A. Proof of Lemma 3.1

We here prove a lower bound on the number of support vectors to achieve generalization bounds of the form which we consider. Importantly, this result holds not only for linear classifiers, but also for randomized classifiers of a more general form than those by our sparsification procedure (Section 4).

Lemma 3.1. *Let $R, \mathcal{L}^*, \epsilon \geq 0$ be given, with $\mathcal{L}^* + \epsilon \leq 1/4$ and with R^2 being an integer. There exists a data distribution $\mathcal{D}$ and a reference vector u such that $\|u\| = R$, $\mathcal{L}_{hinge}(g_u) = \mathcal{L}^*$, and any w which satisfies:*

$$\mathcal{L}_{0/1}(g_w) \leq \mathcal{L}^* + \epsilon$$

must necessarily be supported on at least $R^2/2$ vectors. Furthermore, the claim also holds for randomized classification rules that predict 1 with probability $\psi(g_u(x))$ for some $\psi : \mathbb{R} \to [0,1]$.

Proof. We define $\mathcal{D}$ such that i is sampled uniformly at random from the set $\{1, \ldots, d\}$, with $d = R^2$, and the feature vector is taken to be $x = e_i$ (the ith standard unit basis vector) with corresponding label distributed according to $\Pr\{y = z\} = 1 - \mathcal{L}^*/2$. The value of $z \in \{\pm 1\}$ will be specified later. Choose $u_i = z$ for all i, so that $\|u\| = R$ and $\mathcal{L}_{hinge}(g_u) = \mathcal{L}^*$.

Take w to be a linear combination of $k < d/2 = R^2/2$ vectors. Then $g_w(x) = 0$ on any x which is not in its support set. Suppose that whenever $g_w(x_i) = 0$ the algorithm predicts the label 1 with probability $p \in [0,1]$ ($p = \psi(0)$ for a randomized classifier). If $p \geq 1/2$ we'll set $z = -1$, and if $p < 1/2$ we'll set $z = 1$. This implies that:

$$\mathcal{L}_{0/1}(g_w) \geq \frac{d-k}{2d} > \frac{1}{4} \geq \mathcal{L}^* + \epsilon$$

which concludes the proof. □

B. Compression Bound

We rely on the following compression bound (Theorem 2 of Shalev-Shwartz (2010)).

Theorem B.1. *Let k and n be fixed, with $n \geq 2k$, and let $A : \left(\mathbb{R}^d \times \{\pm 1\}\right)^k \to \mathcal{H}$ be a mapping which receives a list of k labeled training examples, and returns a classification vector $w \in \mathcal{H}$. Use $S \in [n]^k$ to denote a list of k training indices, and let w_S be the result of applying A to the training elements indexed by S. Finally, let $\ell : \mathbb{R} \to [0,1]$ be a loss function bounded below by 0 and above by 1, with $\mathcal{L}(g_w)$ and $\hat{\mathcal{L}}(g_w)$ the expected loss, and empirical loss on the training set, respectively. Then, with probability $1 - \delta$, for all S:*

$$\mathcal{L}(g_{w_S}) \leq \hat{\mathcal{L}}(g_{w_S}) + \sqrt{\frac{32\hat{\mathcal{L}}(g_{w_S})\left(k \log n + \log \frac{1}{\delta}\right)}{n}} + \frac{8\left(k \log n + \log \frac{1}{\delta}\right)}{n}$$

Proof. Consider, for some fixed δ', the probability that there exists a $S \subseteq \{1, \ldots, n\}$ of size k such that:

$$\mathcal{L}(g_{w_S}) \geq \hat{\mathcal{L}}_{\text{test}}(g_{w_S}) + \sqrt{\frac{2\hat{\mathcal{L}}_{\text{test}}(g_{w_S}) \log \frac{1}{\delta'}}{(n-k)}} + \frac{4 \log \frac{1}{\delta'}}{n-k}$$

where $\hat{\mathcal{L}}_{\text{test}}(g_w) = \frac{1}{n-k} \sum_{i \notin S} \ell(y_i g_w(x_i))$ is the empirical loss on the *complement* of S. It follows from Bernstein's inequality that, for a *particular* S, the above holds with probability at most δ'. By the union bound:

$$n^k \delta' \geq \Pr\left\{ \exists S \in [n]^k : \mathcal{L}(g_{w_S}) \geq \hat{\mathcal{L}}_{\text{test}}(g_{w_S}) + \sqrt{\frac{2\hat{\mathcal{L}}_{\text{test}}(g_{w_S}) \log \frac{1}{\delta'}}{(n-k)}} + \frac{4 \log \frac{1}{\delta'}}{n-k} \right\}$$

Let $\delta = n^k \delta'$. Notice that $(n-k)\hat{\mathcal{L}}_{\text{test}}(g_{w_S}) \leq n\hat{\mathcal{L}}(g_{w_S})$, so:

$$\delta \geq \Pr\left\{ \exists S \in [n]^k : \mathcal{L}(g_{w_S}) \geq \frac{n\hat{\mathcal{L}}(g_{w_S})}{n-k} + \sqrt{\frac{2n\hat{\mathcal{L}}(g_{w_S}) \log \frac{n^k}{\delta}}{(n-k)^2}} + \frac{4 \log \frac{n^k}{\delta}}{n-k} \right\}$$

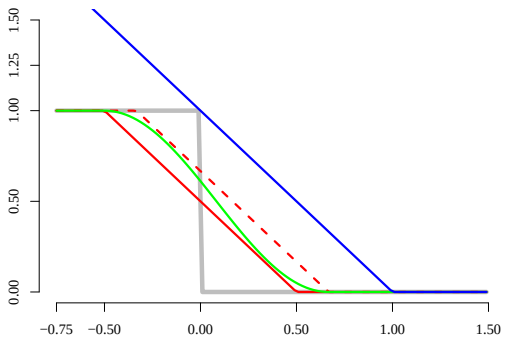

Figure 3. Illustration of the how our smooth loss relates to the slant and hinge losses. Our smooth loss (green) upper bounds the slant-loss, and lower bounds the slant-loss when shifted by 1/6, and the hinge-loss when shifted by 1/3.

Because $\hat{\mathcal{L}}\left(g_{w_S}\right) \leq 1$ and $k \leq n$, it follows that $k\hat{\mathcal{L}}\left(g_{w_S}\right) \leq 2n\log n$, and therefore that $\frac{k}{n-k}\hat{\mathcal{L}}\left(g_{w_S}\right) \leq \sqrt{\frac{2nk\hat{\mathcal{L}}\left(g_{w_S}\right)\log n}{(n-k)^2}} \leq \sqrt{\frac{2n\hat{\mathcal{L}}\left(g_{w_S}\right)\log\frac{n^k}{\delta}}{(n-k)^2}}$. Hence:

$$\delta \geq \Pr\left\{\exists S \in [n]^k : \mathcal{L}\left(g_{w_S}\right) \geq \hat{\mathcal{L}}\left(g_{w_S}\right) + \sqrt{\frac{8n\hat{\mathcal{L}}\left(g_{w_S}\right)\log\frac{n^k}{\delta}}{\left(n-k\right)^2}} + \frac{4\log\frac{n^k}{\delta}}{n-k}\right\}$$

Using the assumption that $n \geq 2k$ completes the proof. □

C. Concentration-based Analysis

In this section, we will prove a bound comparable to that of Theorem 4.4, but using a proof technique based on a smooth loss, rather than a compression bound. In order to accomplish this, we must first modify the objective of Problem 4.1 by adding a norm-constraint:

$$\begin{aligned} \text{minimize}: & f(\tilde{w}) = \max_{i:y_i\langle w,\Phi(x_i)\rangle>0}\left(h_i - y_i\left\langle\tilde{w},\Phi(x_i)\right\rangle\right) \\ \text{subject to}: & \|\tilde{w}\| \leq \|w\| \end{aligned} \tag{C.1}$$

Here, as before, $h_i = \min\left(1, y_i\left\langle w,\Phi(x_i)\right\rangle\right)$. Like Problem 4.1, this objective can be optimized using subgradient descent, although one must add a step in which the current iterate is projected onto the ball of radius $\|w\|$ after every iteration. Despite this change, an ϵ-suboptimal solution can still be found in $\|w\|^2/\epsilon^2$ iterations.

The concentration-based version of our main theorem follows:

Theorem C.1. *Let $R \in \mathbb{R}_+$ be fixed. With probability $1-\delta$ over the training sample, uniformly over all pairs $w, \tilde{w} \in \mathcal{H}$ such that $\|w\| \leq R$ and $\tilde{w}$ has objective function $f(\tilde{w}) \leq 1/3$ in Problem C.1:*

$$\mathcal{L}_{0/1}\left(\tilde{g}_{\tilde{w}}\right) \leq \hat{\mathcal{L}}_{hinge}\left(g_w\right) + O\left(\sqrt{\frac{\hat{\mathcal{L}}_{hinge}\left(g_w\right)R^2\log^3 n}{n}} + \sqrt{\frac{\hat{\mathcal{L}}_{hinge}\left(g_w\right)\log\frac{1}{\delta}}{n}} + \frac{R^2\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{n}\right)$$

Proof. Because our bound is based on a smooth loss, we begin by defining the bounded 4-smooth loss $\ell_{smooth}(z)$ to be 1 if $z < -1/2$, 0 if $z > 2/3$, and $1/2\left(1+\cos\left(\pi/2\left(1+1/7\left(12z-1\right)\right)\right)\right)$ otherwise. This function is illustrated in Figure C—notice that it upper-bounds the slant-loss, and lower-bounds the hinge loss even when shifted by 1/3. Applying Theorem 1 of Srebro et al. (2010) to this smooth loss yields that, with probability $1-\delta$, uniformly over all $\tilde{w}$ such that $\|\tilde{w}\| \leq R$:

$$\mathcal{L}_{smooth}\left(g_{\tilde{w}}\right) \leq \hat{\mathcal{L}}_{smooth}\left(g_{\tilde{w}}\right) + O\left(\sqrt{\frac{\hat{\mathcal{L}}_{smooth}\left(g_{\tilde{w}}\right)R^2\log^3 n}{n}} + \sqrt{\frac{\hat{\mathcal{L}}_{smooth}\left(g_{\tilde{w}}\right)\log\frac{1}{\delta}}{n}} + \frac{R^2\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{n}\right)$$

Just as the empirical slant-loss of a $\tilde{w}$ with $f(\tilde{w}) \leq 1/2$ is upper bounded by the empirical hinge loss of w, the empirical smooth loss of a $\tilde{w}$ with $f(\tilde{w}) \leq 1/3$ is upper-bounded by the same quantity. As was argued in the proof of Lemma 4.1, this follows directly from Problem C.1, and the definition of the smooth loss. Combining this with the facts that the slant-loss lower bounds the smooth loss, and that $\mathcal{L}_{slant}\left(g_{\tilde{w}}\right) = \mathcal{L}_{0/1}\left(\tilde{g}_{\tilde{w}}\right)$, completes the proof. □

It's worth pointing out that the addition of a norm-constraint to the objective function (Problem C.1) is only necessary because we want the theorem to apply to any $\tilde{w}$ with $f(\tilde{w}) \leq 1/3$. If we restrict ourselves to $\tilde{w}$ which are found using subgradient descent with the suggested step size and iteration count, then applying the triangle inequality to the sequence of steps yields that $\|\tilde{w}\| \leq O\left(\|w\|\right)$, and the above bound still holds (albeit with a different constant hidden inside the big-Oh notation).

D. Unregularized Bias (Alternative)

In Section 4.6, we discussed a simple extension of our algorithm to a SVM problem with an unregularized bias term, in which we took our sparse classifier $\tilde{w}, \tilde{b}$ to have the same bias as our target classifier w, b (i.e. $\tilde{b} = b$). In this section, we discuss an alternative, in which we optimize over $\tilde{b}$ during our subgradient descent procedure. The relevant optimization problem (analogous to Problem 4.1) is:

$$\begin{aligned} \text{minimize}: & f(\tilde{w}, \tilde{b}) = \max_{i: y_i \langle w, \Phi(x_i)\rangle > 0} \left(h_i - y_i\left(\langle \tilde{w}, \Phi(x_i)\rangle + \tilde{b}\right)\right) \\ \text{with}: & h_i = \min\left(1, y_i(\langle w, \Phi(x_i)\rangle + b)\right) \end{aligned} \tag{D.1}$$

A $1/2$-approximation may once more be found using subgradient descent. The difference is that, before finding a subgradient, we will implicitly optimize over $\tilde{b}$. It can be easily observed that the optimal $\tilde{b}$ will ensure that:

$$\max_{i: y_i > 0 \wedge \langle w, \Phi(x_i)\rangle > 0} \left(h_i - (\langle \tilde{w}, \Phi(x_i)\rangle + \tilde{b})\right) = \max_{i: y_i < 0 \wedge \langle w, \Phi(x_i)\rangle < 0} \left(h_i + (\langle \tilde{w}, \Phi(x_i)\rangle + \tilde{b})\right) \tag{D.2}$$

In other words, $\tilde{b}$ will be chosen such that the maximal violation among the set of positive examples will equal that among the negative examples. Hence, during optimization, we may find the most violating *pair* of one positive and one negative example, and then take a step on both elements. The resulting subgradient descent algorithm is:

1. Find the training indices $i_+ : y_i > 0 \wedge \langle w, \Phi(x_i)\rangle + b > 0$ and $i_- : y_i < 0 \wedge \langle w, \Phi(x_i)\rangle + b < 0$ which maximize $h_i - y_i \left\langle \tilde{w}^{(t-1)}, \Phi(x_i)\right\rangle$
2. Take the subgradient step $\tilde{w}^{(t)} \leftarrow \tilde{w}^{(t-1)} + \eta(\Phi(x_{i_+}) - \Phi(x_{i_-}))$.

Once optimization has completed, $\tilde{b}$ may be computed from Equation D.2. As before, this algorithm will find a $1/2$-approximation in $4\|w\|^2$ iterations.

E. Sample Complexity of SVM

In this appendix, we provide a brief proof of a claim based on Lemma D.1 of the appendix of Cotter et al. (2012b), which is the long version of Cotter et al. (2012a). This result, which follows almost immediately from Theorem 1 of Srebro et al. (2010), establishes the sample complexity bound claimed in Section 2.

Lemma E.1. (See Lemma D.1 of Cotter et al. (2012b)) *Let u be an arbitrary linear classifier, and suppose that we sample a training set of size n, with n given by the following equation, for parameters $\epsilon > 0$ and $\delta \in (0, 1)$:*

$$n = \tilde{O}\left(\left(\frac{\mathcal{L}_{hinge}\left(g_u\right) + \epsilon}{\epsilon}\right)\frac{\left(\|u\| + \log\frac{1}{\delta}\right)^2}{\epsilon}\right) \tag{E.1}$$

Let $\hat{w}^* = \underset{\|\hat{w}\| \leq \|u\|}{\operatorname{argmin}} \hat{\mathcal{L}}_{hinge}\left(\hat{w}\right)$. Then, with probability $1 - 2\delta$ over the i.i.d. training sample $x_i, y_i : i \in \{1, \ldots, n\}$, we have that $\mathcal{L}_{0/1}\left(g_{\hat{w}^*}\right) \leq \mathcal{L}_{hinge}\left(g_u\right) + 2\epsilon$.

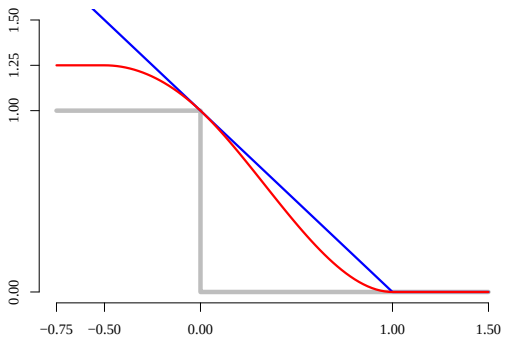

Figure 4. Plot of a smooth and bounded function (red) which upper bounds the 0/1 loss and lower bounds the hinge loss.

To prove this, we will first prove two helper lemmas: Lemma E.2 is a direct application of Theorem 1 of Srebro et al. (2010) to a smooth function which is intermediate between the 0/1 and hinge losses (this is similar to Theorem 5 of Srebro et al. (2010)); Lemma E.3 analyzes the empirical error of a single hypothesis by a direct application of Bernstein's inequality. Combining these two lemmas (Section E.2) then gives the claimed result.

E.1. Helper Lemmas

Lemma E.2. *Suppose that we sample a training set of size n, with n given by the following equation, for parameters $L, B, \epsilon > 0$ and $\delta \in (0,1)$:*

$$n = \tilde{O}\left(\left(\frac{L}{\epsilon}\right)\frac{\left(B + \sqrt{\log\frac{1}{\delta}}\right)^2}{\epsilon}\right) \tag{E.2}$$

Then, with probability $1-\delta$ over the i.i.d. training sample $x_i, y_i : i \in \{1,\ldots,n\}$, uniformly for all linear classifiers w satisfying:

$$\|w\| \leq B, \quad \hat{\mathcal{L}}_{hinge}\left(g_w\right) \leq L \tag{E.3}$$

we have that $\mathcal{L}_{0/1}\left(g_w\right) \leq L + \epsilon$.

Proof. For a *smooth* loss function, Theorem 1 of Srebro et al. (2010) bounds the expected loss in terms of the empirical loss, plus a factor depending on (among other things) the sample size. Neither the 0/1 nor the hinge losses are smooth, so we will define a bounded and smooth loss function which upper bounds the 0/1 loss and lower-bounds the hinge loss. The particular function which we use doesn't matter, since its smoothness parameter and upper bound will ultimately be absorbed into the big-Oh notation—all that is needed is the *existence* of such a function. One such is:

$$\phi\left(x\right) = \begin{cases} 5/4 & \cdots & x < -1/2 \\ -x^2 - x + 1 & \cdots & -1/2 \leq x < 0 \\ x^3 - x^2 - x + 1 & \cdots & 0 \leq x < 1 \\ 0 & \cdots & x \geq 1 \end{cases}$$

This function, illustrated in Figure E.1, is 4-smooth and $5/4$-bounded. If we define $\mathcal{L}_\phi\left(g_w\right)$ and $\hat{\mathcal{L}}_\phi\left(g_w\right)$ as the expected and empirical ϕ-losses, respectively, then the aforementioned theorem gives that, with probability $1-\delta$ uniformly over all w such that $\|w\| \leq B$:

$$\mathcal{L}_\phi\left(g_w\right) \leq \hat{\mathcal{L}}_\phi\left(g_w\right) + O\left(\frac{B^2\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{n} + \sqrt{\hat{\mathcal{L}}_\phi\left(g_w\right)}\left(\sqrt{\frac{B^2\log^3 n}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)\right)$$

Because ϕ is lower-bounded by the 0/1 loss and upper-bounded by the hinge loss, we may replace $\mathcal{L}_\phi(g_w)$ with $\mathcal{L}_{0/1}(g_w)$ on the LHS of the above bound, and $\hat{\mathcal{L}}_\phi(g_w)$ with L on the RHS. Setting the big-Oh expression to ϵ and solving for n then gives the desired result. □

Lemma E.3. *Let u be an arbitrary linear classifier, and suppose that we sample a training set of size n, with n given by the following equation, for parameters $\epsilon > 0$ and $\delta \in (0,1)$:*

$$n = 2\left(\frac{\mathcal{L}_{hinge}(g_u) + \epsilon}{\epsilon}\right)\frac{\|u\| \log \frac{1}{\delta}}{\epsilon} \tag{E.4}$$

Then, with probability $1-\delta$ over the i.i.d. training sample $x_i, y_i : i \in \{1,\ldots,n\}$, we have that $\hat{\mathcal{L}}_{hinge}(g_u) \leq \mathcal{L}_{hinge}(g_u) + \epsilon$.

Proof. The hinge loss is upper-bounded by $\|u\|$ (by assumption, $\|x\| \leq 1$ with probability 1), from which it follows that $\mathrm{Var}_{x,y}(\ell(y\langle u, x\rangle)) \leq \|u\| \mathcal{L}_{hinge}(g_u)$. Hence, by Bernstein's inequality:

$$\begin{aligned}\Pr\left\{\hat{\mathcal{L}}_{hinge}(g_u) > \mathcal{L}_{hinge}(g_u) + \epsilon\right\} &\leq \exp\left(-\frac{n}{\|u\|}\left(\frac{\epsilon^2/2}{\mathcal{L}_{hinge}(g_u) + \epsilon/3}\right)\right) \\ &\leq \exp\left(-\frac{n}{2\|u\|}\left(\frac{\epsilon^2}{\mathcal{L}_{hinge}(g_u) + \epsilon}\right)\right)\end{aligned}$$

Setting the LHS to δ and solving for n gives the desired result. □

E.2. Proof of Lemma E.1

Proof. Lemma E.3 gives that $\hat{\mathcal{L}}_{hinge}(g_u) \leq \mathcal{L}_{hinge}(g_u) + \epsilon$ provided that Equation E.4 is satisfied. Take $L = \mathcal{L}_{hinge}(g_u) + \epsilon$ and $B = \|u\|$, and observe that $\hat{w}^*$ satisfies Equation E.3 because $\hat{\mathcal{L}}_{hinge}(g_{\hat{w}^*}) \leq \hat{\mathcal{L}}_{hinge}(g_u) \leq \mathcal{L}_{hinge}(g_u) + \epsilon = L$ and $\|\hat{w}^*\| \leq \|u\| = B$. Therefore, Lemma E.2 gives that $\mathcal{L}_{0/1}(g_{\hat{w}^*}) \leq \mathcal{L}_{hinge}(g_u) + 2\epsilon$, provided that Equation E.2 is also satisfied. Equation E.1 is what results from combining these two bounds and simplifying. Each lemma holds with probability $1-\delta$, so this result holds with probability $1-2\delta$. □