]} \lambda_i^2 \, \|x_i\|^2 - \sum_{i \neq k \in [m]} \lambda_i \lambda_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle \\
&\geq \sum_{i \in [m]} \lambda_i^2 \, \|x_i\|^2 - \phi \sum_{i \neq k \in [m]} \lambda_i \lambda_k \, .
\end{aligned}$$

Similarly $\|\widetilde{\mathbf{w}}\| \geq \|\underline{\widetilde{\mathbf{w}}}\| - \widetilde{\epsilon}$ and

$$\begin{aligned}
\|\underline{\widetilde{\mathbf{w}}}\|^2 = \left\| \sum_{i \in [m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i \right\|^2 &= \langle \sum_{i \in [m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i, \sum_{i \in [m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i \rangle \\
&\geq \sum_{i \in [m]_{-l}} \widetilde{\lambda}_i^{\,2} \, \|x_i\|^2 - \phi \sum_{i \neq k \in [m]_{-l}} \widetilde{\lambda}_i \widetilde{\lambda}_k \, .
\end{aligned}$$

We now upper bound the inner product $\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}} \rangle$, having

$$
\begin{aligned}
\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}} \rangle &= \langle \sum_{i\in[m]} \lambda_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i \rangle = \\
&= \langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i \rangle + \langle \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i, \lambda_l y_l \mathbf{x}_l \rangle \\
&\leq |\langle \sum_{i\in[m]_{-l}} \lambda_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i \rangle| + |\langle \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \mathbf{x}_i, \lambda_l y_l \mathbf{x}_l \rangle| \\
&\leq \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \left\|\mathbf{x}_i\right\|^2 + \sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda}_i \lambda_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_l \langle \mathbf{x}_i, \mathbf{x}_l \rangle \\
&\leq \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \left\|\mathbf{x}_i\right\|^2 + \phi \sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda}_i \lambda_k + \phi \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_l
\end{aligned}
$$

Plug it all in, we get for the first summon at 8

$$
\frac{\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}} \rangle}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} \leq
\frac{\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \left\|\mathbf{x}_i\right\|^2 + \phi \sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda}_i \lambda_k + \phi \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_l}{\left(\sqrt{\sum_{i\in[m]} \lambda_i^2 \left\|x_i\right\|^2 - \phi \sum_{i\neq k\in[m]} \lambda_i \lambda_k} - \epsilon\right) \left(\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|x_i\right\|^2 - \phi \sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda_i}\widetilde{\lambda_k}} - \widetilde{\epsilon}\right)} \, .
$$

We first note that by Cauchy–Schwarz

$$
\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \left\|\mathbf{x}_i\right\|^2 \leq \sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|\mathbf{x}_i\right\|^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \left\|\mathbf{x}_i\right\|^2} \, ,
$$

and

$$
\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \leq \sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2} \, .
$$

We now reduce the nominator and denominator by $\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|\mathbf{x}_i\right\|^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \left\|\mathbf{x}_i\right\|^2}$. We denote $b = (1 + 1.2\epsilon_d + 2.15\epsilon + 2.2\delta), a = (1 - 0.6\epsilon_d - 1.1\epsilon)$, and use Lemma B.2 in which for all i, $a < \lambda_i \left\|\mathbf{x}_i\right\|^2 < b$. We calculate the summons in the nominator after reduction, having

$$
\frac{\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \left\|\mathbf{x}_i\right\|^2}{\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|\mathbf{x}_i\right\|^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \left\|\mathbf{x}_i\right\|^2}} \leq 1 \, ,
$$

$$
\frac{\phi \sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda}_i \lambda_k}{\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|\mathbf{x}_i\right\|^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \left\|\mathbf{x}_i\right\|^2}} \leq \frac{\phi \sqrt{\sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda_i}^2} \sqrt{\sum_{i\neq k\in[m]_{-l}} \lambda_i^2}}{\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|\mathbf{x}_i\right\|^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \left\|\mathbf{x}_i\right\|^2}} \leq \frac{\epsilon_d}{3.6} \, ,
$$

$$
\frac{\phi \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_l}{\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\lambda_i}^2 \left\|\mathbf{x}_i\right\|^2} \sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \left\|\mathbf{x}_i\right\|^2}} \leq \frac{\phi \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_l}{\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i \lambda_i \left\|\mathbf{x}_i\right\|^2} \leq \frac{1.2b\epsilon_d}{4ma} \, .
$$

and for the denominator we have

$$\frac{\sum_{i\in[m]} \lambda_i^2 \|x_i\|^2}{\sum_{i\in[m]} \lambda_i^2 \|x_i\|^2} = 1 \,,$$

$$\frac{\phi \sum_{i\neq k\in[m]_{-l}} \lambda_i\lambda_k}{\sum_{i\in[m]_{-l}} \lambda_i^2 \|x_i\|^2} \leq \frac{\phi\sqrt{\sum_{i\neq k\in[m]_{-l}} {\lambda_i}^2}\sqrt{\sum_{i\neq k\in[m]_{-l}} \lambda_i^2}}{\sum_{i\in[m]_{-l}} \lambda_i^2 \|x_i\|^2} \leq \frac{\phi(m-1)\sum_{i\in[m]_{-l}} {\lambda_i}^2}{\sum_{i\in[m]_{-l}} \lambda_i^2 \|x_i\|^2} \leq \frac{\epsilon_d}{3.6} \,,$$

$$\frac{\epsilon}{\sqrt{\sum_{i\in[m]_{-l}} \lambda_i^2 \|x_i\|^2}} \leq \frac{\epsilon}{0.9a\sqrt{m}} \,,$$

the same for $\widetilde{\lambda}_i$ and $\widetilde{\epsilon}$, and finally

$$\frac{{\widetilde{\lambda}_l}^2 \|x_l\|^2}{\sum_{i\in[m]_{-l}} {\widetilde{\lambda}_i}^2 \|x_i\|^2} \leq \frac{2.4b}{0.91a^2m} \leq \frac{2.64b}{am} \,.$$

Plug it all in we have

$$\begin{aligned}
&\frac{\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}}\rangle}{\|\mathbf{w}\|\,\|\widetilde{\mathbf{w}}\|} \leq \\
&\frac{\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i\lambda_i \|\mathbf{x}_i\|^2 + \phi\sum_{i\neq k\in[m]_{-l}} \widetilde{\lambda}_i\lambda_k + \phi\sum_{i\in[m]_{-l}} \widetilde{\lambda}_i\lambda_l}{\sqrt{\sum_{i\in[m]} \lambda_i^2 \|x_i\|^2 - \phi\sum_{i\neq k\in[m]} \lambda_i\lambda_k - \epsilon}\sqrt{\sum_{i\in[m]} {\widetilde{\lambda}_i}^2 \|x_i\|^2 - \phi\sum_{i\neq k\in[m]} \widetilde{\lambda_i}\widetilde{\lambda_k} - \widetilde{\epsilon}}} \\
&\leq \frac{1 + 0.28\epsilon_d + \frac{1.2b\epsilon_d}{4ma}}{\sqrt{1 - 0.28\epsilon_d - \frac{\epsilon}{0.9a\sqrt{m}}}\sqrt{1 - 0.28\epsilon_d - \frac{\widetilde{\epsilon}}{0.9a\sqrt{m}} + \frac{2.64b}{am}}} \\
&\leq \frac{1 + 0.28\epsilon_d + \frac{1.2b\epsilon_d}{4ma}}{\left(1 - 0.28\epsilon_d - \frac{\epsilon+\widetilde{\epsilon}}{0.9a\sqrt{m}}\right)\sqrt{1 + \frac{2.64b}{am}}}
\end{aligned}$$

for any $0 < x < 1$ we get that

$$\frac{1}{\sqrt{1+x}} \leq 1 - \frac{x}{4}$$

and thus in conclusion we have

$$\begin{aligned}
&\frac{\langle \underline{\mathbf{w}}, \underline{\widetilde{\mathbf{w}}}\rangle}{\|\mathbf{w}\|\,\|\widetilde{\mathbf{w}}\|} \leq \\
&\leq \frac{1 + 0.28\epsilon_d + \frac{1.2b\epsilon_d}{4ma}}{\left(1 - 0.28\epsilon_d - \frac{\epsilon+\widetilde{\epsilon}}{0.9a\sqrt{m}}\right)\sqrt{1 + \frac{2.64b}{am}}} \\
&\leq \frac{1 + 0.28\epsilon_d + \frac{1.2b\epsilon_d}{4ma}}{1 - 0.28\epsilon_d - \frac{\epsilon+\widetilde{\epsilon}}{0.9a\sqrt{m}}}\left(1 - \frac{0.66b}{am}\right) \\
&\leq \left(1 + \frac{0.56\epsilon_d + \frac{1.2b\epsilon_d}{4ma} + \frac{\epsilon+\widetilde{\epsilon}}{0.9a\sqrt{m}}}{1 - 0.28\epsilon_d - \frac{\epsilon+\widetilde{\epsilon}}{0.9a\sqrt{m}}}\right)\left(1 - \frac{0.66b}{am}\right) \\
&\leq 1 - \frac{C}{m} + C(\epsilon_d + \epsilon + \widetilde{\epsilon}) \,,
\end{aligned}$$

which finishes the upper bounded the first summon of the cosine similarity at 8. We now upper bound the second summon, we recall that $\underline{\mathbf{w}} = \mathbf{w} - \mathbf{v}_\epsilon$ and therefore $\|\underline{\mathbf{w}}\| \leq \|\mathbf{w}\| + \epsilon$, and similar for $\widetilde{\mathbf{w}}$, and thus,

$$\frac{\|\mathbf{v}_\epsilon\| \, \|\underline{\widetilde{\mathbf{w}}}\| + \|\mathbf{v}_{\widetilde{\epsilon}}\| \, \|\underline{\mathbf{w}}\| + \|\mathbf{v}_\epsilon\| \, \|\mathbf{v}_{\widetilde{\epsilon}}\|}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} \leq \frac{\epsilon \, \|\widetilde{\mathbf{w}}\| + \epsilon^2 + \widetilde{\epsilon} \, \|\mathbf{w}\| + \widetilde{\epsilon}^2 + \epsilon\widetilde{\epsilon}}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} = \frac{\epsilon}{\|\mathbf{w}\|} + \frac{\widetilde{\epsilon}}{\|\widetilde{\mathbf{w}}\|} + \frac{\epsilon^2 + \widetilde{\epsilon}^2 + \epsilon\widetilde{\epsilon}}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|}$$

We look at the norm lower bound. We note that

$$\|\mathbf{w}\| = \|\underline{\mathbf{w}} + \mathbf{v}_\epsilon\| \geq \|\underline{\mathbf{w}}\| - \epsilon \, ,$$

and

$$\begin{aligned}
\|\underline{\mathbf{w}}\|^2 &= \langle \sum_{i \in [m]} \lambda_i y_i \mathbf{x}_i, \sum_{i \in [m]} \lambda_i y_i \mathbf{x}_i \rangle = \\
&\geq \sum_{i \in [m]} \lambda_i^2 \, \|x_i\|^2 - \phi \sum_{i \neq k \in [m]} \lambda_i \lambda_k \\
&\geq \sum_{i \in [m]} \lambda_i \, [a - \phi m b] \\
&\geq m 0.9 a \, [a - 0.6 \epsilon_d] \\
&\geq m 0.9 a \, [1 - 1.2 \epsilon_d - 1.1 \epsilon] \geq 0.1 m \, ,
\end{aligned}$$

and similarly $\|\underline{\widetilde{\mathbf{w}}}\|^2 \geq 0.1(m-1)$. Plug in to the denominator of the above fraction we get

$$\frac{\epsilon}{\|\mathbf{w}\|} + \frac{\widetilde{\epsilon}}{\|\widetilde{\mathbf{w}}\|} + \frac{\epsilon^2 + \widetilde{\epsilon}^2 + \epsilon\widetilde{\epsilon}}{\|\mathbf{w}\| \, \|\widetilde{\mathbf{w}}\|} \leq \frac{\epsilon}{0.1m - \epsilon} + \frac{\widetilde{\epsilon}}{0.1(m-1) - \widetilde{\epsilon}} + \frac{\epsilon^2 + \widetilde{\epsilon}^2 + \epsilon\widetilde{\epsilon}}{(0.1(m-1) - \epsilon)^2} \leq C_1(\epsilon_d + \epsilon + \widetilde{\epsilon})$$

which means that there exists C such that

$$\text{cossim}(\mathbf{w}, \widetilde{\mathbf{w}}) \leq 1 - \frac{C}{m} + C(\epsilon_d + \epsilon + \widetilde{\epsilon}) \, ,$$

Thus, concluding the proof. □

C Proofs for Section 4

C.1 Lemmas for Proof C.2 of Theorem 4.1

Lemma C.1. *Let $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$ such that $\forall i \in [m], \mathbf{x}_i \in \mathbb{R}^d$ and let $\{\mathbf{w}_j\}_{j=1}^n$, $\forall j \in [n], \mathbf{w}_j \in \mathbb{R}^d$. Assume the data distribution $\mathcal{D}$ satisfies Assumption 2.3 for some ψ, ϕ. Given $l \in [m]$ and $c \in \mathbb{R}$, for $j \in [n]$ and $r \in [m]_{-l}$, we denote*

$$\Delta_{r,j} = \sum_{k \in [m]_{-l}} c \langle \mathbf{x}_k, \mathbf{x}_r \rangle \, \text{sign}(\langle \mathbf{x}_k, \mathbf{w}_j \rangle) \, .$$

Then,

$$\begin{aligned}
&\mathbf{w}_j^\top \mathbf{x}_r \geq 0 \Rightarrow \\
&\qquad c(1-\psi) - (m-2)c\phi \leq \Delta_{r,j} \leq c(1+\psi) + (m-2)c\phi \\
&\mathbf{w}_j^\top \mathbf{x}_r < 0 \Rightarrow \\
&\qquad -c(1+\psi) - (m-2)c\phi \leq \Delta_{r,j} \leq -c(1-\psi) + (m-2)c\phi
\end{aligned}$$

Proof:

$$\sum_{k\in[m]_{-l}} c\langle \mathbf{x}_k, \mathbf{x}_r\rangle \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle) =$$
$$= c\left\|\mathbf{x}_r\right\|^2 \operatorname{sign}(\langle \mathbf{x}_r, \mathbf{w}_j\rangle) + \sum_{k\in[m]_{-l}, k\neq r} c\langle \mathbf{x}_k, \mathbf{x}_r\rangle \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle)$$

From Assumption 2.3 we know that $(1-\psi) \leq \left\|\mathbf{x}_r\right\|^2 \leq (1+\psi)$, for $k \neq r$, $-\phi \leq \langle \mathbf{x}_k, \mathbf{x}_r\rangle \leq \phi$ which finishes the proof. □

Lemma C.2. *Let $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$ such that $\forall i \in [m], \mathbf{x}_i \in \mathbb{R}^d$ and let $\{\mathbf{w}_j\}_{j=1}^n$ $\forall j \in [n], \mathbf{w}_j \in \mathbb{R}^d$. Assume the data distribution $\mathcal{D}$ satisfies Assumption 2.3 for some $\psi \leq 0.1, \phi \leq \frac{\epsilon_d}{4mn}$. Given $l \in [m]$, and $c = \frac{\epsilon_d}{2mn}$, for $j \in [n]$ and $r \in [m]_{-l}$, we denote*

$$\Delta_j = \sum_{k\in[m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle)$$

Then for $j \in [n]$,

$$\left\||u_j|\lambda_l \sigma'_{l,j}\Delta_j\right\| \leq \frac{22\epsilon_d}{\sqrt{mn}}$$

Proof: We first look at the norm of Δ_j, having

$$\left\|\Delta_j\right\|^2 = \left\|\sum_{k\in[m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle)\right\|^2 =$$
$$= \langle \sum_{k\in[m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle), \sum_{k\in[m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle)\rangle$$
$$\leq c^2 \langle \sum_{k\in[m]_{-l}} \mathbf{x}_k, \sum_{k\in[m]_{-l}} \mathbf{x}_k\rangle$$
$$\leq c^2 \left[\sum_{k\in[m]_{-l}} \left\|\mathbf{x}_i\right\|^2 + \sum_{s\neq k\in[m]_{-l}} \langle \mathbf{x}_k, \mathbf{x}_s\rangle\right]$$
$$\leq c^2 \left(m(1+\psi) + m^2\phi\right)$$

we plug in $\psi \leq 0.1, \phi \leq \frac{\epsilon_d}{4mn}, c = \frac{\epsilon_d}{2mn}$ and get

$$\left\|\Delta_j\right\|^2 \leq \frac{\epsilon_d^2}{4m^2n^2}\left(1.1m + m^2\frac{\epsilon_d}{4mn}\right) = \frac{\epsilon_d^2\left(1.1 + \frac{\epsilon_d}{4n}\right)}{4mn^2}$$

and

$$\left\|\Delta_j\right\| \leq \frac{\epsilon_d\sqrt{1.1 + \frac{\epsilon_d}{n}}}{2\sqrt{m}n}$$

From Lemma C.3 we have that $\max_{i\in[m]} \lambda_i \leq 20.4n$. As for all $j \in [n], |u_j| = \frac{1}{\sqrt{n}}$, and $\sigma'_{l,j} \geq 0$, joining all together we have

$$\begin{aligned}\left\||u_j|\lambda_l\sigma'_{l,j}\Delta_j\right\| = |u_j|\lambda_l\sigma'_{l,j}\left\|\Delta_j\right\| &\leq \frac{1}{\sqrt{n}}20.4n\frac{\epsilon_d\sqrt{1.1+\frac{\epsilon_d}{n}}}{2\sqrt{m}n} \\ &\leq \frac{1}{\sqrt{n}}20.4\frac{\epsilon_d\sqrt{1.1+\frac{\epsilon_d}{n}}}{2\sqrt{m}} \\ &\leq \frac{\epsilon_d\left(20.4 + \frac{1}{2}\sqrt{1.1+\frac{\epsilon_d}{n}}\right)}{\sqrt{nm}} \leq \frac{22\epsilon_d}{\sqrt{mn}},\end{aligned}$$

as desired. □

Lemma C.3. *Let $N(\boldsymbol{\theta}, \mathbf{x}) = \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x})$ be a two-layer fully connected neural network, trained on $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, and let $0 < \epsilon_d, \epsilon, \delta \leq 1$ such that $\boldsymbol{\theta}$ is an (ϵ, δ)-approximate KKT point for the margin maximization problem for S according to Definition 2.1 for $\lambda_1, ..., \lambda_m$, and S satisfies Assumption 2.3 for $\psi = 0, 1$, and $\phi \leq \frac{\epsilon_d}{4mn}$. Assume $\forall j \in [n], u_j \sim \mathcal{U}\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}$. Then, For $i \in [m]$ we have*

$$\max\left\{\sum_{j\in J_+} u_j^2 \lambda_i \sigma'_{i,j}, \sum_{j\in J_-} u_j^2 \lambda_i \sigma'_{i,j}\right\} \leq 2.5 + 5.25\epsilon + 2.4\delta \leq 10.2 \ ,$$

and therefore also

$$\sum_{j=1}^{n} u_j^2 \lambda_i \sigma'_{i,j} \leq 5 + 10.5\epsilon + 4.8\delta \leq 20.4 \ ,$$

and

$$\lambda_i \leq n\left(5 + 10.5\epsilon + 4.8\delta\right) \leq 20.4n \ .$$

Proof: Let $J_+ = \{j \in [n] : u_j > 0\}$ and $J_- = \{j \in [n] : u_j < 0\}$. Denote $\alpha_+ = \max_{i\in[m]}\left(\sum_{j\in J_+} u_j^2 \lambda_i \sigma'_{i,j}\right)$ and $\alpha_- = \max_{i\in[m]}\left(\sum_{j\in J_-} u_j^2 \lambda_i \sigma'_{i,j}\right)$. w.l.o.g. we assume $\alpha_+ \geq \alpha_-$ (the other direction is proven similarly). We denote $\alpha = \alpha_+ = \max_{i\in[m]}\left(\sum_{j\in J_+} u_j^2 \lambda_i \sigma'_{i,j}\right)$, and $k = \arg\max_{i\in[m]}\left(\sum_{j\in J_+} u_j^2 \lambda_i \sigma'_{i,j}\right)$. If $\lambda_k = 0$ the claim follows.

Using the stationarity condition in Definition 2.1 for $\boldsymbol{\theta}$, we denote $\mathbf{v}_\epsilon = \boldsymbol{\theta} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i)$, and $\mathbf{v}_{\epsilon,j} = \mathbf{w}_j - \sum_{i=1}^{m} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i$, such that $\mathbf{v}_\epsilon$ is the concatenation of all $\mathbf{v}_{\epsilon,j}$ and $\|\mathbf{v}_\epsilon\| = \epsilon$. Using this notation we have for all $j \in [n]$ the inner product

$$\begin{aligned}\mathbf{w}_j^\top \mathbf{x}_k =& u_j \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k\rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k\rangle \\ =& u_j \lambda_k y_k \sigma'_{k,j} \|\mathbf{x}_k\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k\rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k\rangle \ .\end{aligned}$$

To upper bound α, we use the Complementarity Slackness condition in Definition 2.1 to first bound the margin, and then solve for α. First, since for all $j \in [n]$ and $k \in [m]$, $|u_j| = \frac{1}{\sqrt{n}}$ and $\sigma'_{k,j} \leq 1$, we get that $\alpha \leq \lambda_k \frac{1}{n} \sum_{j=1}^{n} \sigma'_{k,j} \leq \lambda_k$, so $\frac{1}{\lambda_k} \leq \frac{1}{\alpha}$.

Then, using the Complementarity Slackness condition for $\boldsymbol{\theta}$ we get that $y_k N(\boldsymbol{\theta}, \mathbf{x}_k) \leq 1 + \frac{\delta}{\lambda_k} \leq 1 + \frac{\delta}{\alpha}$. To use the α notation we express the margin with in terms of sums over J_+ and J_-

$$1 + \frac{\delta}{\alpha} \geq y_k N(\boldsymbol{\theta}, \mathbf{x}_k) = y_k \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) = y_k \left[\sum_{j\in J_+} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) + \sum_{j\in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k)\right] \ .$$

Now, to divide both sides of the inequality by y_k, we need to know its sign. We separate to two cases for y_k:

Case 1: $y_k = 1$
We lower bound the margin

$$\begin{aligned}1 + \frac{\delta}{\alpha} \geq N(\boldsymbol{\theta}, \mathbf{x}_k) &= \sum_{j\in J_+} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) + \sum_{j\in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) \\ &\geq \sum_{j\in J_+} u_j \mathbf{w}_j^\top \mathbf{x}_k + \sum_{j\in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) \ ,\end{aligned}$$

Where the last inequality hold since for all $y \in \mathbb{R}$, $y \leq \sigma(y)$. We lower bound separately the first summand, getting

$$\begin{aligned}\sum_{j\in J_+} u_j \mathbf{w}_j^\top \mathbf{x}_k = \sum_{j\in J_+} u_j \left(u_j \lambda_k \sigma'_{k,j} \|\mathbf{x}_k\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle \right) \\ \geq (1-\psi) \sum_{j\in J_+} u_j^2 \lambda_k \sigma'_{k,j} - \phi \sum_{j\in J_+} \sum_{i=1, i\neq k}^{m} u_j^2 \lambda_i \sigma'_{i,j} - \sum_{j\in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \\ \geq (1-\psi)\alpha - \phi(m-1)\alpha - \sum_{j\in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \,.\end{aligned}$$

Using Cauchy–Schwarz inequality we have

$$\sum_{j\in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| = \frac{1}{\sqrt{n}} \sum_{j\in J_+} |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \leq \frac{1}{\sqrt{n}} \|\mathbf{v}_\epsilon\| \sqrt{n} \max_{p\in[m]} \|x_p\| \leq \epsilon \sqrt{1+\psi} \,,$$

getting

$$\sum_{j\in J_+} u_j \mathbf{w}_j^\top \mathbf{x}_k \geq (1-\psi)\alpha - \phi(m-1)\alpha - \epsilon\sqrt{1+\psi} \,.$$

Bounding the second summand we have

$$\begin{aligned}\sum_{j\in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) &= \sum_{j\in J_-} u_j \sigma \left(u_j \lambda_k \sigma'_{k,j} \|\mathbf{x}_k\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle \right) \\ &\geq \sum_{j\in J_-} u_j \sigma \left(u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \right) \\ &\geq \sum_{j\in J_-} u_j \sigma \left(|u_j| \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} |\langle \mathbf{x}_i, \mathbf{x}_k \rangle| + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \right) \\ &\geq \sum_{j\in J_-} u_j \sigma \left(|u_j| \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} \phi + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \right) \\ &\geq -\phi \sum_{j\in J_-} \sum_{i=1, i\neq k}^{m} u_j^2 \lambda_i \sigma'_{i,j} - \sum_{j\in J_-} |u_j| |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \\ &\geq -\phi(m-1)\alpha - \epsilon\sqrt{1+\psi} \,,\end{aligned}$$

and combining the two results we have

$$\begin{aligned}1 + \frac{\delta}{\alpha} \geq 1 + \frac{\delta}{\lambda_k} \geq y_k N(\boldsymbol{\theta}, \mathbf{x}_k) &\geq \sum_{j\in J_+} u_j \mathbf{w}_j^\top \mathbf{x}_k + \sum_{j\in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) \\ &\geq (1-\psi)\alpha - \phi(m-1)\alpha - \phi(m-1)\alpha \\ &= \alpha\left((1-\psi) - 2\phi(m-1)\right) - 2\epsilon\sqrt{1+\psi} \,,\end{aligned}$$

getting

$$\alpha^2 \left((1-\psi) - 2\phi(m-1)\right) - \alpha\left(1 + 2\epsilon\sqrt{1+\psi}\right) - \delta \leq 0 \,.$$

Note, for our setting $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4mn}$ so we get that

$$(1-\psi) - 2\phi(m-1) \geq 0.9 - 2\frac{\epsilon_d}{4mn}(m-1) \geq 0.9 - \frac{\epsilon_d}{2n} > 0 \,,$$

hence solving for α we get

$$\alpha \leq \frac{1 + 2\epsilon\sqrt{1+\psi} + \sqrt{(1+2\epsilon\sqrt{1+\psi})^2 + 4\delta\left((1-\psi) - 2\phi(m-1)\right)}}{2\left((1-\psi) - 2\phi(m-1)\right)} \ .$$

Case 2: $y_k = -1$ is very similar.
First we have

$$-1 - \frac{\delta}{\alpha} = N(\boldsymbol{\theta}, \mathbf{x}_k) \leq \sum_{j \in J_+} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) + \sum_{j \in J_-} u_j \mathbf{w}_j^\top \mathbf{x}_k \ ,$$

for the first summand we get

$$\begin{aligned}
\sum_{j \in J_+} u_j \sigma\left(\mathbf{w}_j^\top \mathbf{x}_k\right) &= \sum_{j \in J_+} u_j \sigma\left(-u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle\right) \\
&\leq \sum_{j \in J_+} u_j \sigma\left(u_j \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} |\langle \mathbf{x}_i, \mathbf{x}_k \rangle| + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle|\right) \\
&\leq \sum_{j \in J_+} u_j \sigma\left(u_j \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} \phi + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle|\right) \\
&\leq \phi \sum_{i=1, i\neq k}^{m} \sum_{j \in J_+} u_j^2 \lambda_i \sigma'_{i,j} + \sum_{j \in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \\
&\leq \phi(m-1)\alpha + \epsilon\sqrt{1+\psi}
\end{aligned}$$

and for the second

$$\begin{aligned}
\sum_{j \in J_-} u_j \mathbf{w}_j^\top \mathbf{x}_k &= \sum_{j \in J_-} u_j \left(-u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle\right) \\
&\leq -(1-\psi) \sum_{j \in J_-} u_j^2 \lambda_k \sigma'_{k,j} + \phi \sum_{j \in J_-} \sum_{i=1, i\neq k}^{m} u_j^2 \lambda_i \sigma'_{i,j} + \sum_{j \in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \\
&\leq -(1-\psi)\alpha + \phi(m-1)\alpha + \epsilon\sqrt{1+\psi}
\end{aligned}$$

combining the two results leads to the same upper bound

$$\alpha \leq \frac{1 + 2\epsilon\sqrt{1+\psi} + \sqrt{(1+2\epsilon\sqrt{1+\psi})^2 + 4\delta\left((1-\psi) - 2\phi(m-1)\right)}}{2\left((1-\psi) - 2\phi(m-1)\right)}$$

We plug in $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4mn}$, $\epsilon_d \leq 1$, and get

$$\begin{aligned}
\alpha &\leq \frac{1 + 2\epsilon\sqrt{1+\psi} + \sqrt{(1+2\epsilon\sqrt{1+\psi})^2 + 4\delta\left((1-\psi) - 2\phi(m-1)\right)}}{2\left((1-\psi) - 2\phi(m-1)\right)} \\
&\leq \frac{1 + 2.1\epsilon + \sqrt{(1+2.1\epsilon)^2 + 4\delta(0.9)}}{2(0.9 - 2\frac{\epsilon_d}{4mn}(m-1))} \\
&\leq \frac{1 + 2.1\epsilon + (1+2.1\epsilon) + 1.9\delta}{2(0.9 - 2\frac{\epsilon_d}{4})} \\
&\leq \frac{2 + 4.2\epsilon + 1.9\delta}{0.8} \leq 2.5 + 5.25\epsilon + 2.4\delta \leq 10.2
\end{aligned}$$

meaning for all $i \in [m]$ we have

$$\max\left\{\sum_{j\in J_+} u_j^2\lambda_i\sigma'_{i,j}, \sum_{j\in J_-} u_j^2\lambda_i\sigma'_{i,j}\right\} \leq 2.5 + 5.25\epsilon + 2.4\delta$$

so

$$\sum_{j\in[n]} u_j^2\lambda_i\sigma'_{i,j} \leq 5 + 10.5\epsilon + 4.8\delta \leq 20.4$$

using the fact that for all $j \in [n]$ and $k \in [m]$, $|u_j| = \frac{1}{\sqrt{n}}$ and $\sigma'_{k,j} \leq 1$ we also get that

$$\lambda_i \leq \frac{5 + 10.5\epsilon + 4.8\delta}{\sum\limits_{j\in[n]} u_j^2\sigma'_{i,j}} \leq \frac{5 + 10.5\epsilon + 4.8\delta}{\frac{1}{n}} \leq n\left(5 + 10.5\epsilon + 4.8\delta\right) \leq 20.4n$$

□

Lemma C.4. *Let $N(\boldsymbol{\theta}, \mathbf{x}) = \sum\limits_{j=1}^{n} u_j\sigma(\mathbf{w}_j^\top\mathbf{x})$ be a two-layer fully connected neural network, trained on $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, and let $0 < \epsilon_d, \epsilon, \delta \leq 1$ such that $\boldsymbol{\theta}$ is an (ϵ, δ)-approximate KKT point for the margin maximization problem (2) for S according to Definition 2.1 for $\lambda_1, ..., \lambda_m$, and S satisfies Assumption 2.3 for $\psi = 0, 1$, and $\phi \leq \frac{\epsilon_d}{4mn}$. Assume $\forall j \in [n], u_j \sim \mathcal{U}\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}$. We denote $\alpha_{\max} = \max_{i\in[m]}\left(\max\left\{\sum\limits_{j\in J_+} u_j^2\lambda_i\sigma'_{i,j}, \sum\limits_{j\in J_-} u_j^2\lambda_i\sigma'_{i,j}\right\}\right)$. Then, For $i \in [m]$ we have*

$$\min\left\{\sum_{j\in J_+} u_j^2\lambda_i\sigma'_{i,j}, \sum_{j\in J_-} u_j^2\lambda_i\sigma'_{i,j}\right\} \geq 0.45 - 2.32\frac{\epsilon_d}{n} - 0.96\epsilon$$

and therefore also

$$\sum_{j=1}^{n} u_j^2\lambda_i\sigma'_{i,j} \geq 0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon$$

and

$$\lambda_i \geq 0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon$$

Proof: Let $J_+ = \{j \in [n] : u_j > 0\}$ and $J_- = \{j \in [n] : u_j < 0\}$. Denote $\alpha_+ = \min_{i\in[m]}\left(\sum\limits_{j\in J_+} u_j^2\lambda_i\sigma'_{i,j}\right)$ and $\alpha_- = \min_{i\in[m]}\left(\sum\limits_{j\in J_-} u_j^2\lambda_i\sigma'_{i,j}\right)$. w.l.o.g. we assume $\alpha_+ \leq \alpha_-$ (the other direction is proven similarly). We denote $\alpha = \alpha_+ = \min_{i\in[m]}\left(\sum\limits_{j\in J_+} u_j^2\lambda_i\sigma'_{i,j}\right)$, and $k = \arg\min_{i\in[m]}\left(\sum\limits_{j\in J_+} u_j^2\lambda_i\sigma'_{i,j}\right)$.

Using the stationarity condition in Definition 2.1 for $\boldsymbol{\theta}$, we denote $\mathbf{v}_\epsilon = \boldsymbol{\theta} - \sum_{i=1}^{m}\lambda_i y_i\nabla_{\boldsymbol{\theta}}N(\boldsymbol{\theta}, \mathbf{x}_i)$, and $\mathbf{v}_{\epsilon,j} = \mathbf{w}_j - \sum\limits_{i=1}^{m} u_j\lambda_i y_i\sigma'_{i,j}\mathbf{x}_i$, such that $\mathbf{v}_\epsilon$ is the concatenation of all $\mathbf{v}_{\epsilon,j}$ and $\|\mathbf{v}_\epsilon\| = \epsilon$. Using this notation we have for all $j \in [n]$ the inner product

$$\begin{aligned}\mathbf{w}_j^\top\mathbf{x}_k =& u_j\sum_{i=1}^{m}\lambda_i y_i\sigma'_{i,j}\langle\mathbf{x}_i, \mathbf{x}_k\rangle + \langle\mathbf{v}_{\epsilon,j}, \mathbf{x}_k\rangle \\ =& u_j\lambda_k y_k\sigma'_{k,j}\|\mathbf{x}_k\|^2 + u_j\sum_{i=1, i\neq k}^{m}\lambda_i y_i\sigma'_{i,j}\langle\mathbf{x}_i, \mathbf{x}_k\rangle + \langle\mathbf{v}_{\epsilon,j}, \mathbf{x}_k\rangle\,.\end{aligned}$$

To lower bound α, we use the primal feasibility condition in Definition 2.1 to first bound the margin, and then solve for α. To use the α notation we express the margin with in terms of sums over J_+ and J_-

$$1+\frac{\delta}{\alpha} \geq y_k N(\boldsymbol{\theta}, \mathbf{x}_k) = y_k \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) = y_k \left[\sum_{j\in J_+} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) + \sum_{j\in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) \right] .$$

Now, to divide both sides of the inequality by y_k, we need to know its sign. We separate to two cases for y_k:

Case 1: $y_k = 1$
We upper bound the margin

$$1 \leq N(\boldsymbol{\theta}, \mathbf{x}_k) \leq \sum_{j\in J_+} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) + \sum_{j\in J_-} u_j \mathbf{w}_j^\top \mathbf{x}_k$$

Where the last inequality hold since for all $y \in \mathbb{R}$, $y \leq \sigma(y)$. We lower bound separately the first summand, getting

$$\begin{aligned}
\sum_{j\in J_+} u_j \sigma\left(\mathbf{w}_j^\top \mathbf{x}_k\right) &= \sum_{j\in J_+} u_j \sigma\left(u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle \right) \\
&\leq \sum_{j\in J_+} u_j \sigma\left(u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} |\langle \mathbf{x}_i, \mathbf{x}_k \rangle| + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \right) \\
&\leq \sum_{j\in J_+} u_j \sigma\left(u_j \lambda_k \sigma'_{k,j} (1+\psi) + u_j \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} \phi + |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \right) \\
&\leq (1+\psi) \sum_{j\in J_+} u_j^2 \lambda_k \sigma'_{k,j} + \phi \sum_{i=1, i\neq k}^{m} \sum_{j\in J_+} u_j^2 \lambda_i \sigma'_{i,j} + \sum_{j\in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| .
\end{aligned}$$

Using Cauchy–Schwarz inequality we have

$$\sum_{j\in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| = \frac{1}{\sqrt{n}} \sum_{j\in J_+} |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \leq \frac{1}{\sqrt{n}} \left\|\mathbf{v}_\epsilon\right\| \sqrt{n} \max_{p\in[m]} \left\|x_p\right\| \leq \epsilon\sqrt{1+\psi} ,$$

getting

$$\sum_{j\in J_+} u_j \sigma\left(\mathbf{w}_j^\top \mathbf{x}_k\right) \leq (1+\psi)\alpha + \phi(m-1)\alpha_{\max} + \epsilon\sqrt{1+\psi} .$$

For the upper bound of the second summand we have

$$\begin{aligned}
\sum_{j\in J_-} u_j \mathbf{w}_j^\top \mathbf{x}_k &= \sum_{j\in J_-} u_j \left(u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i\neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle \right) \\
&\leq \sum_{j\in J_-} u_j \left(u_j \lambda_k \sigma'_{k,j} (1+\psi) + u_j \sum_{i=1, i\neq k}^{m} \lambda_i \sigma'_{i,j} \phi - |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \right) \\
&\leq (1+\psi)\alpha + \phi(m-1)\alpha_{\max} + \epsilon\sqrt{1+\psi} ,
\end{aligned}$$

and combining the two results we have

$$\begin{aligned}
1 \leq N(\boldsymbol{\theta}, \mathbf{x}_k) &\leq \sum_{j\in J_+} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) + \sum_{j\in J_-} u_j \mathbf{w}_j^\top \mathbf{x}_k \\
&\leq 2(1+\psi)\alpha + 2\phi(m-1)\alpha_{\max} + 2\epsilon\sqrt{1+\psi}
\end{aligned}$$

and solving for α we have

$$\alpha \geq \frac{1 - 2\phi(m-1)\alpha_{\max} - 2\epsilon\sqrt{1+\psi}}{2(1+\psi)} \ .$$

Case 2: $y_k = -1$
First we have

$$-1 \geq N(\boldsymbol{\theta}, \mathbf{x}_k) \geq \sum_{j \in J_+} u_j \mathbf{w}_j^\top \mathbf{x}_k + \sum_{j \in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k)$$

we get for the first summand

$$\begin{aligned}
\sum_{j \in J_+} u_j \left(\mathbf{w}_j^\top \mathbf{x}_k\right) &= \sum_{j \in J_+} u_j \left(-u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i \neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle \right) \\
&\geq -(1+\psi) \sum_{j \in J_+} u_j^2 \lambda_k \sigma'_{k,j} - \phi \sum_{i=1, i \neq k}^{m} \sum_{j \in J_+} u_j^2 \lambda_i \sigma'_{i,j} - \sum_{j \in J_+} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \\
&\geq -(1+\psi)\alpha - \phi(m-1)\alpha_{\max} - \epsilon\sqrt{1+\psi}
\end{aligned}$$

And for the second summand

$$\begin{aligned}
\sum_{j \in J_-} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_k) &= \sum_{j \in J_-} u_j \sigma \left(-u_j \lambda_k \sigma'_{k,j} \left\|\mathbf{x}_k\right\|^2 + u_j \sum_{i=1, i \neq k}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle \right) \\
&\geq -(1+\psi) \sum_{j \in J_+} u_j^2 \lambda_k \sigma'_{k,j} - \phi \sum_{i=1, i \neq k}^{m} \sum_{j \in J_+} u_j^2 \lambda_i \sigma'_{i,j} + \sum_{j \in J_-} u_j |\langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_k \rangle| \\
&\geq -(1+\psi)\alpha - \phi(m-1)\alpha_{\max} - \epsilon\sqrt{1+\psi}
\end{aligned}$$

combining the two results leads to the same lower bound

$$\alpha \geq \frac{1 - 2\phi(m-1)\alpha_{\max} - 2\epsilon\sqrt{1+\psi}}{2(1+\psi)} \ .$$

From C.3 we have that $\alpha_{\max} \leq 10.2$, and we plug in $\psi \leq 0.1$ and $\phi \leq \frac{\epsilon_d}{4mn}$, getting

$$\begin{aligned}
\alpha &\geq \frac{1 - 2\phi(m-1)\alpha_{\max} - 2\epsilon\sqrt{1+\psi}}{2(1+\psi)} \\
&\geq \frac{1 - 2\frac{\epsilon_d}{4mn}(m-1)10.2 - 2.1\epsilon}{2.2} \\
&\geq \frac{1 - 5.1\frac{\epsilon_d}{n} - 2.1\epsilon}{2.2} \geq 0.45 - 2.32\frac{\epsilon_d}{n} - 0.96\epsilon
\end{aligned}$$

meaning for all $i \in [m]$ we have

$$\min\left\{ \sum_{j \in J_+} u_j^2 \lambda_i \sigma'_{i,j}, \sum_{j \in J_-} u_j^2 \lambda_i \sigma'_{i,j} \right\} \geq 0.45 - 2.32\frac{\epsilon_d}{n} - 0.96\epsilon$$

so

$$\sum_{j=1}^{n} u_j^2 \lambda_i \sigma'_{i,j} \geq 0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon$$

using the fact that for all $j \in [n]$ and $k \in [m]$, $|u_j| = \frac{1}{\sqrt{n}}$ and $\sigma'_{k,j} \leq 1$ we also get that

$$\lambda_i \geq \frac{0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon}{\frac{1}{n}\sum_{j=1}^{n}\sigma'_{i,j}} \geq 0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon$$

□

Lemma C.5. *Let $N(\boldsymbol{\theta}, \mathbf{x}) = \sum_{j=1}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x})$ be a two-layer fully connected neural network, trained on $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, and let $0 < \epsilon_d, \epsilon, \delta \leq 1$ such that $\boldsymbol{\theta}$ is an (ϵ, δ)-approximate KKT point for the margin maximization problem (2) for S according to Definition 2.1 for $\lambda_1, ..., \lambda_m$, and S satisfies Assumption 2.3 for $\psi = 0, 1$, and $\phi \leq \frac{\epsilon_d}{4mn}$. Given $l \in [m]$, we denote by $\hat{\boldsymbol{\theta}}$ the parameters created by performing gradient ascent on the first layer weights, for the data sample $(\mathbf{x}_l, y_l) \in S$ with step size determined by λ_l (3). We denote by $\widetilde{\boldsymbol{\theta}}$ the weight vector such that for $j \in [n]$*

$$\widetilde{\mathbf{w}}_j = \hat{\mathbf{w}}_j + |u_j|\lambda_l \sigma'_{l,j} \Delta_j \ ,$$

for $\Delta_j = \sum_{k \in [m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j \rangle)$ and $c = \frac{\epsilon_d}{2mn}$. Then, for all $r \in [m]_{-l}$ and $j \in [n]$,

$$\operatorname{sign}(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) = \operatorname{sign}(\mathbf{w}_j^\top \mathbf{x}_r)$$

Proof: Let $r \in [m]_{-l}$, and $j \in [n]$. Looking at the inner product, we denote

$$\Delta_{r,j} = \langle \Delta_j, \mathbf{x}_r \rangle = \sum_{k \in [m]_{-l}} c\langle \mathbf{x}_k, \mathbf{x}_r \rangle \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j \rangle) \ ,$$

and have

$$\begin{aligned} \mathbf{w}_j^\top \mathbf{x}_r &= u_j \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_r \rangle = \\ &= u_j \sum_{i \in [m]_{-l}} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_r \rangle + u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_r \rangle \end{aligned}$$

and

$$\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r = u_j \sum_{i \in [m]_{-l}} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_r \rangle + |u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} + \langle \mathbf{v}_{\epsilon,j}, \mathbf{x}_r \rangle \ ,$$

where one can see that the difference between the inner products is

$$\mathbf{w}_j^\top \mathbf{x}_r - \widetilde{\mathbf{w}}_j^\top \mathbf{x}_r = u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} = \lambda_l \sigma'_{l,j} \left(u_j \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \Delta_{j,r} \right) \ .$$

To show they have the same sign, it's enough to show that the difference is either negative to positive, depending on $\mathbf{w}_j^\top \mathbf{x}_r$ sign. If it is positive, we show the difference in negative, hence $\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r$ is bigger and also positive, and if it's negative we show the a positive difference to conclude equal sign.

Note, if $\lambda_l = 0$ we are done, and particularly we have not change $\boldsymbol{\theta}$ by unlearning or adding our fix, meaning $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$. In addition, if $\sigma'_{l,j} = 0$ for some j, we haven't change the neuron $\mathbf{w}_j$, and the claim follows. For the rest of the proof we assume $\lambda_l > 0$ and $\sigma'_{l,j} = 1$, so to show the difference's sign it's enough to show the sign of $(u_j \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \Delta_{j,r})$.

Case 1: $\mathbf{w}_j^\top \mathbf{x}_r \geq 0$. We show that $(u_j \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \Delta_{j,r}) \leq 0$

By Lemma C.1

$$|u_j| \Delta_{j,r} \geq |u_j| \left(c(1 - \psi) - (m - 2)c\phi \right)$$

And using Assumption 2.3 we get that $|\langle \mathbf{x}_l, \mathbf{x}_r \rangle| \leq \phi$ we have

$$
\begin{aligned}
u_j\langle \mathbf{x}_l, \mathbf{x}_r\rangle - |u_j|\Delta_{j,r} &\le |u_j|\phi - |u_j|\left(c(1-\psi) - (m-2)c\phi\right) \\
&\le |u_j|\left(\phi - (c(1-\psi) - (m-2)c\phi)\right) \ .
\end{aligned}
$$

We left to show that $(\phi - c(1-\psi) + (m-2)c\phi) \le 0$ and indeed plugging in $\psi = 0.1$, $\phi \le \frac{\epsilon_d}{4mn}$, $c = \frac{\epsilon_d}{2mn}$ we have

$$
\phi - c(1-\psi) + (m-2)c\phi \le \frac{\epsilon_d}{4mn} - \frac{\epsilon_d}{2mn}(0.9) + (m-2)\frac{\epsilon_d}{2mn}\frac{\epsilon_d}{4mn} \le \frac{0.25\epsilon_d - 0.45\epsilon_d + 0.125\epsilon_d^2}{mn} < 0
$$

which finishes this case.

Case 2: $\mathbf{w}_j^T\mathbf{x}_r < 0$. We show that $(u_j\langle \mathbf{x}_l, \mathbf{x}_r\rangle - |u_j|\Delta_{j,r}) \ge 0$

By Lemma C.1

$$
|u_j|\Delta_{j,r} \le |u_j|\left(-c(1-\psi) + (m-2)c\phi\right)
$$

And using Assumption 2.3 we get that $|\langle \mathbf{x}_l, \mathbf{x}_r\rangle| \le \phi$ we have

$$
\begin{aligned}
u_j\langle \mathbf{x}_l, \mathbf{x}_r\rangle - |u_j|\Delta_{j,r} &\ge -|u_j|\phi - |u_j|\left(-c(1-\psi) + (m-2)c\phi\right) \\
&\ge |u_j|\left(-\phi + (c(1-\psi) - (m-2)c\phi)\right) \ .
\end{aligned}
$$

Now, It's enough to show that $-\phi + c(1-\psi) + (m-2)c\phi \ge 0$, which has already proven in the previous case.

□

Lemma C.6. *Let $0 < \epsilon_d, \epsilon, \delta \le 0.4$. Let $N(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{n} u_j\sigma(\mathbf{w}_j^\top \mathbf{x})$ be a two-layer fully connected neural network, trained on $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, and assume that $\boldsymbol{\theta}$ is an (ϵ, δ)-approximate KKT point for the margin maximization problem (2) for S according to Definition 2.1 for $\lambda_1, ..., \lambda_m$, and S satisfies Assumption 2.3 for $\psi \le 0.1$ and $\phi \le \frac{\epsilon_d}{4mn}$. Given $l \in [m]$, we denote by $\hat{\boldsymbol{\theta}}$ the parameters created by performing gradient ascent on the first layer weights, for the data sample $(\mathbf{x}_l, y_l) \in S$ with step size λ_l (3). We denote by $\widetilde{\boldsymbol{\theta}}$ the weight vector such that for $j \in [n]$*

$$
\widetilde{\mathbf{w}}_j = \hat{\mathbf{w}}_j + |u_j|\lambda_l\sigma'_{l,j}\Delta_j \ ,
$$

for $\Delta_j = \sum_{k\in[m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle)$ and $c = \frac{\epsilon_d}{2mn}$. Then, for all $r \in [m]_{-l}$,

$$
-\frac{9\epsilon_d}{mn} \le y_r\left[N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r)\right] \le \frac{9\epsilon_d}{mn} \ ,
$$

Proof: Let $r \in [m]_{-l}$. We look at the margins for $\mathbf{x}_r$ with respect to $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ and get the difference

$$
y_r\left[N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r)\right] = y_r\left[\sum_{j=1}^{n} u_j\sigma(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) - \sum_{j=1}^{n} u_j\sigma(\mathbf{w}_j^\top \mathbf{x}_r)\right] \ .
$$

From Lemma C.5 we get that for $j \in [n]$, $\operatorname{sign}(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) = \operatorname{sign}(\mathbf{w}_j^\top \mathbf{x}_r)$. Then, if $\mathbf{w}_j^\top \mathbf{x}_r < 0$ we get that $\sigma(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) = \sigma(\mathbf{w}_j^\top \mathbf{x}_r) = 0$. Otherwise, $\mathbf{w}_j^\top \mathbf{x}_r \ge 0$, and we get that $\sigma(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) = \widetilde{\mathbf{w}}_j^\top \mathbf{x}_r$ and $\sigma(\mathbf{w}_j^\top \mathbf{x}_r) = \mathbf{w}_j^\top \mathbf{x}_r$. We denote $J_+ = \{j \in [n] : \mathbf{w}_j^\top \mathbf{x}_r > 0 \text{ and } u_j > 0\}$ and $J_- = \{j \in [n] : \mathbf{w}_j^\top \mathbf{x}_r > 0 \text{ and } u_j < 0\}$, and get

$$
\begin{aligned}
&\sum_{j=1}^{n} u_j\sigma(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) - \sum_{j=1}^{n} u_j\sigma(\mathbf{w}_j^\top \mathbf{x}_r) = \\
&= \sum_{j=1}^{n} u_j\left(\sigma(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) - \sigma(\mathbf{w}_j^\top \mathbf{x}_r)\right) \\
&= \sum_{j\in J_+} u_j\left(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r - \mathbf{w}_j^\top \mathbf{x}_r\right) - \sum_{j\in J_-} |u_j|\left(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r - \mathbf{w}_j^\top \mathbf{x}_r\right)
\end{aligned}
$$

Following Definition 2.1, we denote $\mathbf{v}_\epsilon = \boldsymbol{\theta} - \sum_{i=1}^m \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i)$ and for $j \in [n]$ we denote,

$$\mathbf{w}_j = \sum_{i=1}^m \lambda_i y_i \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_{\epsilon,j} = u_j \sum_{i=1}^m \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i + \mathbf{v}_{\epsilon,j} \ ,$$

such that $\mathbf{v}_\epsilon = (\mathbf{v}_{\epsilon,1}, ..., \mathbf{v}_{\epsilon,n})$ a concatenation of all $\mathbf{v}_{\epsilon,j}$'s vectors. Following the unlearning step in 3 for $(\mathbf{x}_l, y_l)$, we denote

$$\hat{\mathbf{w}}_j = \sum_{i \in [m]_{-l}} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i + \mathbf{v}_{\epsilon,j} \ ,$$

and get

$$\widetilde{\mathbf{w}}_j = \sum_{i \in [m]_{-l}} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i + |u_j| \lambda_l \sigma'_{l,j} \Delta_j + \mathbf{v}_{\epsilon,j} \ .$$

When we look at the difference $\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r - \mathbf{w}_j^\top \mathbf{x}_r$, we get that for $j \in J_+ \cup J_-$

$$\begin{aligned}
&0 \leq \widetilde{\mathbf{w}}_j^\top \mathbf{x}_r - \mathbf{w}_j^\top \mathbf{x}_r = \\
&= \left[u_j \sum_{i \in [m]_{-l}} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_r \rangle + |u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} + \mathbf{v}_{\epsilon,j} \right] - \left[u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle + u_j \sum_{i \in [m]_{-l}} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_r \rangle + \mathbf{v}_{\epsilon,j} \right] \\
&= |u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} - u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle \ .
\end{aligned}$$

We now use this equality for the margin difference, getting

$$\begin{aligned}
&N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r) = \\
&= \sum_{j \in J_+} u_j \left(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r - \mathbf{w}_j^\top \mathbf{x}_r \right) - \sum_{j \in J_-} |u_j| \left(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r - \mathbf{w}_j^\top \mathbf{x}_r \right) \\
&= \sum_{j \in J_+} u_j \left(|u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} - u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) - \sum_{j \in J_-} |u_j| \left(|u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} - u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) \\
&= \sum_{j \in J_+} u_j^2 \lambda_l \sigma'_{l,j} \left(\Delta_{j,r} - y_l \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) - \sum_{j \in J_-} u_j^2 \lambda_l \sigma'_{l,j} \left(\Delta_{j,r} + y_l \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) \ .
\end{aligned}$$

We denote $\alpha_- = \sum_{j \in J_-} u_j^2 \lambda_l \sigma'_{l,j}$ and by $\alpha_+ = \sum_{j \in J_+} u_j^2 \lambda_l \sigma'_{l,j}$. So, we get that

$$\begin{aligned}
N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r) =& \alpha_+ \left(\Delta_{j,r} - y_l \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) - \alpha_- \left(\Delta_{j,r} + y_l \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) \\
=& \alpha_+ \Delta_{j,r} - y_l \alpha_+ \langle \mathbf{x}_l, \mathbf{x}_r \rangle - \alpha_- \Delta_{j,r} - y_l \alpha_- \langle \mathbf{x}_l, \mathbf{x}_r \rangle \\
=& \alpha_+ \Delta_{j,r} - \alpha_- \Delta_{j,r} - y_l \left(\alpha_+ \langle \mathbf{x}_l, \mathbf{x}_r \rangle + \alpha_- \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) \ .
\end{aligned}$$

Since $\alpha_-, \alpha_+, \Delta_{j,r} \geq 0$, for the upper bounds we get

$$\begin{aligned}
N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r) =& \alpha_+ \Delta_{j,r} - \alpha_- \Delta_{j,r} - y_l \left(\alpha_+ \langle \mathbf{x}_l, \mathbf{x}_r \rangle + \alpha_- \langle \mathbf{x}_l, \mathbf{x}_r \rangle \right) \\
\leq& \alpha_+ \Delta_{j,r} + \alpha_+ \phi + \alpha_- \phi \ .
\end{aligned}$$

From Lemma C.3 we get that $\alpha_-, \alpha_+ \leq 10.2$, from Lemma C.1 we get that $\Delta_{j,r} \leq c(1+\psi) + (m-2)c\phi$. Together with plugging in $\psi = 0.1$, $\phi \leq \frac{\epsilon_d}{4mn}$, $c = \frac{\epsilon_d}{2mn}$ and $\epsilon_d \leq 1$, we get

$$
\begin{aligned}
N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r)\alpha_+\Delta_{j,r} + \alpha_+\phi + \alpha_-\phi \\
&\leq \alpha_+ \left(c(1+\psi) + (m-2)c\phi + \phi\right) + \alpha_-\phi \\
&\leq 10.2\left(\frac{1.1\epsilon_d}{2mn} + (m-2)\frac{\epsilon_d}{2mn}\frac{\epsilon_d}{4mn} + \frac{\epsilon_d}{4mn}\right) + 10.2\frac{\epsilon_d}{4mn} \\
&\leq 10.2\left(\frac{1.1\epsilon_d}{2mn} + \frac{\epsilon_d}{2n}\frac{\epsilon_d}{4mn} + \frac{\epsilon_d}{4mn}\right) + \frac{2.55\epsilon_d}{mn} \\
&\leq \frac{\epsilon_d}{mn}\left[5.61 + \frac{0.125\epsilon_d}{n} + 0.25 + 2.55\right] \leq \frac{9\epsilon_d}{mn} \,.
\end{aligned}
$$

For the lower bound of the margin we get

$$
\begin{aligned}
N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r) =& \alpha_+\Delta_{j,r} - \alpha_-\Delta_{j,r} - y_l\left(\alpha_+\langle \mathbf{x}_l, \mathbf{x}_r\rangle + \alpha_-\langle \mathbf{x}_l, \mathbf{x}_r\rangle\right) \\
\geq& -\alpha_-\Delta_{j,r} - \alpha_+\phi - \alpha_-\phi \,,
\end{aligned}
$$

and the same calculations we did for the upper bound will yield

$$
N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r) \geq -\frac{9\epsilon_d}{mn} \,.
$$

□

C.2 Proof for Theorem 4.1

Proof: Note, for readability of the proof we denote ϵ_1 by ϵ and δ_1 by δ.

Using the stationarity condition in Definition 2.1 for $\boldsymbol{\theta}$, we denote $\mathbf{v}_\epsilon = \boldsymbol{\theta} - \sum_{i=1}^m \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i)$ and for $j \in [n]$ we denote,

$$
\mathbf{w}_j = \sum_{i=1}^m \lambda_i y_i \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_{\epsilon,j} = u_j \sum_{i=1}^m \lambda_i y_i \sigma'_{i,j}\mathbf{x}_i + \mathbf{v}_{\epsilon,j}
$$

where $\mathbf{v}_\epsilon = (\mathbf{v}_{\epsilon,1}, ..., \mathbf{v}_{\epsilon,n})$ the concatenation of all $\mathbf{v}_{\epsilon,j}$ and $\|\mathbf{v}_\epsilon\| = \epsilon$.

Let $l \in [m]$, we wish to take a negative gradient step of size β, such that

$$
\beta \nabla_{\boldsymbol{\theta}} \ell(y_l N(\boldsymbol{\theta}, \mathbf{x}_l)) = -\lambda_l y_l \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_l)
$$

so we pick a step size $\beta = \frac{-\lambda_l}{\ell'(y_l N(\boldsymbol{\theta}, \mathbf{x}_l))}$. We denote by $\hat{\boldsymbol{\theta}}$ the parameters created by performing gradient ascent on the first layer weights, for the data sample $(\mathbf{x}_l, y_l) \in S$ with step size β (3). As a result, for all $j \in [n]$ we have

$$
\begin{aligned}
\hat{\mathbf{w}}_j =& \mathbf{w}_j - \lambda_l y_l \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_l) \\
=& \sum_{i=1}^m \lambda_i y_i \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_{\epsilon,j} - \lambda_l y_l \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_l) = \sum_{i \in [m]_{-l}} u_j \lambda_i y_i \sigma'_{i,j}\mathbf{x}_i + \mathbf{v}_{\epsilon,j} \,.
\end{aligned}
$$

Given $\hat{\boldsymbol{\theta}}$ and the unlearned sample index $l \in [m]$, we denote $c = \frac{\epsilon_d}{2mn}$, and for $j \in [n]$, we denote:

$$
\Delta_j := \sum_{k \in [m]_{-l}} c\mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j\rangle) \,.
$$

Using Δ_j, we define a slightly modified weight vector $\widetilde{\boldsymbol{\theta}}$, such that for $j \in [n]$,

$$
\widetilde{\mathbf{w}}_j = \hat{\mathbf{w}}_j + |u_j|\lambda_l \sigma'_{l,j}\Delta_j \,.
$$

C.2.1 Proof of $\widetilde{\theta}$ has the direction of a $(\epsilon + \frac{9\epsilon_d\epsilon}{m-9\epsilon_d} + \frac{23\epsilon_d}{\sqrt{m}}, \delta + \frac{9\epsilon_d\delta}{m-9\epsilon_d} + \frac{22.6\epsilon_d}{m})$-approximate KKT point of the margin maximization problem (2) w.r.t. $S \setminus \{\mathbf{x}_l, y_l\}$

It is enough to prove that $\widetilde{\boldsymbol{\theta}}$ is an $(\epsilon + \frac{22\epsilon_d}{\sqrt{m}}, \delta + \frac{184\epsilon_d}{m}, \frac{9\epsilon_d}{mn})$-approximate KKT for the margin maximization problem (2) w.r.t. $S \setminus (\mathbf{x}_l, y_l)$ with the corresponding $\{\lambda_i\}_{i\in[m]_{-l}}$, according to Definition B.1. Then, using Lemma B.3, we conclude the approximation parameters for $\frac{1}{1-\frac{9\epsilon_d}{mn}}\widetilde{\boldsymbol{\theta}}$, for the stationarity parameter, for $\epsilon_d \leq 0.01$ we have

$$\frac{1}{1-\frac{9\epsilon_d}{mn}}\left(\epsilon + \frac{22\epsilon_d}{\sqrt{m}}\right) \leq \left(1 + \frac{9\epsilon_d}{m-9\epsilon_d}\right)\left(\epsilon + \frac{22\epsilon_d}{\sqrt{m}}\right) \leq \epsilon + \frac{9\epsilon_d\epsilon}{m-9\epsilon_d} + \frac{23\epsilon_d}{\sqrt{m}} \ ,$$

For the complementarity slackness parameter we use the upper bound for $\max_p \lambda_p$ from C.3, and have

$$\frac{1}{1-\frac{9\epsilon_d}{mn}}\left(\delta + \frac{184\epsilon_d}{m}\right) + \max_p \lambda_p \frac{\frac{9\epsilon_d}{mn}}{1-\frac{9\epsilon_d}{mn}} \leq \delta + \frac{9\epsilon_d\delta}{m-9\epsilon_d} + \frac{22.6\epsilon_d}{m} \ ,$$

Finally we conclude that $\frac{1}{1-\frac{9\epsilon_d}{mn}}\widetilde{\boldsymbol{\theta}}$ is a $(\epsilon + \frac{9\epsilon_d\epsilon}{m-9\epsilon_d} + \frac{23\epsilon_d}{\sqrt{m}}, \delta + \frac{9\epsilon_d\delta}{m-9\epsilon_d} + \frac{22.6\epsilon_d}{m})$-approximate KKT for the margin maximization problem (2) w.r.t. $S \setminus \{\mathbf{x}_l, y_l\}$, according to Definition 2.1. We note that $\widetilde{\boldsymbol{\theta}}$ and $\frac{1}{1-\hat{\gamma}}\widetilde{\boldsymbol{\theta}}$ has the same direction, which finishes the proof.

We start by showing $\widetilde{\boldsymbol{\theta}}$ is an $(\epsilon + \frac{22\epsilon_d}{\sqrt{m}}, \delta + \frac{184\epsilon_d}{m}, \frac{9\epsilon_d}{mn})$-approximate KKT.

(1) Dual Feasibility: For all $r \in [m]_{-l}$, $\lambda_r \geq 0$.

Directly from dual feasibility for $\boldsymbol{\theta}$ (Definition 2.1).

(2) Stationarity: $\left\|\widetilde{\boldsymbol{\theta}} - \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i)\right\| \leq \epsilon + \frac{22\epsilon_d}{\sqrt{m}}$.

From stationarity for $\boldsymbol{\theta}$ (Definition 2.1) we get that $\boldsymbol{\theta} = \sum_{i=1}^m \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_\epsilon$. By the difinition of $\hat{\boldsymbol{\theta}}$ we get that $\hat{\boldsymbol{\theta}} = \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_\epsilon$. For readability, we first denote $\mathbf{u} = (|u_1|\lambda_l\sigma'_{l,1}\Delta_1, .., |u_n|\lambda_l\sigma'_{l,n}\Delta_n)$, such that $\mathbf{u} \in \mathbb{R}^{m\times n}$, and note that one can write $\widetilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \mathbf{u}$. Thus,

$$\left\|\widetilde{\boldsymbol{\theta}} - \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i)\right\| = \left\|\sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_\epsilon + \mathbf{u} - \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i)\right\|$$

In Lemma C.5 we showed that for $j \in [n], i \in [m]$, $\mathbb{1}_{\{\widetilde{\mathbf{w}}_j^T\mathbf{x}_j \geq 0\}} = \mathbb{1}_{\{\mathbf{w}_j^T\mathbf{x}_j \geq 0\}}$. Then, for $j \in [n]$ we have

$$\nabla_{\mathbf{w}_j} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i) = u_j \mathbb{1}_{\{\widetilde{\mathbf{w}}_j^T\mathbf{x}_j \geq 0\}}\mathbf{x}_i = u_j \mathbb{1}_{\{\mathbf{w}_j^T\mathbf{x}_j \geq 0\}}\mathbf{x}_i = \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_i) \ ,$$

which leads to

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{\theta}} - \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i)\right\| = \\
&= \left\|\sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_\epsilon + \mathbf{u} - \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i)\right\| \\
&= \|\mathbf{v}_\epsilon + \mathbf{u}\| \leq \|\mathbf{v}_\epsilon\| + \|\mathbf{u}\| \ .
\end{aligned}$$

Using the upper bound from Lemma C.2, for $\|\mathbf{u}\|$ we have

$$\|\mathbf{u}\| = \left\|(|u_1|\lambda_l\sigma'_{l,1}\Delta_1, .., |u_n|\lambda_l\sigma'_{l,n}\Delta_n)\right\| = \sqrt{\sum_{j=1}^{n}\left\|u_j\lambda_l\sigma'_{l,j}\Delta_j\right\|^2} \leq$$
$$\leq \sqrt{n}\max_{j\in[n]}|u_j|\lambda_l\sigma'_{l,j}\,\|\Delta_j\|$$
$$\leq \sqrt{n}\frac{22\epsilon_d}{\sqrt{mn}} \leq \frac{22\epsilon_d}{\sqrt{m}}\,,$$

and plugging it in we have

$$\left\|\widetilde{\boldsymbol{\theta}} - \sum_{i\in[m]_{-l}}\lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i)\right\| \leq \|\mathbf{v}_\epsilon\| + \|\mathbf{u}\| \leq \epsilon + \frac{22\epsilon_d}{\sqrt{m}}\,.$$

as desired.

From Lemma C.6 we get that $-\frac{9\epsilon_d}{mn} \leq y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - y_r N(\boldsymbol{\theta}, \mathbf{x}_r) \leq \frac{9\epsilon_d}{mn}$. Using it we prove the next conditions.

(3) Complementarity Slackness: For all $r \in [m]_{-l}$, $\lambda_r\left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - 1\right) \leq \delta + \frac{184\epsilon_d}{m}$.

Let $r \in [m]_{-l}$. If $\lambda_r = 0$ we are done. Otherwise, from complementarity slackness condition for $\boldsymbol{\theta}$ we get that $\lambda_r\left(y_r N(\boldsymbol{\theta}, \mathbf{x}_r) - 1\right) \leq \delta$. We use the fact that $y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - \frac{9\epsilon_d}{mn} \leq y_r N(\boldsymbol{\theta}, \mathbf{x}_r)$ to get that

$$\delta \geq \lambda_r\left(y_r N(\boldsymbol{\theta}, \mathbf{x}_r) - 1\right) = \lambda_r\left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - \frac{9\epsilon_d}{mn} - 1\right) = \lambda_r\left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - 1\right) - \lambda_r\frac{9\epsilon_d}{mn}$$
$$\geq \lambda_r\left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - 1\right) - \max_p \lambda_p\frac{9\epsilon_d}{mn}$$

and conclude that

$$\lambda_r\left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - 1\right) \leq \delta + \max_p \lambda_p\frac{9\epsilon_d}{mn}\,.$$

From Lemma C.3 we have an upper bound $\max_p \lambda_p \leq 20.4n$, so we get that

$$\lambda_r\left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - 1\right) \leq \delta + \max_p \lambda_p\frac{9\epsilon_d}{mn} \leq \delta + 20.4n\frac{9\epsilon_d}{nm} \leq \delta + \frac{184\epsilon_d}{m}\,.$$

(4) Primal Feasibility: For all $r \in [m]_{-l}$, $y_i N(\mathbf{x}_i, \widetilde{\boldsymbol{\theta}}) \geq 1 - \frac{9\epsilon_d}{mn}$.

Let $r \in [m]_{-l}$. From primal feasibility for $\boldsymbol{\theta}$ (Definition 2.1) we get that $y_r N(\boldsymbol{\theta}, \mathbf{x}_r) \geq 1$, and from Lemma C.6 we have that

$$y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - y_r N(\boldsymbol{\theta}, \mathbf{x}_r) \geq -\frac{9\epsilon_d}{mn}$$

which concludes the proof.

C.2.2 Proof of $\text{cossim}(\hat{\theta}, \widetilde{\theta}) \geq 1 - \frac{82\epsilon_d}{m}$

We begin with looking at the inner product $\langle\hat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}\rangle$. For readability, we first denote $\mathbf{u} = (|u_1|\lambda_l\sigma'_{l,1}\Delta_1, .., |u_n|\lambda_l\sigma'_{l,n}\Delta_n)$, such that $\mathbf{u} \in \mathbb{R}^{m\times n}$, and note that one can write $\widetilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \mathbf{u}$ and

$$\langle\hat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}\rangle = \langle\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}} + \mathbf{u}\rangle = \left\|\hat{\boldsymbol{\theta}}\right\|^2 + \langle\hat{\boldsymbol{\theta}}, \mathbf{u}\rangle \geq \left\|\hat{\boldsymbol{\theta}}\right\|^2 - |\langle\hat{\boldsymbol{\theta}}, \mathbf{u}\rangle| \geq \left\|\hat{\boldsymbol{\theta}}\right\|^2 - \left\|\hat{\boldsymbol{\theta}}\right\|\|\mathbf{u}\|\,,$$

where the last transition is due to Cauchy–Schwarz inequality. We now look at the weights vectors norm and get

$$\left\|\widetilde{\boldsymbol{\theta}}\right\| = \left\|\hat{\boldsymbol{\theta}} + \mathbf{u}\right\| \leq \left\|\hat{\boldsymbol{\theta}}\right\| + \|\mathbf{u}\|$$

which leads to

$$\left\|\hat{\boldsymbol{\theta}}\right\| \left\|\widetilde{\boldsymbol{\theta}}\right\| = \left\|\hat{\boldsymbol{\theta}}\right\| \left(\left\|\hat{\boldsymbol{\theta}}\right\| + \|\mathbf{u}\|\right) = \left\|\hat{\boldsymbol{\theta}}\right\|^2 + \left\|\hat{\boldsymbol{\theta}}\right\| \|\mathbf{u}\|$$

We are now ready to lower bound the cosine similarity, having

$$\begin{aligned}
\operatorname{cossim}(\hat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}) =& \frac{\langle \hat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}} \rangle}{\left\|\hat{\boldsymbol{\theta}}\right\| \left\|\widetilde{\boldsymbol{\theta}}\right\|} \\
\geq& \frac{\left\|\hat{\boldsymbol{\theta}}\right\|^2 - \left\|\hat{\boldsymbol{\theta}}\right\| \|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|^2 + \left\|\hat{\boldsymbol{\theta}}\right\| \|\mathbf{u}\|} \\
\geq& 1 - \frac{2\left\|\hat{\boldsymbol{\theta}}\right\| \|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|^2 + \left\|\hat{\boldsymbol{\theta}}\right\| \|\mathbf{u}\|} \\
\geq& 1 - \frac{2\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|} \ .
\end{aligned}$$

To finish the proof, we upper bound $\frac{\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|}$. We note that we can upper bound the norm of $\mathbf{u}$ using the upper bound from Lemma C.2:

$$\begin{aligned}
\|\mathbf{u}\| = \left\|(|u_1|\lambda_l\sigma'_{l,1}\Delta_1, .., |u_n|\lambda_l\sigma'_{l,n}\Delta_n)\right\| =& \sqrt{\sum_{j=1}^{n} \left\|u_j\lambda_l\sigma'_{l,j}\Delta_j\right\|^2} \leq \\
\leq& \sqrt{n} \max_{j\in[n]} |u_j|\lambda_l\sigma'_{l,j} \|\Delta_j\| \\
\leq& \sqrt{n}\frac{22\epsilon_d}{\sqrt{mn}} \leq \frac{22\epsilon_d}{\sqrt{m}} \ ,
\end{aligned}$$

We now show a lower bound for $\left\|\hat{\boldsymbol{\theta}}\right\|$, using that for all $j \in [n]$, $|u_j| = \frac{1}{\sqrt{n}}$, and for Assumption 2.3, for all $i, k \in [m]$ $\|\mathbf{x}_i\|^2 \geq (1-\psi)$, and $|\langle \mathbf{x}_i, \mathbf{x}_k \rangle| \leq \phi$. We have

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}\right\|^2 &= \sum_{j\in[n]} \|\hat{\mathbf{w}}_j\|^2 \\
&= \sum_{j\in[n]} \left\| \sum_{i\in[m]_{-l}} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i \right\|^2 \\
&= \sum_{j\in[n]} \langle \sum_{i\in[m]_{-l}} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i, \sum_{i\in[m]_{-l}} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i \rangle \\
&\geq \sum_{j\in[n]} \left(\sum_{i\in[m]_{-l}} u_j^2 \lambda_i^2 \sigma'_{i,j} \|\mathbf{x}_i\|^2 - \sum_{i\in[m]_{-l}} \sum_{k\neq i\in[m]_{-l}} u_j^2 \lambda_i \lambda_k \sigma'_{i,j} \sigma'_{k,j} \langle \mathbf{x}_i, \mathbf{x}_k \rangle \right) \\
&\geq \frac{1}{n} \sum_{j\in[n]} \left((1-\psi) \sum_{i\in[m]_{-l}} \lambda_i^2 \sigma'_{i,j} - \phi \sum_{i\in[m]_{-l}} \sum_{k\neq i\in[m]_{-l}} \lambda_i \lambda_k \sigma'_{i,j} \sigma'_{k,j} \right) \\
&\geq \frac{1}{n} \left((1-\psi) \sum_{i\in[m]_{-l}} \lambda_i^2 \sum_{j\in[n]} \sigma'_{i,j} - \phi \sum_{i\in[m]_{-l}} \sum_{k\neq i\in[m]_{-l}} \lambda_i \lambda_k \sum_{j\in[n]} \sigma'_{i,j} \right)
\end{aligned}
$$

We note that using Lemma C.4 and Lemma C.3, for all i, we have

$$
\left(0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon\right) \leq \sum_{j=1}^{n} u_j^2 \lambda_i \sigma'_{i,j} \leq 20.4
$$

hence since $|u_j| = \frac{1}{\sqrt{n}}$

$$
\left(0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon\right) n \leq \lambda_i \sum_{j=1}^{n} \sigma'_{i,j} \leq 20.4n
$$

Using these bounds we have

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}\right\|^2 &\geq \frac{1}{n} \left((1-\psi) \sum_{i\in[m]_{-l}} \lambda_i \left(0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon\right) n - \phi \sum_{i\in[m]_{-l}} \sum_{k\neq i\in[m]_{-l}} \lambda_i 20.4n \right) \\
&\geq \left((1-\psi) \left(0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon\right) \sum_{i\in[m]_{-l}} \lambda_i - 20.4\phi \sum_{i\in[m]_{-l}} \sum_{k\neq i\in[m]_{-l}} \lambda_i \right) \\
&\geq \sum_{i\in[m]_{-l}} \lambda_i \left[(1-\psi) \left(0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon\right) - 20.4\phi(m-2) \right]
\end{aligned}
$$

Plugging in $\psi \leq 0.1$, $\phi \leq \frac{\epsilon_d}{4mn}$, $\epsilon < 1$ and $\epsilon_d \leq 0.01$ we have

$$
\begin{aligned}
&\geq \sum_{i\in[m]_{-l}} \lambda_i \left[0.9\left(0.9-4.64\frac{\epsilon_d}{n}-1.92\epsilon\right)-20.4\phi(m-2)\right] \\
&\geq (m-1)\left(0.9-4.64\frac{\epsilon_d}{n}-1.92\epsilon\right)\left[0.9\left(0.9-4.64\frac{\epsilon_d}{n}-1.92\epsilon\right)-\frac{20.4\epsilon_d}{4n}\right] \\
&\geq (m-1)\left(0.9(0.9-4.64\frac{\epsilon_d}{n}-1.92\epsilon)^2-\frac{5.1\epsilon_d}{n}\left(0.9-4.64\frac{\epsilon_d}{n}-1.92\epsilon\right)\right) \\
\geq&(m-1)\left(0.72+19\frac{\epsilon_d^2}{n^2}+3\epsilon+8\frac{\epsilon_d\epsilon}{n}-7.6\frac{\epsilon_d}{n}-3.2\epsilon-\frac{4.6\epsilon_d}{n}\right) \\
\geq&(m-1)\left(0.72-12.2\frac{\epsilon_d}{n}-0.2\epsilon\right) \\
\geq&0.3(m-1)
\end{aligned}
$$

and of course

$$
\left\|\hat{\boldsymbol{\theta}}\right\| \geq \sqrt{0.3(m-1)}\,.
$$

We can know join the upper bound for $\|\mathbf{u}\|$ and lower bound of $\left\|\hat{\boldsymbol{\theta}}\right\|$ getting

$$
\begin{aligned}
\frac{\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|} \leq& \frac{\frac{22\epsilon_d}{\sqrt{m}}}{\sqrt{0.3(m-1)}} \\
\leq& \frac{41\epsilon_d}{m}
\end{aligned}
$$

and finally,

$$
\begin{aligned}
\operatorname{cossim}(\hat{\boldsymbol{\theta}},\widetilde{\boldsymbol{\theta}}) \geq& 1-\frac{2\,\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|} \\
\geq& 1-\frac{82\epsilon_d}{m}\,,
\end{aligned}
$$

as desired. □

C.3 Proof for forgetting subset of points using $\mathcal{A}_{\text{K-GA}}$ – two layer networks

We formalize and prove the statement for unlearning a subset of data points. Recall that the term *successful unlearning* here is the natural extension of Definition 2.2 to unlearning a subset, rather than a single point.

Theorem C.1. *In the same settings as Theorem 4.1, let $S_{forget} \subseteq S$ a subset of size k. Then, the extended algorithm $\mathcal{A}_{k\text{-}GA}$, with appropriate coefficients $\{\beta_r\}$, is a (ϵ,δ,τ)-**successful** unlearning algorithm w.r.t. $\boldsymbol{\theta}$ and S, where $\epsilon = \epsilon_1 + \frac{9\epsilon_d\epsilon_1}{\frac{m}{k}-9\epsilon_d} + \frac{23k\epsilon_d}{\sqrt{m}}$, $\delta = \delta_1 + \frac{9\epsilon_d\delta_1}{\frac{m}{k}-9\epsilon_d} + \frac{22.6k\epsilon_d}{m}$ and $\tau = \frac{82k\epsilon_d}{m-k}$.*

Proof: Let a forget set $S_f \subset S$ such that $|S_f| = k$. We denote $I_f = \{i : (\mathbf{x}_i, y_i) \in S_f\}$. We denote $S_r = S \setminus S_f$ and $I_r = \{i : (\mathbf{x}_i, y_i) \in S_r\}$. This proof widely relies the proof in C.2.

Using the stationarity condition in Definition 2.1 for $\boldsymbol{\theta}$, we denote $\mathbf{v}_\epsilon = \boldsymbol{\theta} - \sum_{i=1}^m \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i)$ and for $j \in [n]$ we denote,

$$
\mathbf{w}_j = \sum_{i=1}^m \lambda_i y_i \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_{\epsilon,j} = u_j \sum_{i=1}^m \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i + \mathbf{v}_{\epsilon,j}
$$

where $\mathbf{v}_\epsilon = (\mathbf{v}_{\epsilon,1}, ..., \mathbf{v}_{\epsilon,n})$ the concatenation of all $\mathbf{v}_{\epsilon,j}$ and $\|\mathbf{v}_\epsilon\| = \epsilon$.

According to the algorithm $\mathcal{A}_{\text{k-GA}}$, we take a step consists of the sum of k gradients w.r.t. data points in S_f with the following sizes- for any $(\mathbf{x}_l, y_l) \in S_f$, we take a step size $\beta = \frac{-\lambda_l}{\ell'(y_l N(\boldsymbol{\theta}, \mathbf{x}_l))}$. As a result, for all $j \in [n]$ we have

$$\begin{aligned}\hat{\mathbf{w}}_j =& \mathbf{w}_j - \lambda_l y_l \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_l) \\ =& \sum_{i=1}^{m} \lambda_i y_i \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_{\epsilon,j} - \sum_{l \in I_f} \lambda_l y_l \nabla_{\mathbf{w}_j} N(\boldsymbol{\theta}, \mathbf{x}_l) = \sum_{i \in I_r} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i + \mathbf{v}_{\epsilon,j} \ .\end{aligned}$$

Given $\hat{\boldsymbol{\theta}}$ and the unlearned sample indices $l \in I_f$, we denote $c = \frac{\epsilon_d}{2mn}$, and for $j \in [n]$, we denote:

$$\Delta_j := \sum_{k \in S_r} c \mathbf{x}_k \operatorname{sign}(\langle \mathbf{x}_k, \mathbf{w}_j \rangle) \ .$$

Using Δ_j, we define a slightly modified weight vector $\widetilde{\boldsymbol{\theta}}$, such that for $j \in [n]$,

$$\widetilde{\mathbf{w}}_j = \hat{\mathbf{w}}_j + \sum_{l \in I_f} |u_j| \lambda_l \sigma'_{l,j} \Delta_j \ .$$

The first main challenge of this proof is Lemma C.5, that is proven for a single point unlearning. However, browsing through the proof one can see that its main observation is about the difference between the inner product of some training sample $\mathbf{x}_r$ in either the original or the fixed unlearn weight voters. Looking at the difference our case -

$$\begin{aligned}\langle \mathbf{w}_j, \mathbf{x}_r \rangle - \langle \widetilde{\mathbf{w}}_j, \mathbf{x}_r \rangle = \sum_{l \in I_f} u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle - \sum_{l \in I_f} |u_j| \lambda_l \sigma'_{l,j} \Delta_j = \\ \sum_{l \in I_f} \left(u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \lambda_l \sigma'_{l,j} \Delta_j \right) \ ,\end{aligned}$$

one can see that for any $l \in I_f$:

$$u_j \lambda_l y_l \sigma'_{l,j} \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \lambda_l \sigma'_{l,j} \Delta_{j,r} = \lambda_l \sigma'_{l,j} \left(u_j \langle \mathbf{x}_l, \mathbf{x}_r \rangle - |u_j| \Delta_{j,r} \right) \ ,$$

which is the exact same modification that in Lemma C.5 is proven to not effect the sign. Thus, using Lemma C.5 for any $l \in S_f$ will conclude in

$$\operatorname{sign}(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_r) = \operatorname{sign}(\mathbf{w}_j^\top \mathbf{x}_r) \ .$$

The next important issue we need to address to use the similar proof for forgetting multiple points is the norm of the fix. If we denote $\mathbf{u} = (\sum_{l \in I_f} |u_1| \lambda_l \sigma'_{l,1} \Delta_1, .., \sum_{l \in I_f} |u_n| \lambda_l \sigma'_{l,n} \Delta_n)$ we get a factor k in the upper bound for $\|\mathbf{u}\|$, using Lemma C.2:

$$\begin{aligned}\left\| \sum_{l \in I_f} |u_j| \lambda_l \sigma'_{l,j} \Delta_j \right\| = \sum_{l \in I_f} |u_j| \lambda_l \sigma'_{l,j} \|\Delta_j\| \leq & k \frac{1}{\sqrt{n}} 20.4 n \frac{\epsilon_d \sqrt{1.1 + \frac{\epsilon_d}{n}}}{2\sqrt{mn}} \\ \leq & k \frac{1}{\sqrt{n}} 20.4 \frac{\epsilon_d \sqrt{1.1 + \frac{\epsilon_d}{n}}}{2\sqrt{m}} \\ \leq & \frac{k \epsilon_d \left(20.4 + \frac{1}{2}\sqrt{1.1 + \frac{\epsilon_d}{n}}\right)}{\sqrt{nm}} \leq \frac{22 k \epsilon_d}{\sqrt{mn}} \ ,\end{aligned}$$

Lastly, we add a factor k for the margin difference, by straightforward accumulating the margin difference for each $l \in I_f$, getting

$$-\frac{9k\epsilon_d}{mn} \leq y_r \left[N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - N(\boldsymbol{\theta}, \mathbf{x}_r) \right] \leq \frac{9k\epsilon_d}{mn} \ ,$$

We now ready to prove the multi-point version.

PROOF OF $\widetilde{\theta}$ HAS THE DIRECTION OF A $(\epsilon + \frac{9\epsilon_d\epsilon}{\frac{m}{k}-9\epsilon_d} + \frac{23k\epsilon_d}{\sqrt{m}}, \delta + \frac{9\epsilon_d\delta}{\frac{m}{k}-9\epsilon_d} + \frac{22.6k\epsilon_d}{m})$-APPROXIMATE KKT POINT OF THE MARGIN MAXIMIZATION PROBLEM (2) W.R.T. $S \setminus \{\mathbf{x}_l, y_l\}$:

(1) DUAL FEASIBILITY: FOR ALL $r \in [m]_{-l}$, $\lambda_r \geq 0$.

Same. Directly from dual feasibility for $\boldsymbol{\theta}$ (Definition 2.1).

(2) STATIONARITY: $\left\| \widetilde{\boldsymbol{\theta}} - \sum_{i \in I_r} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i) \right\| \leq \epsilon + \frac{22k\epsilon_d}{\sqrt{m}}$.

We showed that for $j \in [n], i \in [m]$, $\mathbb{1}_{\{\widetilde{\mathbf{w}}_j^T \mathbf{x}_j \geq 0\}} = \mathbb{1}_{\{\mathbf{w}_j^T \mathbf{x}_j \geq 0\}}$, thus similarly having

$$\begin{aligned}
&\left\| \widetilde{\boldsymbol{\theta}} - \sum_{i \in I_r} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i) \right\| = \\
&= \left\| \sum_{i \in I_r} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_\epsilon + \mathbf{u} - \sum_{i \in I_r} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i) \right\| \\
&= \|\mathbf{v}_\epsilon + \mathbf{u}\| \leq \|\mathbf{v}_\epsilon\| + \|\mathbf{u}\| \ .
\end{aligned}$$

Using the upper bound from we showed, we have

$$\begin{aligned}
\|\mathbf{u}\| = \left\| (\sum_{l \in I_f} |u_1| \lambda_l \sigma'_{l,1} \Delta_1, .., \sum_{l \in I_f} |u_n| \lambda_l \sigma'_{l,n} \Delta_n) \right\| &= \sqrt{\sum_{j=1}^{n} \left\| \sum_{l \in I_f} u_j \lambda_l \sigma'_{l,j} \Delta_j \right\|^2} \leq \\
&\leq \sqrt{n} \max_{j \in [n]} \sum_{l \in I_f} |u_j| \lambda_l \sigma'_{l,j} \|\Delta_j\| \\
&\leq \sqrt{n} \frac{22k\epsilon_d}{\sqrt{mn}} \leq \frac{22k\epsilon_d}{\sqrt{m}} \ ,
\end{aligned}$$

(3) COMPLEMENTARITY SLACKNESS: FOR ALL $r \in [m]_{-l}$, $\lambda_r \left(y_r N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_r) - 1 \right) \leq \delta + \frac{184k\epsilon_d}{m}$.

Same proof using the modified margin difference $\frac{9k\epsilon_d}{mn}$.

(4) PRIMAL FEASIBILITY: FOR ALL $r \in [m]_{-l}$, $y_i N(\mathbf{x}_i, \widetilde{\boldsymbol{\theta}}) \geq 1 - \frac{9k\epsilon_d}{mn}$.

Same.

To conclude, $\widetilde{\boldsymbol{\theta}}$ is an $(\epsilon + \frac{22k\epsilon_d}{\sqrt{m}}, \delta + \frac{184k\epsilon_d}{m}, \frac{9k\epsilon_d}{mn})$-approximate KKT for the margin maximization problem (2) w.r.t. S_r (Definition B.1). Using Lemma B.3 we conclude that $\frac{1}{1-\frac{9k\epsilon_d}{mn}} \widetilde{\boldsymbol{\theta}}$ is an $(\epsilon + \frac{9\epsilon_d\epsilon}{\frac{m}{k}-9\epsilon_d} + \frac{23k\epsilon_d}{\sqrt{m}}, \delta + \frac{9\epsilon_d\delta}{\frac{m}{k}-9\epsilon_d} + \frac{22.6k\epsilon_d}{m})$-approximate KKT for the margin maximization problem (2) w.r.t. S_r according to Definition 2.1, which finish the proof.

PROOF OF $\text{cossim}(\hat{\theta}, \widetilde{\theta}) \geq 1 - \frac{82k\epsilon_d}{m-k}$:

For the cosine similarly, by noting that $\widetilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \mathbf{u}$, we have that (same as C.2)

$$\text{cossim}(\hat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}) \geq 1 - \frac{2\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|} \ .$$

we have $\|\mathbf{u}\| \leq \frac{22k\epsilon_d}{\sqrt{m}}$ and for $\left\|\hat{\boldsymbol{\theta}}\right\|$, we can follow that same proof with only replace $\sum_{i\in[m]_{-l}} \lambda_i$ with $\sum_{i\in I_r} \lambda_i$, which will slightly effect the norm, having

$$\left\|\hat{\boldsymbol{\theta}}\right\|^2 \geq 0.3(m-k) \ .$$

Thus, we get for the ratio:

$$\frac{\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|} \leq \frac{\frac{22k\epsilon_d}{\sqrt{m}}}{\sqrt{0.3(m-k)}} \leq \frac{41k\epsilon_d}{m-k}$$

Joining it all together we have

$$\text{cossim}(\hat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}) \geq 1 - \frac{2\,\|\mathbf{u}\|}{\left\|\hat{\boldsymbol{\theta}}\right\|} \geq 1 - \frac{82k\epsilon_d}{m-k} \ ,$$

which conclude the proof. □

C.4 The Identity is an Unsuccessful Unlearning Algorithm

Similarly to the linear case, we complement Theorem 4.1 by providing the following remark, that shows that keeping the original network is not a successful unlearning algorithm. Particularly, we show that for the network in Theorem 4.1, its cosine similarity to any (ϵ, δ)-approximate KKT point for $S \setminus \{(\mathbf{x}_l, y_l)\}$ is relatively large (see proof in Appendix C.4).

Remark C.1. *In the same settings as 4.1, the algorithm $\mathcal{A}_I(\boldsymbol{\theta}, S, r) = \boldsymbol{\theta}$, is (ϵ, δ, ρ)-successful only for $\rho \geq \frac{C}{m} + C(\epsilon_d + \epsilon + \widetilde{\epsilon})$ for some $C > 0$.*

Proof: In this section we show that the original network $\boldsymbol{\theta}$ is not a good candidate for the unlearning tasks according to the (ϵ, δ, τ)-successful definition (Definition 2.2). Formally, we look at the simple unlearning algorithm $\mathcal{A}_I(\boldsymbol{\theta}, S, r) = \boldsymbol{\theta}$. We show that $\boldsymbol{\theta}$ will have a small cosine-similarity with any KKT point w.r.t. the retain set $S \setminus (\mathbf{x}_l, y_l)$. Namely, that $\mathcal{A}_I$ is $(\epsilon', \delta', \tau')$ successful for τ' that is at least $O(\frac{1}{mn}) - O(\frac{\epsilon_d}{n})$.

Next, we show for $\tau > 0$. Let $\widetilde{\boldsymbol{\theta}}$ be an $(\widetilde{\epsilon}, \widetilde{\delta})$-approximate KKT point w.r.t. $S \setminus (\mathbf{x}_l, y_l)$. We show that $\tau \geq O(\frac{1}{mn}) - O(\frac{\epsilon_d}{n})$.

From stationarity for $\boldsymbol{\theta}$ w.r.t. S, and for $\widetilde{\boldsymbol{\theta}}$ w.r.t. $S \setminus (\mathbf{x}_l, y_l)$ we get that

$$\boldsymbol{\theta} = \sum_{i\in[m]} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i) + \mathbf{v}_\epsilon \ ,$$

and

$$\widetilde{\boldsymbol{\theta}} = \sum_{i\in[m]_{-l}} \widetilde{\lambda}_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i) + \mathbf{v}_{\widetilde{\epsilon}} \ .$$

We denote $\alpha_i = \sum_{j\in[n]} u_j \lambda_i \sigma'_{i,j}$ and $\widetilde{\alpha}_i = \sum_{j\in[n]} u_j \widetilde{\lambda}_i \widetilde{\sigma}'_{i,j}$, and $\underline{\boldsymbol{\theta}} = \boldsymbol{\theta} - \mathbf{v}_\epsilon, \underline{\widetilde{\boldsymbol{\theta}}} = \widetilde{\boldsymbol{\theta}} - \mathbf{v}\widetilde{\epsilon}$

By Cauchy–Schwarz inequality we have

$$\begin{aligned}
&\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \rangle = \langle \underline{\boldsymbol{\theta}} + \mathbf{v}_\epsilon, \underline{\widetilde{\boldsymbol{\theta}}} + \mathbf{v}_{\widetilde{\epsilon}} \rangle = \\
&\leq \langle \underline{\boldsymbol{\theta}}, \underline{\widetilde{\boldsymbol{\theta}}} \rangle + |\langle \mathbf{v}_\epsilon, \underline{\widetilde{\boldsymbol{\theta}}} \rangle| + |\langle \mathbf{v}_{\widetilde{\epsilon}}, \underline{\boldsymbol{\theta}} \rangle| \\
&\leq \langle \underline{\boldsymbol{\theta}}, \underline{\widetilde{\boldsymbol{\theta}}} \rangle + \epsilon \left\|\underline{\widetilde{\boldsymbol{\theta}}}\right\| + \widetilde{\epsilon}\,\|\underline{\boldsymbol{\theta}}\| \ .
\end{aligned}$$

For the inner product between the sums, we have

$$
\begin{aligned}
\langle \underline{\boldsymbol{\theta}}, \underline{\widetilde{\boldsymbol{\theta}}} \rangle &= \langle \sum_{i\in[m]_{-l}} \lambda_i y_i \nabla_{\boldsymbol{\theta}} N(\boldsymbol{\theta}, \mathbf{x}_i), \sum_{i\in[m]} \widetilde{\lambda}_i y_i \nabla_{\boldsymbol{\theta}} N(\widetilde{\boldsymbol{\theta}}, \mathbf{x}_i) \rangle = \\
&= \sum_{j\in[n]} \langle \sum_{i\in[m]} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i, \sum_{i\in[m]_{-l}} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i \rangle \\
&= \langle \sum_{i\in[m]} \sum_{j\in[n]} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i, \sum_{i\in[m]_{-l}} \sum_{j\in[n]} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i \rangle \\
&= \langle \sum_{i\in[m]} \alpha_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i y_i \mathbf{x}_i \rangle \\
&\leq |\langle \sum_{i\in[m]} \alpha_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i y_i \mathbf{x}_i \rangle| \\
&\leq \sum_{i\in[m]_{-l}} \alpha_i \widetilde{\alpha}_i \left\|\mathbf{x}_i\right\|^2 + \sum_{i\neq k\in[m]_{-l}} \alpha_i \widetilde{\alpha}_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \sum_{i\in[m]_{-l}} \alpha_l \widetilde{\alpha}_i \langle \mathbf{x}_l, \mathbf{x}_i \rangle \\
&\leq \sum_{i\in[m]_{-l}} \alpha_i \widetilde{\alpha}_i \left\|\mathbf{x}_i\right\|^2 + \phi \sum_{i\neq k\in[m]_{-l}} \alpha_i \widetilde{\alpha}_k + \phi \sum_{i\in[m]_{-l}} \alpha_l \widetilde{\alpha}_i
\end{aligned}
$$

For lower bounds of the norms we perform similar calculations. We note that $\left\|\widetilde{\boldsymbol{\theta}}\right\| \geq \left\|\underline{\widetilde{\boldsymbol{\theta}}}\right\| - \epsilon$, and

$$
\begin{aligned}
\left\|\underline{\widetilde{\boldsymbol{\theta}}}\right\|^2 &= \sum_{j\in[n]} \|\widetilde{\mathbf{w}}_j\|^2 \\
&= \sum_{j\in[n]} \left\| \sum_{i\in[m]_{-l}} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i \right\|^2 \\
&= \sum_{j\in[n]} \langle \sum_{i\in[m]_{-l}} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i, \sum_{i\in[m]_{-l}} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i \rangle \\
&= \langle \sum_{i\in[m]_{-l}} \sum_{j\in[n]} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i, \sum_{i\in[m]_{-l}} \sum_{j\in[n]} u_j \widetilde{\lambda}_i y_i \widetilde{\sigma}'_{i,j} \mathbf{x}_i \rangle \\
&= \langle \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i y_i \mathbf{x}_i \rangle \\
&\leq |\langle \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i y_i \mathbf{x}_i, \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i y_i \mathbf{x}_i \rangle| \\
&\geq \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i^2 \left\|\mathbf{x}_i\right\|^2 - \sum_{i\neq k\in[m]_{-l}} |\widetilde{\alpha}_i \widetilde{\alpha}_k| \langle \mathbf{x}_i, \mathbf{x}_k \rangle \\
&\geq \sum_{i\in[m]_{-l}} \widetilde{\alpha}_i^2 \left\|\mathbf{x}_i\right\|^2 - \phi \sum_{i\neq k\in[m]_{-l}} |\widetilde{\alpha}_i \widetilde{\alpha}_k|
\end{aligned}
$$

and similarly

$$\begin{aligned}
\|\underline{\boldsymbol{\theta}}\|^2 &= \sum_{j\in[n]} \|\mathbf{w}_j\|^2 \\
&= \sum_{j\in[n]} \left\| \sum_{i\in[m]} u_j \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i \right\|^2 \\
&\geq \sum_{i\in[m]} \alpha_i^2 \|\mathbf{x}_i\|^2 - \sum_{i\neq k\in[m]} |\alpha_i \alpha_k| \langle \mathbf{x}_i, \mathbf{x}_k \rangle \\
&\geq \alpha_l^2 \|\mathbf{x}_l\|^2 + \sum_{i\in[m]_{-l}} \alpha_i^2 \|\mathbf{x}_i\|^2 - \phi \sum_{i\neq k\in[m]_{-l}} |\alpha_i \alpha_k|
\end{aligned}$$

Plug it all in the cosine similarity definition we get

$$\operatorname{cossim}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \frac{\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \rangle}{\|\boldsymbol{\theta}\| \left\|\widetilde{\boldsymbol{\theta}}\right\|} \leq \frac{\langle \underline{\boldsymbol{\theta}}, \underline{\widetilde{\boldsymbol{\theta}}} \rangle}{\|\boldsymbol{\theta}\| \left\|\widetilde{\boldsymbol{\theta}}\right\|} + \frac{\epsilon \left\|\underline{\widetilde{\boldsymbol{\theta}}}\right\| + \widetilde{\epsilon} \|\underline{\boldsymbol{\theta}}\|}{\|\boldsymbol{\theta}\| \left\|\widetilde{\boldsymbol{\theta}}\right\|} + \frac{\epsilon \widetilde{\epsilon}}{\|\boldsymbol{\theta}\| \left\|\widetilde{\boldsymbol{\theta}}\right\|}$$

bounding the second fraction we have

$$\frac{\epsilon \left\|\underline{\widetilde{\boldsymbol{\theta}}}\right\| + \widetilde{\epsilon} \|\underline{\boldsymbol{\theta}}\|}{\|\boldsymbol{\theta}\| \left\|\widetilde{\boldsymbol{\theta}}\right\|} \leq \frac{\epsilon}{\|\boldsymbol{\theta}\|} + \frac{\widetilde{\epsilon}}{\left\|\widetilde{\boldsymbol{\theta}}\right\|}$$

and note that using Lemma C.4 and Lemma C.3, if we denote $l = \left(0.9 - 4.64\frac{\epsilon_d}{n} - 1.92\epsilon\right)$ for all $i \in [m]$

$$-l\sqrt{n} \leq \alpha_i, \widetilde{\alpha}_i \leq 20.4\sqrt{n}$$

$$\begin{aligned}
\|\boldsymbol{\theta}\|^2 &\geq \sum_{i\in[m]} \alpha_i^2 \|\mathbf{x}_i\|^2 - \phi \sum_{i\neq k\in[m]_{-l}} \alpha_i \alpha_k \\
&\geq \left(\sum_{i\in[m]} |\alpha_i| \right) \left(0.9 l\sqrt{n} - \phi 20.4 m\sqrt{n}\right) \\
&\geq ml\sqrt{n}(0.9l\sqrt{n} - \frac{5.1\epsilon_d}{\sqrt{n}}) \\
&\geq mln(0.9l - \frac{5.1\epsilon_d}{n}) \geq \frac{C}{mn}
\end{aligned}$$

and similarly $\|\boldsymbol{\theta}\|^2 > \frac{C}{mn}$ then

$$\frac{\epsilon}{\|\boldsymbol{\theta}\|} + \frac{\widetilde{\epsilon}}{\left\|\widetilde{\boldsymbol{\theta}}\right\|} \leq \frac{C(\epsilon + \widetilde{\epsilon})}{\sqrt{mn}}$$

bounding the first fraction we have

$$\frac{\langle \underline{\boldsymbol{\theta}}, \underline{\widetilde{\boldsymbol{\theta}}} \rangle}{\|\boldsymbol{\theta}\| \left\|\widetilde{\boldsymbol{\theta}}\right\|} \leq$$

$$\frac{\sum_{i\in[m]_{-l}} \alpha_i \widetilde{\alpha}_i \|\mathbf{x}_i\|^2 + \phi \sum_{i\neq k\in[m]_{-l}} \alpha_i \widetilde{\alpha}_k + \phi \sum_{i\in[m]_{-l}} \alpha_l \widetilde{\alpha}_i}{\sqrt{\sum_{i\in[m]_{-l}} \widetilde{\alpha}_i^2 \|\mathbf{x}_i\|^2 - \phi \sum_{i\neq k\in[m]_{-l}} \widetilde{\alpha}_i \widetilde{\alpha}_k - \epsilon} \sqrt{\alpha_l^2 \|\mathbf{x}_l\|^2 + \sum_{i\in[m]_{-l}} \alpha_i^2 \|\mathbf{x}_i\|^2 - \phi \sum_{i\neq k\in[m]_{-l}} \alpha_i \alpha_k - \widetilde{\epsilon}}}$$

We lower bound the norm of the parameter

$$\begin{aligned}
\|\underline{\boldsymbol{\theta}}\|^2 &= \sum_{j\in[n]} \|\mathbf{w}_j\|^2 \\
&\geq \sum_{i\in[m]} \alpha_i^2 \|\mathbf{x}_i\|^2 - \phi \sum_{i\neq k\in[m]} |\alpha_i \alpha_k| \\
&\geq (\sum_{i\in[m]} \alpha_i)[a - \phi m b] \geq m 0.9 a [a - \frac{0.6\epsilon_d}{n}]
\end{aligned}$$

As $a - \frac{0.6\epsilon_d}{n} > C$ for some $C > 0$, we note we get a similar equation as in the linear case (B.3), and skip to the result, having

$$\text{cossim}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \leq 1 - \frac{C}{m} + C(\epsilon_d + \epsilon + \widetilde{\epsilon}) \ .$$

□

D Appendix for Section 6

D.1 Proofs for settings properties

We first show this dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim \mathcal{D}_{MG}^m$ satisfy the conditions we discuss in our paper:

1. For all $\mathbf{x}_i \in S$, $\|\mathbf{x}_i\|^2 \in [1-\psi, 1+\psi]$ for $\psi = 0.1$.
2. For all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$ s.t. $i \neq j$, $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq \phi$

For a sample $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, we first show that $\mathbf{x}_i$'s norm is a bounded constant. Denote $\mathbf{x}_i = \boldsymbol{\mu}_i + \boldsymbol{\zeta}_i$ for $\|\boldsymbol{\mu}_i\| = \cdot d^{-\frac{1}{4}+\alpha}$ for $\alpha \in (0, \frac{1}{4})$, and $\boldsymbol{\zeta}_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$.

We show tighter bounds for $\|\boldsymbol{\zeta}_i\|^2$.

Lemma D.1. *Let $i \in [m]$. Then, w.p. $\geq 1 - (2e^{-\frac{d}{1700}})$, $\|\boldsymbol{\zeta}_i\|^2 \in [0.95, 1.05]$.*

Proof: For the lower bound, similar to Lemma A.1, we have for $w \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$

$$\Pr\left[n - \left\|\frac{w}{\sigma}\right\|^2 \geq 2\sqrt{nt}\right] \leq e^{-t} \ .$$

We let $t = \frac{1}{1600} \cdot n$, $\sigma^2 = \frac{1}{d}$ and $n = d$ and get

$$\Pr\left[\|w\|^2 \leq \frac{95}{100}\right] \leq e^{-\frac{d}{1600}} \ .$$

as desired. For the upper bound, similar to Lemma A.2, we have for $w \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$

$$\Pr\left[\left\|\frac{w}{\sigma}\right\|^2 - n \geq 2\sqrt{nt} + 2t\right] \leq e^{-t} .$$

We let $t = \frac{1}{1700} \cdot n$, $\sigma^2 = \frac{1}{d}$ and $n = d$ and get

$$\Pr\left[\|w\|^2 \geq 1.05\right] \leq e^{-\frac{d}{1700}} .$$

□

Lemma D.2. *w.p.* $1 - (2e^{-\frac{d}{1700}})$*, for sufficiently large d,* $\|\mathbf{x}_i\|^2 \in [0.9, 1.1]$.

Proof: We denote $\mathbf{x}_i = \boldsymbol{\mu}_i + \boldsymbol{\zeta}_i$, such that $\boldsymbol{\zeta}_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$. From Lemma D.1 we get that w.p. $1 - (2e^{-\frac{d}{1700}})$, $\|\boldsymbol{\zeta}_i\|^2 \in [0.95, 1.05]$.

As for $\|\boldsymbol{\mu}_i\|$, we note that $\|\boldsymbol{\mu}_i\|^2 = d^{2(-\frac{1}{4}+\alpha)} = d^{(-\frac{1}{2}+2\alpha)}$, therefore if enough to take d such that

$$d^{2\alpha-\frac{1}{2}} \leq 0.01 \iff d^{\frac{1}{2}-\alpha} \geq 100 \iff \log(d) \geq \frac{\log(100)}{\frac{1}{2} - \alpha}$$

Then, for such d we have,

$$\|\mathbf{x}_i\|^2 = \|\boldsymbol{\mu}_i + \boldsymbol{\zeta}_i\|^2 = \|\boldsymbol{\mu}_i\|^2 + \|\boldsymbol{\zeta}_i\|^2 + 2\langle \boldsymbol{\mu}_i, \boldsymbol{\zeta}_i \rangle$$
$$\|\boldsymbol{\mu}_i\|^2 + \|\boldsymbol{\zeta}_i\|^2 - 2|\langle \boldsymbol{\mu}_i, \boldsymbol{\zeta}_i \rangle| \leq \|\mathbf{x}\|^2 \leq \|\boldsymbol{\mu}_i\|^2 + \|\boldsymbol{\zeta}_i\|^2 + 2|\langle \boldsymbol{\mu}_i, \boldsymbol{\zeta}_i \rangle|$$
$$2|\langle \boldsymbol{\mu}_i, \boldsymbol{\zeta}_i \rangle| \leq 2\|\boldsymbol{\mu}_i\| \|\boldsymbol{\zeta}_i\| \leq 2 \cdot 0.01 \cdot 1.05 = 0.021$$

and therefore,

$$0.9 < 0.929 \leq \|\mathbf{x}_i\|^2 \leq 1.081 < 1.1$$

as desired. □

Next, we look at two samples $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \sim \mathcal{D}_{MG}$, showing that if $i \neq j$, $\mathbf{x}_i, \mathbf{x}_j$ are almost orthogonal.

Lemma D.3. *Let* $i \neq j$*, and let* $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \sim \mathcal{D}_{MG}$*. Then, for sufficiently large d, w.p.* $\geq 1 - e^{-d/500} + 6d^{-\frac{\log(d)}{2}}$*:*

$$|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| - \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle \in [-2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} - 1.1\frac{\log(d)}{\sqrt{d}}, 2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} + 1.1\frac{\log(d)}{\sqrt{d}}]$$

Proof: Let $\mathbf{x}_i, \mathbf{x}_j$ data points. We denote $\mathbf{x}_i = \boldsymbol{\mu}_i + \boldsymbol{\zeta}_i$ and $\mathbf{x}_j = \boldsymbol{\mu}_j + \boldsymbol{\zeta}_j$ We look at -

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \boldsymbol{\mu}_i + \boldsymbol{\zeta}_i, \boldsymbol{\mu}_j + \boldsymbol{\zeta}_j \rangle = \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle + \langle \boldsymbol{\mu}_i, \boldsymbol{\zeta}_j \rangle + \langle \boldsymbol{\zeta}_i, \boldsymbol{\mu}_j \rangle + \langle \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j \rangle$$

Since $\boldsymbol{\mu}_i \in \mathbb{R}^n$ and $\boldsymbol{\zeta}_j \sim \mathcal{N}(0, \frac{1}{d} I_d)$, we get from Lemma A.4 for $t = \frac{\log(d)}{\sqrt{d}}$ that w.p. $\geq 1 - 2d^{-\frac{\log(d)}{2}}$

$$|\langle \boldsymbol{\mu}_i, \boldsymbol{\zeta}_j \rangle| \leq \|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}}$$

From the same argument $|\langle \boldsymbol{\mu}_j, \boldsymbol{\zeta}_i \rangle| \leq \|\boldsymbol{\mu}_j\| \frac{\log(d)}{\sqrt{d}}$.

Finally, From Lemma A.6 we get that w.p. $\geq 1 - (e^{-d/500} + 2d^{-\frac{\log(d)}{2}})$, $|\langle \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j \rangle| \leq 1.1\frac{\log(d)}{\sqrt{d}}$. Combining all together,

$$\Pr\left[|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| - \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle \geq 2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} + 1.1\frac{\log(d)}{\sqrt{d}}\right] \leq e^{-d/500} + 6d^{-\frac{\log(d)}{2}}$$

and the claim follows. □

Lemma D.4. *For d large enough and $\|\boldsymbol{\mu}_+\| = \frac{\log(d)}{d^\alpha}$, for $\alpha \in (0, \frac{1}{4})$,*

$$\|\boldsymbol{\mu}_+\|^2 > 2\|\boldsymbol{\mu}_+\| \frac{\log(d)}{\sqrt{d}} + 1.1\frac{\log(d)}{\sqrt{d}}$$

Proof:

$$\begin{aligned}
&\|\boldsymbol{\mu}_+\|^2 - 2\|\boldsymbol{\mu}_+\| \frac{\log(d)}{\sqrt{d}} - 1.1\frac{\log(d)}{\sqrt{d}} \\
=&\frac{1}{d^{\frac{1}{2}-2\alpha}} - 2\frac{\log(d)}{d^{\alpha+3/4}} - 1.1\frac{\log(d)}{\sqrt{d}} \\
=&d^{-\frac{1}{2}}\left(d^{2\alpha} - 2\log(d)d^{-\frac{1}{4}} - 1.1\log(d)\right)
\end{aligned}$$

it's enough to find d such that

$$d^{2\alpha} \geq 2\log(d)d^{-\frac{1}{4}} + 1.1\log(d) \iff 2\alpha \geq \frac{\log\left(2\log(d)d^{-\frac{1}{4}} + 1.1\log(d)\right)}{\log d}$$

which is possible since r.h.s goes to 0 when d goes to infinity.

□

Lemma D.5. *Let a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be such that $\forall i$, $\mathbf{x}_i \in \mathbb{R}^d$ and $(\mathbf{x}_i, y_i) \sim \mathcal{D}_{MG}$, for $m \leq d$ and for sufficiently large d. Then, w.p. $\geq 1 - (2me^{-\frac{d}{1700}} + m^2e^{-d/500} + 2m^2d^{-\frac{\log(d)}{2}})$*

1. *For all $(\mathbf{x}, y) \in S$, $\|\mathbf{x}\|^2 \in [0.9, 1.1]$*
2. *For all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$, $|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \leq \phi$ for $\phi \leq \frac{\epsilon_d}{4mn}$*

Proof:

1. First,
$$\Pr\left[\forall(\mathbf{x}, y) \in S, \|\mathbf{x}\|^2 \in [0.9, 1.1]\right] = \Pr\left[\max_{(\mathbf{x},y)\in S} \|\mathbf{x}\|^2 \in [0.9, 1.1]\right],$$
and the claim follows w.p. $\geq 1 - 2me^{-\frac{d}{1700}}$, directly from using simple union, given Lemma D.2.
2. First,
$$\Pr\left[\forall(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S, |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \leq \frac{\epsilon_d}{4mn}\right] = \Pr\left[\max_{(\mathbf{x}_i,y_i),(\mathbf{x}_j,y_j)\in S} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \leq \frac{\epsilon_d}{4mn}\right].$$
From Lemma D.3 we get that w.p. $\geq 1 - e^{-d/500} + 6d^{-\frac{\log(d)}{2}}$:
$$|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| - \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle \in [-2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} - 1.1\frac{\log(d)}{\sqrt{d}}, 2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} + 1.1\frac{\log(d)}{\sqrt{d}}].$$
Therefore, we get that maximal value for $|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|$ if we take $i \neq j$ such that $y_i = y_j$, resulting in
$$|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \leq \|\boldsymbol{\mu}_i\|^2 + 2\|\boldsymbol{\mu}_i\| \frac{\log(d)}{\sqrt{d}} + 1.1\frac{\log(d)}{\sqrt{d}}$$
From Lemma D.4 one can see its enough to choose d such that
$$2\|\boldsymbol{\mu}_+\|^2 = 2\frac{1}{d^{\frac{1}{2}-2\alpha}} \leq \frac{\epsilon_d}{4mn},$$
which is possible since $\frac{\epsilon_d}{4mn}$ is given constant and $\lim_{d\to\infty} \frac{1}{d^{\frac{1}{2}-2\alpha}} = 0$. Then, from using simple union, the claim follows.

□

For the next lemma, we add few notations for readability.

1. $\phi^+_{\max} = \max_{i,j}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i = y_j\}, \phi^+_{\min} = \min_{i,j}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i = y_j\}$
2. $\phi^-_{\max} = \max_{i,j}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i \neq y_j\}, \phi^-_{\min} = \min_{i,j}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i \neq y_j\}$.

Lemma D.6. *Let a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be such that $\forall i$, $\mathbf{x}_i \in \mathbb{R}^d$ and $(\mathbf{x}_i, y_i) \sim \mathcal{D}_{MG}$. Then, for $m \leq d$ and for sufficiently large d, w.p. $\geq 1 - (me^{-d/500} + 6md^{-\frac{\log(d)}{2}})$, for all $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$:*

$$0 < \phi^+_{\max} = -\phi^-_{\min} = \|\boldsymbol{\mu}_i\| + 2\|\boldsymbol{\mu}_i\|\frac{\log(d)}{\sqrt{d}} + 1.1\frac{\log(d)}{\sqrt{d}} \leq \frac{\epsilon_d}{4mn}$$

$$0 < \phi^+_{\min} = -\phi^-_{\max} = \|\boldsymbol{\mu}_i\| - 2\|\boldsymbol{\mu}_i\|\frac{\log(d)}{\sqrt{d}} - 1.1\frac{\log(d)}{\sqrt{d}}$$

Proof: The proof is directly from Lemma D.3, using simple union bound same as Lemma D.5. Both larger than 0 from Lemma D.4. □

Lemma D.7. *Suppose a two-layer neural network $N(\boldsymbol{\theta}, \mathbf{x}) = \sum_{j=1}^n u_j\sigma(\mathbf{w}_j^\top \mathbf{x})$, trained on a dataset $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\} \sim \mathcal{D}^m_{MG}$, described in Definition 6. Assume that $\boldsymbol{\theta}$ is a KKT point of the margin maximization problem (2) w.r.t. S as in Definition 2.1. Let $(\mathbf{x}_t, y_t) \sim \mathcal{D}$, Then for all $j \in [n]$*

$$\text{sign}(\hat{\mathbf{w}}_j^\top \mathbf{x}_t) = \text{sign}(\mathbf{w}_j^\top \mathbf{x}_t) = y_t\,\text{sign}(u_j)$$

Proof: Let $(\mathbf{x}_t, y_t) \sim \mathcal{D}$. Since $\boldsymbol{\theta}$ is a KKT point, from Definition 2.1 we get that

$$\mathbf{w}_j = u_j\sum_{i=1}^m \lambda_i y_i \sigma'_{i,j}\mathbf{x}_i\ ,\ \mathbf{w}_j^\top \mathbf{x}_t = u_j\sum_{i=1}^m \lambda_i y_i \sigma'_{i,j}\langle \mathbf{x}_i, \mathbf{x}_t\rangle$$

$$\hat{\mathbf{w}}_j = u_j\sum_{i\in[m]_{-l}} \lambda_i y_i \sigma'_{i,j}\mathbf{x}_i\ ,\ \hat{\mathbf{w}}_j^\top \mathbf{x}_t = u_j\sum_{i\in[m]_{-l}} \lambda_i y_i \sigma'_{i,j}\langle \mathbf{x}_i, \mathbf{x}_t\rangle$$

where $\sigma'_{i,j} = \mathbb{1}_{\mathbf{w}_j^T\mathbf{x}_j \geq 0}$.

Case 1: $y_t = 1$.
We note that for all $i \in [m]$, $y_i\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq \phi^+_{\min} > 0$: If $y_i = 1$, $y_i\langle \mathbf{x}_i, \mathbf{x}_t\rangle = \langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq \phi^+_{\min}$, else $y_i = -1$ and $\langle \mathbf{x}_i, \mathbf{x}_t\rangle \leq \phi^-_{\max}$ so $-\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq -\phi^-_{\max} = \phi^+_{\min}$, from Lemma D.6. Therefore, for all $j \in [n]$, $\text{sign}(\hat{\mathbf{w}}_j^\top \mathbf{x}_t) = \text{sign}(\mathbf{w}_j^\top \mathbf{x}_t) = \text{sign}(u_j) = y_t\,\text{sign}(u_j)$.

Case 2: $y_t = -1$.
We note that for all $i \in [m]$, $y_i\langle \mathbf{x}_i, \mathbf{x}_t\rangle \leq \phi^-_{\max} < 0$: If $y_i = 1$, $y_i\langle \mathbf{x}_i, \mathbf{x}_t\rangle = \langle \mathbf{x}_i, \mathbf{x}_t\rangle \leq \phi^-_{\max}$, else $y_i = -1$ and $\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq \phi^+_{\min}$ so $-\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq -\phi^+_{\min} = \phi^-_{\max}$, from Lemma D.6. Therefore, for all $j \in [n]$, $\text{sign}(\hat{\mathbf{w}}_j^\top \mathbf{x}_t) = \text{sign}(\mathbf{w}_j^\top \mathbf{x}_t) = -\,\text{sign}(u_j) = y_t\,\text{sign}(u_j)$. □

D.2 Proof for Theorem 6.1

First, we note that according to Lemma D.5, w.p. $\geq 1 - (2me^{-\frac{d}{1700}} + m^2e^{-d/500} + 2m^2d^{-\frac{\log(d)}{2}})$ over the choice of S, S satisfies Assumption 2.3. For readability, the following proof we assume S satisfies Assumption 2.3. Given a data point $(\mathbf{x}_t, y_t) \sim \mathcal{D}_{MG}$, we show that

$$y_t N(\boldsymbol{\theta}, \mathbf{x}_t) = y_t\sum_{j=1}^n u_j\sigma(\mathbf{w}_j^\top \mathbf{x}_t) > 0\ .$$

We denote $\mathbf{x}_t = \boldsymbol{\mu}_t + \boldsymbol{\zeta}_t$, for $\boldsymbol{\zeta}_t \sim \mathcal{N}(0, \frac{1}{d})$. We denote $I^+ = \{i \in [m] : y_i = 1\}$, $I^- = \{i \in [m] : y_i = -1\}$. We also denote $\phi^+_{\max} = \max_{i,j\in[m]}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i = y_j\}, \phi^+_{\min} = \min_{i,j\in[m]}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i = y_j\}$ and $\phi^-_{\max} = \max_{i,j\in[m]}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i \neq y_j\}, \phi^-_{\min} = \min_{i,j\in[m]}\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle : y_i \neq y_j\}$.

Next, from

From Lemma D.6 we get that $\phi^-_{\max} = -\phi^+_{\min}$ and $\phi^-_{\min} = -\phi^+_{\max}$ Since $\boldsymbol{\theta}$ is a KKT point, from Definition 2.1 we get that

$$\mathbf{w}_j = u_j \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i \ , \ \mathbf{w}_j^\top \mathbf{x}_t = u_j \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_t\rangle$$

where $\sigma'_{i,j} = \mathbb{1}_{\mathbf{w}_j^T \mathbf{x}_j \geq 0}$.

Case 1: $y_t = 1$.
We show that $N(\boldsymbol{\theta}, \mathbf{x}_t) > 0$. From Lemma D.7, for all $j \in [n]$, $\operatorname{sign}(\mathbf{w}_j^\top \mathbf{x}_t) = \operatorname{sign}(u_j)$. Hence,

$$\begin{aligned}
N(\boldsymbol{\theta}, \mathbf{x}_t) &= \sum_{j=1, u_j<0}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_t) + \sum_{j=1, u_j\geq 0}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_t) \\
&= \sum_{j=1, u_j\geq 0}^{n} u_j \mathbf{w}_j^\top \mathbf{x}_t \\
&= \sum_{j=1, u_j\geq 0}^{n} u_j^2 \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_t\rangle \\
&= \sum_{i=1}^{m} y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle \sum_{j=1, u_j\geq 0}^{n} u_j^2 \lambda_i \sigma'_{i,j}
\end{aligned}$$

First, we note that for all $i \in [m]$, $y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq \phi^+_{\min} > 0$: If $y_i = 1$, $y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle = \langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq \phi^+_{\min}$, else $y_i = -1$ and $\langle \mathbf{x}_i, \mathbf{x}_t\rangle \leq \phi^-_{\max}$ so $-\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq -\phi^-_{\max} = \phi^+_{\min}$, from Lemma D.6. Next, since S satisfies Assumption 2.3, and $\boldsymbol{\theta}$ satisfies 2.1 for $\epsilon = \delta = 0$ we get from Lemma C.4 that for all $i \in [m]$, $\sum_{j=1, u_j\geq 0}^{n} u_j^2 \lambda_i \sigma'_{i,j} > 0$.

Case 2: $y_t = -1$.
Similarly, we show that $N(\boldsymbol{\theta}, \mathbf{x}_t) < 0$. From Lemma D.7, for all $j \in [n]$, $\operatorname{sign}(\mathbf{w}_j^\top \mathbf{x}_t) = -\operatorname{sign}(u_j)$. Hence,

$$\begin{aligned}
N(\boldsymbol{\theta}, \mathbf{x}_t) &= \sum_{j=1, u_j<0}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_t) + \sum_{j=1, u_j\geq 0}^{n} u_j \sigma(\mathbf{w}_j^\top \mathbf{x}_t) \\
&= \sum_{j=1, u_j<0}^{n} u_j \mathbf{w}_j^\top \mathbf{x}_t \\
&= \sum_{j=1, u_j<0}^{n} u_j^2 \sum_{i=1}^{m} \lambda_i y_i \sigma'_{i,j} \langle \mathbf{x}_i, \mathbf{x}_t\rangle \\
&= \sum_{i=1}^{m} y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle \sum_{j=1, u_j<0}^{n} u_j^2 \lambda_i \sigma'_{i,j}
\end{aligned}$$

We similarly note that that for all $i \in [m]$, $y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle \leq \phi^-_{\max} < 0$: If $y_i = 1$, $y_i \langle \mathbf{x}_i, \mathbf{x}_t\rangle = \langle \mathbf{x}_i, \mathbf{x}_t\rangle \leq \phi^-_{\max}$, else $y_i = -1$ and $\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq \phi^+_{\min}$ so $-\langle \mathbf{x}_i, \mathbf{x}_t\rangle \geq -\phi^+_{\min} = \phi^-_{\max}$, from Lemma D.6. And from Lemma C.4 we get that

$\sum_{j=1,u_j<0}^{n} u_j^2 \lambda_i \sigma'_{i,j} > 0$ and the claim follows.

For showing that

$$y_t N(\hat{\boldsymbol{\theta}}, \mathbf{x}_t) = y_t \sum_{j=1}^{n} u_j \sigma(\hat{\mathbf{w}}_j^\top \mathbf{x}_t) > 0 \ ,$$

the proof is almost identical. In the end of each case we look at

$$\sum_{i \in [m]_{-l}} y_i \langle \mathbf{x}_i, \mathbf{x}_t \rangle \sum_{j=1,u_j \geq 0}^{n} u_j^2 \lambda_i \sigma'_{i,j} \ ,$$

and all the same arguments holds, concluding generalization for $\hat{\boldsymbol{\theta}}$ as well, which finishes the proof.

We note that the same arguments can be used to show generalization for the case of unlearning a forget set $S_{\text{forget}} \subseteq S$ of any size $k < m$ using the extended algorithm $\mathcal{A}_{\text{k-GA}}$, discussed in section 5. In this case, we instead look at

$$\sum_{i \in S \setminus S_{\text{forget}}} y_i \langle \mathbf{x}_i, \mathbf{x}_t \rangle \sum_{j=1,u_j \geq 0}^{n} u_j^2 \lambda_i \sigma'_{i,j} \ ,$$

yet the same arguments hold, concluding generalization.

E EXPERIMENT DETAILS

We take a high dimensional data set, where $m = 10$, $d = 1000$, the data distribution is $\mathcal{N}(0, \frac{1}{d} I_d)$. As mentioned in Example. 2.4, the data satisfies Assumption 2.3 for small value of ϕ and ψ. We experiment with fully-connected ReLU networks, trained using SGD optimizer with binary cross entropy loss that is normalized to have a margin of size 1. In this experiment, for each data point $\mathbf{x}_i \in S$, we calculate λ_i, and unlearn it using the gradient ascent algorithm $\mathcal{A}_{\text{GA}}$ with step size $\alpha \lambda_i$ for $\alpha \in [0, 1.5]$, resulting in $\widetilde{\boldsymbol{\theta}}_i(\alpha)$. For each $\widetilde{\boldsymbol{\theta}}_i(\alpha)$ we calculate the corresponding ϵ, δ for its KKT conditions with respect to $S \setminus (\mathbf{x}_i, y_i)$. In Figure 1, we sample one point from S, preform the unlearning algorithm for all 10 networks, and average the results.

We test for a two-layer fully-connected ReLU network $\boldsymbol{\theta}$ as in Eq. 1, with $n = 400$. We initialize the network with small initialization for the first layer by dividing its standard deviation by a factor of 10^5. We train with full batch size for 10^5 epochs, using SGD optimizer with a 10^{-5} wight decay factor.