Near-Tight Margin-Based Generalization Bounds for Support Vector Machines

Allan Grønlund [*1] **Lior Kamma** [*1] **Kasper Green Larsen** [*1]

Abstract

Support Vector Machines (SVMs) are among the most fundamental tools for binary classification. In its simplest formulation, an SVM produces a hyperplane separating two classes of data using the largest possible margin to the data. The focus on maximizing the margin has been well motivated through numerous generalization bounds. In this paper, we revisit and improve the classic generalization bounds in terms of margins. Furthermore, we complement our new generalization bound by a nearly matching lower bound, thus almost settling the generalization performance of SVMs in terms of margins.

1. Introduction

Since their introduction (Vapnik, 1982; Cortes & Vapnik, 1995) *Support Vector Machines (SVMs)* have continued to be among the most popular classification algorithms. In the most basic setup an SVM produces, upon receiving a training data set, a classifier by finding a maximum margin hyperplane separating the data. More formally, given a training data set $S = \{x_1, \ldots, x_m\}$ of m samples in $\mathbb{R}^d$, each with a label $y_i \in \{-1, +1\}$, an SVM finds a unit vector $w \in \mathbb{R}^d$ such that $y_i \langle x_i, w \rangle \geq \theta$ for all i, with the largest possible value of the margin θ. Note that one often includes a bias parameter b such that one instead requires $y_i(\langle x_i, w \rangle + b) \geq \theta$. As b has no relevance on this work we ignore it for notational simplicity. The predicted label on a new data data point $x \in \mathbb{R}^d$, is simply $\text{sign}(\langle x, w \rangle)$. When the data is linearly separable, that is there exists a vector w with $y_i \langle x_i, w \rangle > 0$ for all i, then the maximum margin hyperplane w is the solution to the following convex optimization problem, which is often referred to as the *hard margin SVM*.

$$\begin{aligned} &\min_{w} \ \|w\|_2^2 \\ &s.t.\ y_i \langle x_i, w \rangle \geq 1 \ \ \forall i. \end{aligned} \tag{1}$$

Note that the maximum margin hyperplane is not necessarily a vector w of unit norm. If we however let $w^* = w/\|w\|_2$, then by linearity, we get a unit vector w^* such that $y_i \langle x_i, w^* \rangle \geq 1/\|w\|_2$ for all i. That is, the margin becomes at least $1/\|w\|_2$ for all (x_i, y_i).

As data is typically not linearly separable, one often considers a relaxed variant of the above optimization problem, known as *soft margin SVM* (Cortes & Vapnik, 1995).

$$\begin{aligned} &\min_{w,\xi} \ \|w\|_2^2 + \lambda \sum\nolimits_i \xi_i \\ &s.t.\ y_i \langle x_i, w \rangle \geq 1 - \xi_i \ \ \forall i. \\ &\quad \xi_i \geq 0 \ \ \forall i. \end{aligned} \tag{2}$$

Here $\lambda \geq 0$ is a hyper parameter which, roughly speaking, controls the tradeoff between the magnitude of the margin $\theta = 1/\|w\|_2$ and the number of data points with margin significantly less than θ. The soft margin optimization problem is also convex and can be solved efficiently.

A key reason for the success of SVMs is the extensive study and ubiquitousness of kernels (see e.g. (Boser et al., 1992)). By allowing efficient calculation of inner products in high (or even infinite) dimensional spaces, kernels make it possible to apply SVMs in these spaces through feature transforms without actually having to compute the feature transform, neither during training or prediction. Predictions are efficient since they only need to consider the support vectors. These are the sample data points (x, y) that are not *strictly* on the correct side of the margin of the hyperplane, meaning that $y \langle x, w \rangle \leq \theta$.

Feature transforms, like the application of a kernel, often drastically increase the dimensionality of the input domain, directly increasing the the VC-dimension of the hypothesis set (the set of hyperplanes) the same way. Thus one might worry about overfitting. However, SVMs, even with

[*]Equal contribution [1]Department of Computer Science, Aarhus University, Denmark. Correspondence to: Allan Grønlund <jallan@cs.au.dk>, Lior Kamma <lior.kamma@cs.au.dk>, Kasper Green Larsen <larsen@cs.au.dk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

the Gaussian kernel that maps to an infinite dimensional space, often generalize well to new data points in practice. Explaining this phenomenon has been the focus of much theoretical work, see e.g. (Vapnik, 1982; Bartlett & Shawe-Taylor, 1999; Bartlett & Mendelson, 2002), with probably the most prominent and simplest explanations being based on generalization bounds involving margins. These margin generalization bounds show that, as long as a hypothesis vector has large margins on most training data, then the hypothesis generalizes well to new data, independent of the dimension of the data. Further strengthening these generalization bounds and our understanding of the influence of margins is the focus of this paper. We start by reviewing some of the previous margin-based generalization bounds for SVMs.

1.1. Previous Generalization Bounds

In what follows we review previous generalization bounds for SVMs. We have focused on the most classic bounds, taking only the margin θ, the radius R of the input space, and the number of data samples m into account. We have rephrased the previous theorems to put them all into the same form, allowing for easier comparison between them. Throughout X denotes the input space, $\mathcal{D}$ a distribution over $X \times \{-1, 1\}$, and $\mathcal{L}_{\mathcal{D}}(w)$ the out-of-sample error for a vector w. That is $\mathcal{L}_{\mathcal{D}}(w) = \Pr_{(x,y)\sim\mathcal{D}}[\text{sign}(\langle x, w\rangle) \neq y] = \Pr_{(x,y)\sim\mathcal{D}}[y\langle x, w\rangle \leq 0]$. Given a training set S and a margin θ, $\mathcal{L}_S^\theta(w)$ denotes the in-sample margin error for a vector w, i.e. $\mathcal{L}_S^\theta(w) = \Pr_{(x,y)\sim S}[y\langle x, w\rangle \leq \theta]$, where $(x, y) \sim S$ means a that (x, y) is sampled from S uniformly at random.

The first work trying to explain the generalization performance of SVMs through margins is due to (Bartlett & Shawe-Taylor, 1999). They first consider the linearly separable case/hard margin SVM and prove the following generalization if all samples have margins at least θ:

Theorem 1. *[(Bartlett & Shawe-Taylor, 1999)] Let $d \in \mathbb{N}^+$ and let $R > 0$. Denote by X the ball of radius R in $\mathbb{R}^d$ and let $\mathcal{D}$ be any distribution over $X \times \{-1, 1\}$. For every $\delta > 0$, it holds with probability at least $1 - \delta$ over a set of m samples $S \sim \mathcal{D}^m$, that for every $w \in \mathbb{R}^d$ with $\|w\|_2 \leq 1$, if all samples $(x, y) \in S$ have margin (i.e. $y\langle x, w\rangle$) at least $\theta > 0$, then:*

$$\mathcal{L}_{\mathcal{D}}(w) \leq O\left(\frac{(R/\theta)^2 \ln^2 m + \ln(1/\delta)}{m}\right).$$

They complemented their bound with a generalization bound for the soft margin SVM setting, showing that in addition for all $\theta > 0$,

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^\theta(w) + O\left(\sqrt{\frac{(R/\theta)^2 \ln^2 m + \ln(1/\delta)}{m}}\right).$$

Notice how the generalization error in the soft margin case is larger due as $\sqrt{x} \geq x$ for $x \in [0, 1]$. This fits well with classic VC-dimension generalization bounds for the realizable and non-realizable setting, see e.g. (Vapnik & Chervonenkis, 2015; Ehrenfeucht et al., 1989; Anthony & Bartlett, 2009).

This bound was later improved by (Bartlett & Mendelson, 2002), who showed, using Rademacher complexity, that for all $\theta > 0$,

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^\theta(w) + O\left(\sqrt{\frac{(R/\theta)^2 + \ln(1/\delta)}{m}}\right). \quad (3)$$

Ignoring logarithmic factors and the dependency on δ, both bounds show similar dependencies on the radius of the point set R, the margin θ and the number of samples m. The dependency on R/θ also fits well with the intuition that scaling the data distribution should not change the generalization performance. Finally notice how the soft margin bounds allow one to consider any margin θ, not just the smallest over all samples, and then pay an additive term proportional to the fraction of points in the sample with margin less than θ (i.e. $\mathcal{L}_S^\theta(w) = \Pr_{(x,y)\sim S}[y\langle x, w\rangle \leq \theta]$).

Finally, the work by McAllester (McAllester, 2003), uses a PAC-Bayes argument to give a bound that attempts to interpolate between the hard margin and soft margin case. His bound shows that for all $\theta > 0$, we have:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^\theta(w) &+ O\left(\frac{(R/\theta)^2 \ln m}{m}\right) \\ &+ O\left(\sqrt{\frac{(R/\theta)^2 \ln m}{m} \cdot \mathcal{L}_S^\theta(w)}\right) \\ &+ O\left(\sqrt{\frac{\ln m + \ln(1/\delta)}{m}}\right).\end{aligned} \quad (4)$$

Notice that in the hard margin case, we have $\mathcal{L}_S^\theta(w) = 0$ and thus the above simplifies to $O((R/\theta)^2 \ln(m)/m) + O(\sqrt{(\ln m + \ln(1/\delta))/m})$. The first term is an $O(\ln m)$ factor better than the hard margin bound by Bartlett and Shawe-Taylor (Theorem 1), but unfortunately it is dominated by the $\sqrt{(\ln m + \ln(1/\delta))/m}$ term for all but very small margins, that is $\theta \leq O(R(\ln(m)/m)^{1/4})$.

These classic bounds have not seen any improvements for almost two decades, even though we have no generalization lower bounds that rule out further improvements. Generalization bounds for SVMs that are independent of the dimensionality of the space has also been proved based on the (expected) number of support vectors (Vapnik, 1982).

1.2. Our Contributions

Our first main contribution is an improvement over the known margin-based generalization bounds for a large range

of parameters. Our new generalization bound is as follows:

Theorem 2. *Let $d \in \mathbb{N}^+$ and let $R > 0$. Denote by X the ball of radius R in $\mathbb{R}^d$ and let $\mathcal{D}$ be any distribution over $X \times \{-1, 1\}$. For every $\delta > 0$, it holds with probability at least $1 - \delta$ over a set of m samples $S \sim \mathcal{D}^m$, that for every $w \in \mathbb{R}^d$ with $\|w\|_2 \leq 1$ and every margin $\theta > 0$, we have*

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^{\theta}(w) + O\left(\frac{(R/\theta)^2 \ln m + \ln(1/\delta)}{m} + \sqrt{\frac{(R/\theta)^2 \ln m + \ln(1/\delta)}{m} \cdot \mathcal{L}_S^{\theta}(w)}\right).$$

When comparing our new bound to the previous hard margin bound, i.e. every margin is at least θ, note that the previous strongest results were Theorem 1 and the bound in (4) (setting $\mathcal{L}_S^{\theta}(w) = 0$). Theorem 2 improves the former by a logarithmic factor and improves the additive $O\left(\sqrt{(\ln m + \ln(1/\delta))/m}\right)$ term in the latter to $O(\ln(1/\delta)/m)$. For soft margin the best known bounds are (3) and (4). We improve over the former (3) for any choice of margin θ with $\mathcal{L}_S^{\theta}(w) < 1/\ln m$ and we improve over (4) once again by replacing the additive $O\left(\sqrt{(\ln m + \ln(1/\delta))/m}\right)$ term by $O(\ln(1/\delta)/m)$.

A natural question to ask is whether this new bound is close to optimal. In particular, for $\delta = \Omega(1)$, our new generalization bound simplifies to:

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^{\theta}(w) + O\left(\frac{R^2 \ln m}{\theta^2 m} + \sqrt{\frac{R^2 \ln m \cdot \mathcal{L}_S^{\theta}(w)}{\theta^2 m}}\right).$$

and the generalization bound in (3) becomes:

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^{\theta}(w) + O\left(\sqrt{\frac{R^2}{\theta^2 m}}\right).$$

Summarizing the two, we get:

Corollary 3. *Let $d \in \mathbb{N}^+$ and let $R > 0$. Denote by X the ball of radius R in $\mathbb{R}^d$ and let $\mathcal{D}$ be any distribution over $X \times \{-1, 1\}$. Then it holds with constant probability over a set of m samples $S \sim \mathcal{D}^m$, that for every $w \in \mathbb{R}^d$ with $\|w\|_2 \leq 1$ and every margin $\theta > 0$, we have*

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^{\theta}(w) + O\left(\frac{R^2 \ln m}{\theta^2 m} + \sqrt{\frac{R^2}{\theta^2 m} \cdot \min\{\ln m \cdot \mathcal{L}_S^{\theta}(w), 1\}}\right).$$

At first glance the bound presented in Corollary 3 might seem odd. The first expression inside the O-notation, which intuitively stands for the hard-margin bound, incorporates a $\ln m$ factor, while the second term, which intuitively stands for the soft-margin bound does not. Our second main result, however, demonstrates that Corollary 3 is in fact tight for most ranges of parameters. Specifically, one cannot remove the extra $\ln m$ factor for the hard-margin case.

Theorem 4. *There exists a universal constant $C > 0$ such that for every $R \geq C\theta$, every $m \geq (R^2/\theta^2)^{1.001}$ and every $0 \leq \tau \leq 1$, there exists a distribution $\mathcal{D}$ over $X \times \{-1, +1\}$, where X is the ball of radius R in $\mathbb{R}^u$ for some u, such that with constant probability over a set of m samples $S \sim \mathcal{D}^m$, there exists a vector w with $\|w\|_2 \leq 1$ and $\mathcal{L}_S^{\theta}(w) \leq \tau$ satisfying:*

$$\mathcal{L}_{\mathcal{D}}(w) \geq \mathcal{L}_S^{\theta}(w) + \Omega\left(\frac{R^2 \ln m}{\theta^2 m} + \sqrt{\frac{R^2 \ln(\tau^{-1})\tau}{\theta^2 m}}\right)$$
$$\geq \mathcal{L}_S^{\theta}(w) + \Omega\left(\frac{R^2 \ln m}{\theta^2 m} + \sqrt{\frac{R^2 \ln(\mathcal{L}_S^{\theta}(w)^{-1})\mathcal{L}_S^{\theta}(w)}{\theta^2 m}}\right).$$

Together with Theorem 4, Corollary 3 gives the first completely tight generalization bounds in the hard margin case (by setting $\tau = 0$ in Theorem 4, and defining $0 \ln(0^{-1}) = 0$). For the soft margin SVM case, the bounds are only off from one another by a factor

$$\sqrt{\ln m / \ln(\mathcal{L}_S^{\theta}(w)^{-1})}$$

i.e. they asymptotically match when $\mathcal{L}_S^{\theta}(w) \leq m^{-\varepsilon}$ for an arbitrarily small constant $\varepsilon > 0$. Our generalization lower bound also shows that the previous generalization bound in (3) is tight when $\mathcal{L}_S^{\theta}(w) \geq \varepsilon$ for any constant $\varepsilon > 0$. Thus our main results settle the generalization performance of Support Vector Machines in terms of the classic margin-based parameters for all ranges of $\mathcal{L}_S^{\theta}(w)$ not including $m^{-o(1)} \leq \mathcal{L}_S^{\theta}(w) \leq o(1)$.

We remark that our upper bound generalize to infinite dimension as it only depends on the ability for performing Johnson Lindenstrauss transforms of the data which works for Hilbert spaces in general (Johnson & Lindenstrauss, 1984).

2. Margin-Based Generalization Upper Bound

This section is devoted to the proof of Theorem 2, and we start by recollecting some notation. To this end, let $d \in \mathbb{N}^+$ and let $R, \delta > 0$. Let $\mathcal{D}$ be some distribution over $X \times \{-1, 1\}$, where X is the R-radius ball around the origin in $\mathbb{R}^d$, and let $\mathcal{H}$ denote the unit ball in $\mathbb{R}^d$. Finally, let $\mathcal{E} \subseteq (X \times \{-1, 1\})^m$ include all sequences $S \in (X \times \{-1, 1\})^m$ such that for every $w \in \mathcal{H}$ and $\theta > 0$,

$$\mathcal{L}_{\mathcal{D}}(w) \leq \mathcal{L}_S^{\theta}(w) + O\left(\pi + \sqrt{\pi \mathcal{L}_S^{\theta}(w)}\right),$$

where $\pi = \pi(\delta) = \frac{(R/\theta)^2 \ln m + \ln(1/\delta)}{m}$. In these notations the theorem states that $\Pr_{S\sim\mathcal{D}^m}[\mathcal{E}] \geq 1-\delta$.

Key Tools and Techniques. One known method to prove such bounds (see, e.g. (Schapire et al., 1998; Gao & Zhou, 2013)) is to discretize the set of classifiers (or hyperplanes) and then union bound over the discrete set. When considering hyperplanes in $\mathbb{R}^d$, however, the discretization results in too large a set, which in turn means that the resulting union bound gives too large a probability bound. More specifically, the size of the set depends on the dimension d. In order to overcome this difficulty, and give generalization upper bound for a general d-dimensional distribution $\mathcal{D}$ we first reduce the dimension of the data set to a small dimension while approximately maintaining the geometric structure of the data set. That is, the dot products of a set points $x \in X$ with hyperplanes $w \in \mathcal{H}$ are maintained by the projection with high probability. More specifically, we randomly project both balls X and $\mathcal{H}$ onto a small dimension k, while approximately preserving the inner products. The random linear projection we use is simply a matrix whose every entry is sampled independently from a standard normal distribution. While this projection matrix has been studied in previous applications of dimensionality reduction (Johnson & Lindenstrauss, 1984; Dasgupta & Gupta, 2003), we present some new analysis and give tight bounds that show that inner product values in $X \times \mathcal{H}$ are well-preserved with high probability by the projection. We next discretize the set of hyperplanes in $\mathbb{R}^k$, using techniques inspired by (Alon & Klartag, 2017), and show that it is enough to union bound over the resulting small grid.

We now turn to prove the theorem. Note first that if $\theta > R$ then the bound is trivial, since for every S, $\Pr_{(x,y)\sim S}[y\langle x, w\rangle \leq \theta] = 1$. We may therefore assume hereafter that $\theta \in (0, R]$. Similarly we assume that $m \geq (R/\theta)^2 \ln m + \ln(1/\delta)$. To show that $\mathcal{E}$ occurs with high probability, we next define a sequence $\{\mathcal{E}_k\}_{k\in\mathbb{N}^+}$ of events whose intersection is contained in $\mathcal{E}$ and has probability at least $1-\delta$. In order to define the sequence $\{\mathcal{E}_k\}_{k\in\mathbb{N}^+}$ we start by defining, for every $w \in \mathcal{H}$ and every positive integer $k \in \mathbb{N}^+$, a distribution $\mathcal{Q}_k(w)$ over $\mathbb{R}^d \to \mathbb{R}$. Loosely speaking, every function $g \in \mathrm{supp}(\mathcal{Q}_k(w))$ takes a vector $x \in \mathbb{R}^d$, projects it into $\mathbb{R}^k$ and then takes its inner product with a vector $\tilde{w} \in \mathbb{R}^k$. The vector $\tilde{w}$ is the projection of w into $\mathbb{R}^k$ rounded to a predefined grid in $\mathbb{R}^k$. Formally, we next describe the process that samples $g \sim \mathcal{Q}_k(w)$. First sample a projection matrix $A \in \mathbb{R}^{k\times d}$ from $\mathbb{R}^d$ to $\mathbb{R}^k$. Every entry of A is independently sampled from a normal distribution $\mathcal{N}(0, 1/k)$ with mean 0 and variance $1/k$. Next, we define the vector $\tilde{w}$, which is a randomized rounding of Aw to the grid of vectors in $\mathbb{R}^k$ whose every entry is a whole multiple of $1/\sqrt{k}$. For every $j \in [k]$, let ℓ be the unique integer such that $\ell \leq \sqrt{k}[Aw]_j < \ell + 1$. Set $\tilde{w}_j = \ell/\sqrt{k}$ with probability $(\ell+1) - \sqrt{k}[Aw]_j$ and $\tilde{w}_j = (\ell+1)/\sqrt{k}$ otherwise, independently for every $j \in [k]$ and independently of the choice of A. Finally, define $g : \mathbb{R}^d \to \mathbb{R}$ by $g(x) = \langle Ax, \tilde{w}\rangle$ for every $x \in \mathbb{R}^d$. For every $w \in \mathcal{H}$ and every $g \in \mathrm{supp}(\mathcal{Q}_k(w))$ denote by $A_g \in \mathbb{R}^{k\times d}$ the matrix associated with g. Note that the choice of A_g does not depend on w. If w is clear from context we simply write $\mathcal{Q}_k$ instead of $\mathcal{Q}_k(w)$.

Finally, for every $k \in \mathbb{N}^+$, let Δ_k be the set of all vectors $v \in \mathbb{R}^k$ satisfying that $\|v\|_2^2 \leq 6$ and for every $j \in [k]$, $v_j\sqrt{k}$ is an integer. We are now ready to define the sequence $\{\mathcal{E}_k\}_{k\in\mathbb{N}^+}$ of events.

Definition 1. *Let $k \in \mathbb{N}^+$. For every $A \in \mathbb{R}^{k\times d}$ and $S \in \mathrm{supp}(\mathcal{D}^m)$, we say that A and S are* compatible *if for all $v \in \Delta_k$ and $\ell \in [10k]$,*

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D}}\left[y\langle Ax, v\rangle \leq \frac{\ell R}{10k}\right] \\ &\leq \Pr_{(x,y)\sim S}\left[y\langle Ax, v\rangle \leq \frac{\ell R}{10k}\right] + \frac{8\ln(2^{9k}/\delta)}{m} \\ &+ 4\sqrt{\Pr_{(x,y)\sim S}\left[y\langle Ax, v\rangle \leq \frac{\ell R}{10k}\right]\cdot\frac{\ln(2^{9k}/\delta)}{m}}\,.\end{aligned} \tag{5}$$

Let $\mathcal{C}$ denote the set of all compatible pairs (A, S). Finally, let $\mathcal{E}_k$ be the set of all $S \in \mathrm{supp}(\mathcal{D}^m)$ such that for all $w \in \mathcal{H}$, $\Pr_{g\sim\mathcal{Q}_k}[(A_g, S) \in \mathcal{C}] \geq 1 - 6\cdot 2^{-k/2}$.

The next lemma implies Theorem 2 by simply applying a union bound, since $\sum_k \frac{1}{k(k+1)} = 1$.

Lemma 5. *For every $k \in \mathbb{N}^+$, $\Pr_{S\sim\mathcal{D}^m}[\mathcal{E}_k] \geq 1 - \frac{\delta}{k(k+1)}$, and moreover $\bigcap_{k\in\mathbb{N}^+}\mathcal{E}_k \subseteq \mathcal{E}$.*

We start by proving that for every k, with high probability over $S \sim \mathcal{D}^m$, $S \in \mathcal{E}_k$. The first step is to prove that for every fixed matrix A, a random sample $S \sim \mathcal{D}^m$ is compatible with A with very high probability. Using Markov's inequality we then conclude that a random sample $S \sim \mathcal{D}^m$ is, with very high probability, compatible with most projection matrices $\{A_g\}_{g\in\mathrm{supp}(\mathcal{Q}_k(w))}$ for every $w \in \mathcal{H}$. Formally, we prove the following.

Claim 6. *For every $A \in \mathbb{R}^{d\times k}$, $\Pr_{S\sim\mathcal{D}^m}[(A, S) \in \mathcal{C}] \geq 1 - \delta/2^k$.*

Proof. Let $A \in \mathbb{R}^{d\times k}$, and fix some $v \in \Delta_k$ and $\ell \in [10k]$. First note that if $\Pr_{(x,y)\sim\mathcal{D}}[y\langle Ax, v\rangle \leq \ell R/(10k)] \leq \frac{8\ln(2^{9k}/\delta)}{m}$ then (5) holds for all $S \in \mathrm{supp}(\mathcal{D}^m)$. We can therefore assume that $\Pr_{(x,y)\sim\mathcal{D}}[y\langle Ax, v\rangle \leq \ell R/(10k)] > \frac{8\ln(2^{9k}/\delta)}{m}$. Let $\gamma = \sqrt{\frac{2\ln(2^{9k}/\delta)}{m\Pr_{\mathcal{D}}[y\langle Ax,v\rangle \leq \ell R/(10k)]}}$, then $\gamma \in (0, 1/2)$, and therefore a Chernoff bound then gives that with probability at least $1 - \frac{2\delta}{2^{9k}}$ over the choice of $S \sim \mathcal{D}^m$, $\Pr_{(x,y)\sim S}[y\langle Ax, v\rangle \leq \ell R/(10k)]$

is between $(1-\gamma)\Pr_{(x,y)\sim\mathcal{D}}[y\langle Ax,v\rangle \le \ell R/(10k)]$ and $2\Pr_{(x,y)\sim\mathcal{D}}[y\langle Ax,v\rangle \le \ell R/(10k)]$. Hence since $(1-\gamma)^{-1} \le 1+2\gamma$ we get that with probability at least $1-2\delta/2^{9k}$ over the choice of S we have

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D}}[y\langle Ax,v\rangle \le \ell R/(10k)]\\ &\le (1+2\gamma)\Pr_{(x,y)\sim S}[y\langle Ax,v\rangle \le \ell R/(10k)]\ ,\end{aligned} \tag{6}$$

and moreover,

$$\gamma \le \sqrt{\frac{4\ln(2^{9k}/\delta)}{m\Pr_{(x,y)\sim S}[y\langle Ax,v\rangle \le \ell R/(10k)]}} \tag{7}$$

Plugging (7) into (6) and summing up we get that for every $v\in\Delta_k$ and $\ell\in[10k]$, with probability at least $1-2\delta/2^{9k}$ over the choice of S we have

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D}}[y\langle Ax,v\rangle \le \ell R/(10k)] \le\\ &\Pr_{(x,y)\sim S}[y\langle Ax,v\rangle \le \ell R/(10k)] + \frac{8\ln(2^{9k}/\delta)}{m}\\ &+4\sqrt{\frac{\ln(2^{9k}/\delta)}{m}\Pr_{(x,y)\sim S}[y\langle Ax,v\rangle \le \ell R/(10k)]}\end{aligned} \tag{8}$$

Union bounding over all $v\in\Delta_k$ and $\ell\in[10k]$ we get that $\Pr_{S\sim\mathcal{D}^m}[(A,S)\in\mathcal{C}] \ge 1-10k|\Delta_k|\delta/2^{9k}$. To finish the proof of the claim, we show that $|\Delta_k|\le 2^{6k}$. Let $v\in\Delta_k$, then as $|v_j\sqrt{k}|\in\mathbb{N}$ for all $j\in[k]$ then $\sum_{j\in[k]}|v_j\sqrt{k}| \le \sum_{j\in[k]}|v_j\sqrt{k}|^2 \le 6k$. Therefore the number of possible ways to construct $|v_1\sqrt{k}|,\ldots,|v_k\sqrt{k}|$ is the number of possible solutions to the equation $\sum_{j\in[k+1]}x_j = 6k$ in natural numbers, which is $\binom{7k}{6k} \le 2^{4.5k}$. Taking all possible signs into account gives $|\Delta_k|\le 2^{5.5k}$. We conclude that $\Pr_{S\sim\mathcal{D}^m}[(A,S)\in\mathcal{C}] \ge 1-\delta/2^k$. □

The following corollary follows by applying Markov's inequality. Its proof is deferred to the supplementary material.

Corollary 7. $\Pr_{S\sim\mathcal{D}^m}[\mathcal{E}_k] \ge 1-\delta/(k(k+1))$.

We next prove the second part of Lemma 5, namely that $\bigcap_{k\in\mathbb{N}^+}\mathcal{E}_k \subseteq \mathcal{E}$. We start by introducing some concentration bounds on sums of products of Gaussian random variables.

Lemma 8. *Let $A\in\mathbb{R}^{d\times k}$ be a matrix whose every entry is independently $\mathcal{N}(0,1/k)$ distributed. Then for every $u,v\in\mathbb{R}^d$ and $t\in[0,1/4)$ we have*

1. $\Pr_A[|\|Au\|_2^2 - \|u\|_2^2| > t\|u\|_2^2] \le 2e^{-0.21kt^2}$; *and*
2. $\Pr_A[|\langle Au,Av\rangle - \langle u,v\rangle| > t] \le 4e^{\frac{-kt^2}{7\|u\|_2^2\|v\|_2^2}}$.

The proof of the lemma is quite technically involved, and its proof is thus deferred to the full version of the paper. The next claim shows that with very high probability over the choice of a pair (x,y), either sampled from $\mathcal{D}$ or uniformly at random from a sample S, and the choice of $g\sim\mathcal{Q}_k(w)$, the values $\langle x,w\rangle$ and $g(x)$ cannot be too far apart.

Claim 9. *For all $w\in\mathcal{H},\theta\in(0,R]$ and $k\in\mathbb{N}^+$,*

1. $\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[y\langle x,w\rangle \le 0 \wedge yg(x) \ge 49\theta/100] \le 7e^{-\left(\frac{k}{120}\right)\left(\frac{\theta}{R}\right)^2}$; *and*
2. *For every $S\in\mathrm{supp}(\mathcal{D}^m)$,* $\Pr_{(x,y)\sim S,g\sim\mathcal{Q}_k}[y\langle x,w\rangle \ge \theta \wedge yg(x) \le \theta/2] \le 7e^{-\left(\frac{k}{120}\right)\left(\frac{\theta}{R}\right)^2}$.

Proof. Let $w\in\mathcal{H}$, $\theta>0$ and $k\in\mathbb{N}^+$. Then

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[y\langle x,w\rangle \le 0\wedge yg(x)\ge 49\theta/100] \le\\ &\qquad\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[|y\langle x,w\rangle - yg(x)| > 49\theta/100]\end{aligned}$$

Recall that for every $x\in\mathbb{R}^d$, $g(x)=\langle Ax,\tilde{w}\rangle$, where every entry of $A\in\mathbb{R}^{d\times k}$ is sampled independently from a Gaussian distribution with mean 0 and variance $1/k$, and $\tilde{w}\in\mathbb{R}^k$ is constructed by randomly rounding each entry of Aw independently to a multiple of $1/\sqrt{k}$. By the triangle inequality, the linearity of the dot product, and since $y\in\{-1,1\}$,

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[|y\langle x,w\rangle - yg(x)| > 49\theta/100]\\ &\le \Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[|\langle x,w\rangle - \langle Ax,Aw\rangle| > 49\theta/200]\\ &+ \Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[|\langle Ax,Aw-\tilde{w}\rangle| > 49\theta/200]\end{aligned} \tag{9}$$

To bound the first probability term observe that

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}\left[|\langle x,w\rangle - \langle Ax,Aw\rangle| > \frac{49\theta}{200}\right]\\ &\le \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\Pr_{g\sim\mathcal{Q}_k}\left[\left|\frac{\langle x,w\rangle}{\|x\|_2\|w\|_2} - \frac{\langle Ax,Aw\rangle}{\|x\|_2\|w\|_2}\right| > \frac{49\theta}{200R}\right]\right]\\ &\le 4e^{-\frac{k}{7}\left(\frac{49\theta}{200R}\right)^2}\ ,\end{aligned} \tag{10}$$

where the inequality before last follows from the fact that $\|w\|_2\le 1$ and $\Pr_{(x,y)\sim\mathcal{D}}[\|x\|_2\le R]=1$, and the last inequality is an application of Lemma 8.

To bound the second term in (9), fix $(x,y)\in\mathrm{supp}(\mathcal{D})$ and $A\in\mathbb{R}^{k\times d}$, and denote $Aw=\hat{w}$. Then for every $j\in[k]$ independently $\tilde{w}_j = \frac{\lfloor\sqrt{k}\hat{w}_j\rfloor}{\sqrt{k}}$ with probability $\left\lfloor\sqrt{k}\hat{w}_j\right\rfloor + 1 - \sqrt{k}\hat{w}_j$, and $\tilde{w}_j = \frac{\lfloor\sqrt{k}\hat{w}_j\rfloor+1}{\sqrt{k}}$ otherwise. Therefore for

every $j \in [k]$, $\mathbb{E}[\tilde{w}_j] = \hat{w}_j$, and thus $\mathbb{E}[\langle Ax, Aw - \tilde{w}\rangle] = 0$. A Hoeffding bound then yields

$$\Pr_{g\sim\mathcal{Q}_k}[\,|\,\langle Ax, Aw-\tilde{w}\rangle\,| > 49\theta/200 \mid A_g = A] \leq 2e^{\frac{-2(49\theta/200)^2}{\sum_{j\in[k]}[Ax]_j^2(\hat{w}_j-\tilde{w}_j)^2}} \leq 2e^{-2k\left(\frac{49\theta}{200\|Ax\|_2}\right)^2} .$$

In addition, since $\|x\|_2^2 \leq R^2$ for all $x \in X$,

$$\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[\|Ax\|_2 > \sqrt{1.25}R] \leq e^{-k/80}$$

Finally, from the law of total probability we get that

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}\left[|\langle Ax, Aw-\tilde{w}\rangle| > \frac{49\theta}{200}\right] \\ &\leq 2e^{-2k\left(\frac{49\theta}{200\sqrt{1.25}R}\right)^2} + e^{-k/80}\end{aligned} \tag{11}$$

Plugging (10) and (11) into (9) we get that

$$\Pr_{\substack{(x,y)\sim\mathcal{D}\\ g\sim\mathcal{Q}_k}}[y\,\langle x, w\rangle \leq 0 \wedge yg(x) > \theta/2] \leq 7e^{-\left(\frac{k}{120}\right)\left(\frac{\theta}{R}\right)^2} ,$$

which concludes the first part of the lemma. The proof of the second part is identical, as we did not use any property of the distribution $\mathcal{D}$ other than the fact that $\Pr_{(x,y)\sim\mathcal{D}}[\|x\|_2 \leq R] = 1$. For every $S \in \mathrm{supp}(\mathcal{D}^m)$, it holds that $\Pr_{(x,y)\sim S}[\|x\|_2 \leq R] = 1$, and the result follows. □

The next claim essentially shows that restricting the definition of compatibility of a sample S and a matrix A only to grid points in Δ_k was indeed enough. Intuitively this is due to the fact that with very high probability over the choice of $q \sim \mathcal{Q}_k(w)$, the rounding of $A_g w$ is in the grid. Formally, we show the following.

Claim 10. *For every $S \in \bigcap_{k\in\mathbb{N}} \mathcal{E}_k$, for all $w \in \mathcal{H}, \theta \in (0, R]$ and $k \in \mathbb{N}^+$,*

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[yg(x) \leq 49\theta/100] \\ &\leq \Pr_{(x,y)\sim S,g\sim\mathcal{Q}_k}[yg(x) \leq \theta/2] + 7e^{-\left(\frac{k}{120}\right)\left(\frac{\theta}{R}\right)^2} \\ &+ 30e^{-k/24} + O\left(\frac{k+\ln(1/\delta)}{m}\right. \\ &\left.+\sqrt{\frac{k+\ln(1/\delta)}{m}\cdot\Pr_{(x,y)\sim S,g\sim\mathcal{Q}_k}[yg(x) \leq \theta/2]}\right) ;\end{aligned} \tag{12}$$

Proof. Fix $S \in \bigcap_{k\in\mathbb{N}} \mathcal{E}_k$, $w \in \mathcal{H}$, $\theta \in (0, R]$ and $k \in \mathbb{N}^+$. Clearly, if $\theta \leq 10R/k$ then $7e^{-\left(\frac{k}{120}\right)\left(\frac{\theta}{R}\right)^2} \geq 1$ and therefore (12) holds. Otherwise, let ℓ be the smallest integer such that $49\theta/100 \leq \ell R/(10k)$. As $\theta \leq R$, $\ell \in [10k]$. In addition, $49\theta/100 \leq \ell R/(10k) \leq 49\theta/100 + R/(10k) \leq \theta/2$. Denote by $\mathcal{F}$ the event that $(A_g, S) \in \mathcal{C}$ and $\tilde{w} \in \Delta_k$ (recall that $\tilde{w}$ is the vector Aw, where each entry is rounded to the nearest multiple of $1/\sqrt{k}$). For every $S \in \mathrm{supp}(\mathcal{D}^m)$, $g : \mathbb{R}^d \to \mathbb{R}$ and $\xi > 0$, denote

$$\Lambda_{\mathcal{D}}^{\xi}(g) := \Pr_{(x,y)\sim\mathcal{D}}[yg(x) \leq \xi] ,$$

$$\Lambda_{S}^{\xi}(g) := \Pr_{(x,y)\sim S}[yg(x) \leq \xi] .$$

Therefore

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[yg(x) \leq 49\theta/100] \\ &\leq \Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[yg(x) \leq \ell R/(10k)] \\ &\leq \Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[yg(x) \leq \ell R/(10k) \mid \mathcal{F}] + \Pr_{g\sim\mathcal{Q}_k}[\bar{\mathcal{F}}] \\ &\leq \mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_{\mathcal{D}}^{\ell R/(10k)}(g)\middle|\,\mathcal{F}\right] + \Pr_{g\sim\mathcal{Q}_k}[\bar{\mathcal{F}}] ,\end{aligned} \tag{13}$$

By the definition of compatible pairs and linearity of expectation we get that

$$\begin{aligned}&\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_{\mathcal{D}}^{\ell R/(10k)}(g)\middle|\,\mathcal{F}\right] \\ &\leq \mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_{S}^{\ell R/(10k)}(g)\middle|\,\mathcal{F}\right] + \frac{8\ln(2^{9k}/\delta)}{m} \\ &+ 4\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\sqrt{\Lambda_{S}^{\ell R/(10k)}(g)\cdot\frac{\ln(2^{9k}/\delta)}{m}}\middle|\,\mathcal{F}\right] .\end{aligned}$$

Note that for every non-negative random variable Y and event E, $\mathbb{E}[Y|E] \leq \mathbb{E}[Y]/\Pr[E]$. We therefore turn to bound the probability of $\mathcal{F}$. By a simple union bound,

$$\Pr_{g\sim\mathcal{Q}_k}[\bar{\mathcal{F}}] \leq \Pr_{g\sim\mathcal{Q}_k}[(A_g, S) \notin \mathcal{C}] + \Pr_{g\sim\mathcal{Q}_k}[\tilde{w} \notin \Delta_k] .$$

Since $S \in \mathcal{E}_k$, $\Pr_{g\sim\mathcal{Q}_k}[(A_g, S) \notin \mathcal{C}] \leq 6 \cdot 2^{-k/2}$. Next, for every $j \in [k]$, $|\tilde{w}_j| \leq |[A_g w]_j| + 1/\sqrt{k}$. Therefore $\|\tilde{w}\|_2^2 \leq \|A_g w\|_2^2 + 1 + 2\max\{\|A_g w\|_2^2, 1\}$, and hence if $\|Aw\|_2^2 \leq 1.5$, then $\|\tilde{w}\|_2^2 \leq 6$, and therefore $\tilde{w} \in \Delta_k$. We conclude that $\Pr_{g\sim\mathcal{Q}_k}[\tilde{w} \notin \Delta_k] \leq \Pr_{g\sim\mathcal{Q}_k}[\|A_g w\|_2^2 > 1.5] \leq e^{-k/24}$, and hence $\Pr_{g\sim\mathcal{Q}_k}[\mathcal{F}] \geq 1 - 7e^{-k/24} \geq (1 + 15e^{-k/24})^{-1}$. Since, in addition, $\ell R/(10k) \leq \theta/2$ we get

$$\begin{aligned}&\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_{\mathcal{D}}^{\ell R/(10k)}(g)\middle|\,\mathcal{F}\right] \leq \\ &(1 + 15e^{-k/24})\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_{S}^{\theta/2}(g)\right] + \frac{8\ln(2^{9k}/\delta)}{m} \\ &+ 4(1 + 15e^{-k/24})\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\sqrt{\Lambda_{S}^{\theta/2}(g)\cdot\frac{\ln(2^{9k}/\delta)}{m}}\right] .\end{aligned}$$

Finally, by Jensen's inequality we get

$$\begin{aligned}&\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_{\mathcal{D}}^{\ell R/(10k)}\middle|\,(A_g,S)\in\mathcal{C}\right]\\&\leq\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_S^{\theta/2}\right]+\frac{8\ln(2^{9k}/\delta)}{m}\\&+4\sqrt{\mathbb{E}_{g\sim\mathcal{Q}_k}\left[\Lambda_S^{\theta/2}\right]\cdot\frac{\ln(2^{9k}/\delta)}{m}}+30e^{-k/24}\,.\end{aligned}$$

Plugging into (13) we get (12). □

To finish the proof of Lemma 5, let $S=\langle(x_j,y_j)\rangle_{j\in[m]}\in\bigcap_{k\in\mathbb{N}^+}\mathcal{E}_k$, fix some $w\in\mathcal{H}$ and $\theta>0$, and let $k=\left\lceil 240\left(\frac{R}{\theta}\right)^2\ln m\right\rceil$. We will show that $S\in\mathcal{E}$.

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(w)&=\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[y\langle x,w\rangle\leq 0]\\&\leq\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[yg(x)\leq 49\theta/100]+\frac{1}{m}\end{aligned}\tag{14}$$

Where the last inequality is due to Claim 9, and since $7e^{-\left(\frac{k}{120}\right)\left(\frac{\theta}{R}\right)^2}\leq 7/m^2\leq 1/m$. From Claim 10 we get

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D},g\sim\mathcal{Q}_k}[yg(x)\leq 49\theta/100]\leq\\&\Pr_{(x,y)\sim S,g\sim\mathcal{Q}_k}[yg(x)\leq\theta/2]+O\left(\frac{k+\ln(1/\delta)}{m}\right.\\&\left.+\sqrt{\frac{k+\ln(1/\delta)}{m}\cdot\Pr_{(x,y)\sim S,g\sim\mathcal{Q}_k}[yg(x)\leq\theta/2]}\right)\;;\end{aligned}\tag{15}$$

Similarly to (14) we get that

$$\Pr_{(x,y)\sim S,g\sim\mathcal{Q}_k}[yg(x)\leq\theta/2]\leq\mathcal{L}_S^{\theta}(w)+\frac{1}{m^2}\,.\tag{16}$$

Where the last inequality follows from Claim 9 and the fact that $y\langle x,w\rangle\leq\theta$ is independent of g. Finally, plugging (16) into (15) and then into (14), and assuming that $k+\ln(1/\delta)\leq m$ we get that

$$\begin{aligned}&\Pr_{(x,y)\sim\mathcal{D}}[y\langle x,w\rangle\leq 0]\leq\Pr_{(x,y)\sim S}[y\langle x,w\rangle<\theta]\\&+O\left(\pi+\sqrt{\pi\cdot\Pr_{(x,y)\sim S}[y\langle x,w\rangle<\theta]}\right)\,,\end{aligned}$$

where $\pi=\frac{(R/\theta)^2\ln m+\ln(1/\delta)}{m}$, and therefore $S\in\mathcal{E}$, and the proof of Lemma 5, and thus of Theorem 2, is now complete.

3. Existential Lower Bound

The goal of this section is to prove the generalization lower bound in Theorem 4. Our proof is split into two cases, depending on the magnitude of τ. The results we prove are as follows:

Lemma 11. *There is a universal constant $C>0$ such that for every $R\geq C\theta$ and every $m\geq(R^2/\theta^2)^{1.001}$, there exists a distribution $\mathcal{D}$ over $X\times\{-1,+1\}$, where X is the ball of radius R in $\mathbb{R}^u$ for some u, such that with constant probability over a set of m samples $S\sim\mathcal{D}^m$, there exists a vector w with $\|w\|_2\leq 1$ and $\mathcal{L}_S^{\theta}=0$ satisfying $\mathcal{L}_{\mathcal{D}}\geq\Omega\left(\frac{R^2\ln m}{\theta^2 m}\right)$.*

Lemma 12. *There is a universal constant $C>0$ such that for every $R\geq C\theta$, every $m\geq(R^2/\theta^2)^{1.001}$ and every $R^2\ln(m)/(\theta^2 m)<\tau\leq 1$, there exists a distribution $\mathcal{D}$ over $X\times\{-1,+1\}$, where X is the ball of radius R in $\mathbb{R}^u$ for some u, such that with constant probability over a set of m samples $S\sim\mathcal{D}^m$, there exists a vector w with $\|w\|_2\leq 1$ and $\mathcal{L}_S^{\theta}\leq\tau$ satisfying*

$$\mathcal{L}_{\mathcal{D}}\geq\mathcal{L}_S^{\theta}+\Omega\left(\sqrt{\frac{R^2\tau\ln(\tau^{-1})}{\theta^2 m}}\right)\,.$$

We will first show how to combine Lemma 11 and Lemma 12 to obtain Theorem 4. For any $0\leq\tau\leq 1$, every $R\geq C\theta$ for a large constant $C>0$ and every $m\geq(R^2/\theta^2)^{1.001}$, we can invoke Lemma 11 or Lemma 12 to conclude the existence of a distribution $\mathcal{D}$, such that with constant probability over a choice of m samples $S\sim\mathcal{D}^m$, there is a vector w with $\|w\|_2\leq 1$ and either:

1. $\mathcal{L}_S^{\theta}(w)=0<\tau$ and
$$\mathcal{L}_{\mathcal{D}}(w)\geq\mathcal{L}_S^{\theta}(w)+\Omega(R^2\ln m/(\theta^2 m)).$$
2. $\mathcal{L}_S^{\theta}(w)\leq\tau$ and
$$\mathcal{L}_{\mathcal{D}}(w)\geq\mathcal{L}_S^{\theta}(w)+\Omega(\sqrt{(R^2/\theta^2)\ln(\tau^{-1})\tau/m}).$$

Note that Lemma 12 strictly speaking cannot be invoked for $\tau\leq R^2\ln(m)/(\theta^2 m)$, but for such small values of τ, the expression $\sqrt{(R^2/\theta^2)\ln(\tau^{-1})\tau/m}$ becomes less than $R^2\ln m/(\theta^2 m)$ and the bound follows from Lemma 11 instead. Thus for any $0\leq\tau\leq 1$, with constant probability over S, we may find a w with $\mathcal{L}_S^{\theta}(w)\leq\tau$ and

$$\begin{aligned}&\mathcal{L}_{\mathcal{D}}(w)\leq\mathcal{L}_S^{\theta}(w)\\&+\Omega\left(R^2\ln m/(\theta^2 m)+\sqrt{(R^2/\theta^2)\ln(\tau^{-1})\tau/m}\right)\,.\end{aligned}$$

This concludes the proof of Theorem 4. The following two sections prove the two lemmas.

3.1. Small τ

In this section, we prove Lemma 11. Let m be the number of samples and assume $m\geq(R^2/\theta^2)^{1+\varepsilon}$ where $\varepsilon=0.001$. Assume furthermore that $R\geq C\theta$ for a sufficiently large constant $C>0$. We construct a distribution $\mathcal{D}$ over $\mathbb{R}^{u+1}\times$

$\{-1,+1\}$, where $u = 4e\varepsilon^{-1}m/\ln m$. The distribution $\mathcal{D}$ gives a uniform random point among $\{x_1,\ldots,x_u\}$ where x_i has its $(u+1)$'st and i'th coordinate equal to $R/\sqrt{2}$ and the rest 0. The label is always 1.

Inspired by ideas by (Grønlund et al., 2019), we will show by a coupon-collector argument that with high probability, no more than $u - R^2/\theta^2$ elements of $\{x_1,\ldots,x_u\}$ are included in the sample S. Consider repeatedly sampling elements i.i.d. uniformly at random from $\{x_1,\ldots,x_u\}$. For every $k \in \{1,\ldots,u\}$, let X_k be the number of samples between the time the $(k-1)$'th distinct element is sampled and the time the k'th distinct element is sampled. Then $X_k \sim Geom(p_k)$, where $p_k = (u-k+1)/u$. Denote $X := \sum_{k=1}^{u-t} X_k$ for $t = R^2/\theta^2$. Then:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{u-t}\frac{1}{p_k} = u\cdot\left(\sum_{k=1}^{u}\frac{1}{k} - \sum_{k=1}^{t}\frac{1}{k}\right)\\ &\geq u\cdot(\ln(u)-\ln(t)-1) = u\ln(u/(et)).\end{aligned}$$

For a large enough constant C such that $R > C\theta$, we have $\mathbb{E}[X] \geq em$. To see why this is true, recall that $u = 4e\varepsilon^{-1}m/\ln m$, and $m \geq (R^2/\theta^2)^{1+\varepsilon}$, and therefore

$$\begin{aligned}\mathbb{E}[X] &= u\ln\left(\frac{u}{et}\right) = 4e\varepsilon^{-1}\cdot\frac{m}{\ln m}\cdot\ln\left(\frac{4e\varepsilon^{-1}\cdot\frac{m}{\ln m}}{e\left(\frac{R^2}{\theta^2}\right)}\right)\\ &\geq 4e\varepsilon^{-1}\cdot\frac{m}{\ln m}\cdot\ln\left(\frac{4e\varepsilon^{-1}m^{\varepsilon/(1+\varepsilon)}}{e\ln m}\right)\\ &\geq 4e\varepsilon^{-1}\cdot\frac{\varepsilon}{2(1+\varepsilon)}\frac{m}{\ln m}\cdot\ln m \geq em\end{aligned}$$

where the inequality before last is due to the fact that for large enough $C > 0$, $\ln m < m^{\varepsilon/(2(1+\varepsilon))}$. Denote next $p_* = \min_{k\in[u-t]} p_k = (t+1)/u$, and $\lambda = m/\mathbb{E}[X]$, then $0 < \lambda \leq e^{-1}$, and following known tail bounds on the sum of geometrically-distributed random variables (e.g. (Janson, 2018)) we get:

$$\Pr[X \leq m] = \Pr[X \leq \lambda\mathbb{E}[X]] \leq e^{-p_*\mathbb{E}[X](\lambda-1-\ln\lambda)}.$$

As $\lambda \leq e^{-1}$ we get that $1+\ln\lambda < 0$, and therefore

$$\Pr[X \leq m] \leq e^{-\frac{t+1}{u}\cdot\mathbb{E}[X]\cdot\lambda} \leq e^{-(t+1)\frac{m}{u}} \leq e^{\frac{\varepsilon\ln m}{4e}}.$$

For large enough $C > 0$ we have $e^{-(4e)^{-1}\varepsilon\ln m} < 1/2$. Therefore with constant probability over $S \sim \mathcal{D}^m$, there are at least t elements from $\{x_1,\ldots,x_u\}$ that are not included in S. Assume we are given such an S. Let $x_{i_1},\ldots,x_{i_{t/16}}$ denote some $t/16$ elements that are not in S and consider the vector w having its $(u+1)$'st coordinate set to $\theta\sqrt{2}/R$, coordinates $i_j = -2\sqrt{2}\theta/R$ and remaining coordinates 0. Then $\|w\|_2 = \sqrt{\theta^2 2/R^2 + (t/16)8\theta^2/R^2} \leq \sqrt{1/8+1/2} < 1$. Notice that for all $x_i \in S$, we have $\langle w, x_i\rangle = (\theta\sqrt{2}/R)\cdot R/\sqrt{2} = \theta$. For an x_{i_j} we have $\langle w, x_{i_j}\rangle = (\theta\sqrt{2}/R)\cdot R/\sqrt{2} + (-2\sqrt{2}\theta/R)\cdot R/\sqrt{2} = \theta - 2\theta = -\theta$. Thus $\mathcal{L}_S^\theta(w) = 0$ while $\mathcal{L}_\mathcal{D}(w) = t/(16u) = \Omega(R^2\ln m/(\theta^2 m))$.

3.2. Large τ

In this section, we prove Lemma 12. Let $m \geq (R^2/\theta^2)^{1+\varepsilon}$ be the number of samples with $\varepsilon = 0.001$, and let $R^2\ln(m)/(\theta^2 m) < \tau \leq 1$. We construct a distribution $\mathcal{D}$ over $\mathbb{R}^{u+1}\times\{-1,+1\}$, where $u = R^2/(16\theta^2\tau)$. The distribution $\mathcal{D}$ gives a uniform random point among $\{x_1,\ldots,x_u\}$ where x_i has its $(u+1)$'st and i'th coordinate equal to $R/\sqrt{2}$ and the rest to 0. The label is always 1.

In our lower bound proof, we will find a vector w of the following form. Let $k = e^{-28}\tau u$, and for every subset $T \subseteq \{1,\ldots,u\}$ with $|T| = k$, let w_T be the vector where each coordinate i with $i \in T$ is set to $-1/\sqrt{2k}$, its $(u+1)$'st coordinate is set to $\theta\sqrt{2}/R$ and all remaining coordinates are set to 0. Then $\|w_T\|_2 = \sqrt{1/2 + 2\theta^2/R^2} \leq 1$, as $R > C\theta$ for some sufficiently large $C > 0$. In addition, for every $i \notin T$, $\langle x_i, w_T\rangle = \theta$ and for every $i \in T$ we have $\langle x_i, w_T\rangle = \theta - R/(2\sqrt{k}) \leq -\theta < 0$ if $i \in T$. Clearly for every such subset T, $\mathcal{L}_\mathcal{D}(w_T) = k/u = \tau/e^{28}$. What remains is to argue that with constant probability over S, there exists T where $\mathcal{L}_S^\theta(w_T)$ is significantly smaller than k/u, i.e. there is a large gap between $\mathcal{L}_\mathcal{D}(w_T)$ and $\mathcal{L}_S^\theta(w_T)$.

Fix some set S of m samples from $\mathcal{D}$, let b_i denote the number of times x_i is in the sample. Then for every T we have $\mathcal{L}_S^\theta(w_T) = (\sum_{i\in T} b_i)/m$. Let $T^* \subseteq \{1,\ldots,u\}$ be the set containing the k indices with smallest b_i. We will show that with good probability over the choice of S the k smallest values among $b_1,\ldots,b_u$ are small, and thus $(\sum_{i\in T^*} b_i)/m$ is small.

Consider first a fixed index i. For every $j \in [m]$ let c_j be the indicator for the event that the j'th element in the sample is x_i. Then $c_1,\ldots,c_m$ are independent indicators with success probability $p = 1/u$, and moreover, $b_i = \sum_{j\in[m]} c_i$. We will use the following reverse Chernoff bound to show that b_i is significantly smaller than its expectation m/u with reasonable probability.

Lemma 13. *[(Klein & Young, 2015)] For every $\sqrt{3/(mp)} < \delta < 1/2$,*

$$\Pr\left[\sum_j c_j \leq (1-\delta)mp\right] \geq e^{-9mp\delta^2}.$$

Now set

$$\delta = \sqrt{\ln(u/(2k))/(9m/u)}.$$

Since $u/(2k) = e^{28}\tau^{-1}/2 > e^{27}$ it follows that $\delta > \sqrt{\ln(e^{27})/9(m/u)} = \sqrt{3/(m/u)}$. We have assumed $\tau > R^2\ln(m)/(\theta^2 m)$, and thus $u = R^2/(16\theta^2\tau) <$

$m/(16\ln m)$. Therefore $\delta = \sqrt{\ln(u/(2k))/(9m/u)} \leq \sqrt{\ln(e^{28}\tau^{-1})/(9\cdot 16\ln m)} \leq 1/2$ for a large enough constant $C > 0$ such that $R > C\theta$. Hence we may use Lemma 13 to conclude that $\Pr[b_i \leq (1-\delta)m/u] \geq e^{-\ln(u/(2k))} = 2k/u$.

We will next show that with constant probability there are at least k indices i for which $b_i \leq (1-\delta)m/u$. Let B_i denote the indicator for the event $b_i \leq (1-\delta)m/u$. We will show that with probability at least $1/8$, $B := \sum_i B_i \geq k$. Note first that $\mathbb{E}[B] = \mathbb{E}[\sum_i B_i] = u\mathbb{E}[B_1] \geq 2k$. By the Paley-Zygmund inequality it follows that

$$\Pr[B \geq k] \geq \Pr[B \geq (1/2)\mathbb{E}[B]] \geq \frac{\mathbb{E}[B]^2}{4\mathbb{E}[B^2]} \qquad (17)$$

Consider now $\mathbb{E}[B^2] = \sum_{i,j}\mathbb{E}[B_iB_j]$. For $i \neq j$, we have that the events B_i and B_j are negatively correlated and thus $\mathbb{E}[B_iB_j] \leq \mathbb{E}[B_i]^2 = \mathbb{E}[B_1]^2$. For $i = j$ we have $\mathbb{E}[B_iB_i] = \mathbb{E}[B_i] = \mathbb{E}[B_1]$. Therefore we may bound $\mathbb{E}[B^2] \leq (u^2-u)\mathbb{E}[B_1]^2 + u\mathbb{E}[B_1] \leq \mathbb{E}[B]^2 + \mathbb{E}[B]$. Note that for a large enough $C > 0$, $\mathbb{E}[B] \geq 2k \geq 1$ and thus $\mathbb{E}[B] \leq \mathbb{E}[B]^2$ and we get that $\mathbb{E}[B^2] \leq 2\mathbb{E}[B]^2$. Plugging in (17), we conclude that $\Pr[B \geq k] \geq 1/8$, and hence with probability at least $1/8$ over the random set of samples S, it holds that $(\sum_{i\in T^*} b_i)/m \leq (k(1-\delta)m/u)/m = k(1-\delta)/u$. In this case, we have $\mathcal{L}_{\mathcal{D}}(w_{T^*}) - \mathcal{L}_S^{\theta}(w_{T^*}) \geq k\delta/u = \Omega((R^2/\theta^2)\sqrt{\ln(u/k)/(m/u)}) = \Omega(\sqrt{(R^2/\theta^2)\ln(\tau^{-1})\tau/m})$. Since $\tau = e^{28}k/u = e^{28}\mathcal{L}_{\mathcal{D}}(w_{T^*}) \geq e^{28}\mathcal{L}_S^{\theta}(w_{T^*})$ we have that $\mathcal{L}_S^{\theta}(w_{T^*}) \leq \tau/e^{28} \leq \tau$ which concludes the proof of Lemma 12.

ACKNOWLEDGMENTS

This work was supported by a Villum Young Investigator Grant and an AUFF Starting Grant.

References

Alon, N. and Klartag, B. Optimal compression of approximate inner products and dimension reduction. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pp. 639–650, 2017.

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. ISBN 052111862X, 9780521118620.

Bartlett, P. and Shawe-Taylor, J. Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods-Support Vector Learning*, pp. 43–54. MIT Press, Cambridge, MA, 1999.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92. ACM, 1992.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Dasgupta, S. and Gupta, A. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247 – 261, 1989. ISSN 0890-5401.

Gao, W. and Zhou, Z.-H. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

Grønlund, A., Kamma, L., Larsen, K. G., Mathiasen, A., and Nelson, J. Margin-based generalization lower bounds for boosted classifiers. In *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.

Janson, S. Tail bounds for sums of geometric and exponential variables. *Statistics & Probability Letters*, 135:1 – 6, 2018. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2017.11.017.

Johnson, W. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pp. 189–206. American Mathematical Society, 1984.

Klein, P. N. and Young, N. E. On the number of iterations for dantzig-wolfe optimization and packing-covering approximation algorithms. *SIAM J. Comput.*, 44(4):1154–1172, 2015.

McAllester, D. A. Simplified pac-bayesian margin bounds. In Schölkopf, B. and Warmuth, M. K. (eds.), *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pp. 203–215. Springer, 2003.

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5): 1651–1686, 1998.

Vapnik, V. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 1982. ISBN 0387907335.

Vapnik, V. and Chervonenkis, A. *On the uniform convergence of relative frequencies of events to their probabilities*, pp. 11–30. 01 2015. doi: 10.1007/978-3-319-21852-6_3.