Appendix A Rademacher Complexity and Generalization Bounds

For completeness, we provide the proof of Theorem 1, following the approach of Sabato et al. (2013). We then extend it to prove Theorem 2.

To derive the sample complexity of our hypothesis classes $\mathcal{H}_{2,0}$ and $\mathcal{H}_{1,0}$, we will use the Rademacher complexity. Let $\mathcal{Z}$ be some domain. The empirical Rademacher complexity of a class of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ with respect to a set $S = \{\mathbf{z}_i\} \subseteq \mathcal{Z}$, for $1 \le i \le m$ is

$$\mathcal{R}(\mathcal{F}, S) = \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\boldsymbol{\sigma}_i f(\mathbf{z}_i)\right|\right], \quad (1)$$

where $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_m)$ are m independent uniform $\{\pm 1\}$-valued variables. The empirical Gaussian complexity $\mathcal{G}(\mathcal{F}, S)$ is similarly defined with the entries of $\boldsymbol{\sigma}$ being m independent standard Gaussian random variables. The average Rademacher complexity of $\mathcal{F}$ with respect to a distribution $\mathcal{D}$ over $\mathcal{Z}$ and a sample size m is

$$\mathcal{R}_m(\mathcal{F}, \mathcal{D}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{R}(\mathcal{F}, S)]. \quad (2)$$

Consider a hypothesis class $\mathcal{H}$ and a loss function ℓ. For a hypothesis $h \in \mathcal{H}$, let $h_\ell : \mathcal{X} \times \{\pm 1\} \mapsto \mathbb{R}$ be defined as $h_\ell(x, y) = \ell(y, h(x))$. The resulting function class $\mathcal{H}_\ell$ is $\mathcal{H}_\ell = \{h_\ell | h \in \mathcal{H}\}$. Assume that the range of $\mathcal{H}_\ell$ is $[0, 1]$. Then, from Mendelson (2002), for any $\delta \in (0, 1)$, with probability $1 - \delta$, every $h \in \mathcal{H}$ satisfies that

$$\ell(h, \mathcal{D}) \le \ell(h, S) + 2\mathcal{R}_m(\mathcal{H}_\ell, \mathcal{D}) + \sqrt{\frac{8\ln(2/\delta)}{m}}. \quad (3)$$

Denote the class of ramp-loss functions applied to the hypothesis class $\mathcal{H}$ by

$$\mathrm{RAMP} \circ \mathcal{H} = \{(\mathbf{x}, y) \mapsto \mathrm{ramp}(h, (\mathbf{x}, y)) | h \in \mathcal{H}\}. \quad (4)$$

In addition to the empirical Rademacher complexity, we will use the notion of L_2 covering numbers, defined as follows.

Definition 1. *An ϵ-cover of a subset A of a pseudometric space (S,d) is a set A' such that for each $a \in A$, there exists $a' \in A'$ such that $d(a, a') \le \epsilon$. The ϵ-covering number of A is:*

$$\mathcal{N}(\epsilon, A, d) = \min\{|A'| : A' \text{ is an } \epsilon\text{-cover of } A\}. \quad (5)$$

Appendix B Sample Complexity

We now provide direct proofs for Theorems 1 and 2. In order to bound the Rademacher complexity of the class $\mathrm{RAMP}_\gamma \circ \mathcal{H}_{2,0}$, we will first bound its covering number. To do so, we will express the functions in this class as sums of two functions with respect to $\mathbf{w}_a$ and $\mathbf{w}_b$. We require the three following lemmas from Sabato et al. (2013), reported below for completeness.

Lemma 1 (Sabato et al. (2013), Lemma 8)**.** *Let $(\mathcal{X}, \|\cdot\|_\circ)$ be a normed space. Let $\mathcal{F} \subseteq \mathcal{X}$ be a set, and let $\mathcal{G} : \mathcal{X} \mapsto 2^{\mathcal{X}}$ be a mapping from objects in $\mathcal{X}$ to sets of objects in $\mathcal{X}$. Assume that $\mathcal{G}$ is c-Lipschitz with respect to the Hausdorff distance Δ_H on sets, that is assume that*

$$\forall f_1, f_2 \in \mathcal{X}, \Delta_H(\mathcal{G}(f_1), \mathcal{G}(f_2)) \le c\|f_1 - f_2\|_\circ, \quad (6)$$

where $\Delta_H(\mathcal{G}_1, \mathcal{G}_2) = \sup_{g_1\in\mathcal{G}_1}\inf_{g_2\in\mathcal{G}_2}\|g_1 - g_2\|_\circ$. Let $\mathcal{F}_\mathcal{G} = \{f + g | f \in \mathcal{F}, g \in \mathcal{G}(f)\}$. Then,

$$\mathcal{N}(\epsilon, \mathcal{F}_\mathcal{G}, \circ) \le \mathcal{N}(\epsilon/(2+c), \mathcal{F}, \circ) \cdot \sup_{f\in\mathcal{F}}\mathcal{N}(\epsilon/(2+c), \mathcal{G}(f), \circ). \quad (7)$$

Lemma 2 (Sabato et al. (2013), Lemma 9)**.** *Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a function and let $Z \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class over some domain $\mathcal{X}$. Let $\mathcal{G} : \mathbb{R}^{\mathcal{X}} \mapsto 2^{\mathbb{R}^{\mathcal{X}}}$ be the mapping defined by*

$$\mathcal{G}(f) \triangleq \{x \mapsto [\![f(x) + z(x)]\!] - f(x) | z \in Z\}. \quad (8)$$

Then, $\mathcal{G}$ is 1-Lipschitz with respect to the Hausdorff distance.

Lemma 3 (Sabato et al. (2013), Lemma 10)**.** *Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a function and let $Z \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class over some domain $\mathcal{X}$. Let $\mathcal{G}(f)$ be defined as in (8). Then, the pseudo-dimension of $\mathcal{G}(f)$ is at most the pseudo-dimension of Z.*

Our next lemma requires the definition of the notions of pseudo-shattering (Pollard, 2012) and pseudo-dimension.

Definition 2. *Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \mapsto \mathbb{R}$, and $\gamma > 0$. The set $\{\mathbf{x}_a, \ldots, \mathbf{x}_m\} \subseteq \mathcal{X}$ is pseudo-shattered by $\mathcal{F}$ with the witness $r \in \mathbb{R}^m$ if for all $y \in \{\pm 1\}^m$ there is an $f \in \mathcal{F}$ such that $\forall i \in [1, \ldots, m]$, $y[i](f(\mathbf{x}_i) - r[i]) > 0$.*

The pseudo-dimension *pdim* of a hypothesis class is the size of the largest set that is pseudo-shattered by this class.

Lemma 4. *Let $\mathcal{H} = \{\mathbf{x} \mapsto \langle\mathbf{w}, \mathbf{x}\rangle | \|\mathbf{w}\|_0 \le k\}$. Then,*

$$pdim(\mathcal{H}) = O(k \log d). \quad (9)$$

Equipped with these lemmas, we can now derive an upper bound on the Rademacher complexity of $\mathrm{RAMP} \circ \mathcal{H}_{2,0}$ in the following theorem. Theorem 1 then follows directly from Proposition 1.

Theorem 1. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Assume that all samples are such that $\|\mathbf{x}_i\|_2 \leq 1$. Then,*

$$\mathcal{R}(RAMP \circ \mathcal{H}_{2,0}, \mathcal{D}) \leq \sqrt{\frac{O(k \log d + B^2 \log^2(m))}{m}}. \tag{10}$$

Proof. In this proof, all absolute constants are assumed to be positive and are denoted by C or C_i for some integer i. Their values may change from line to line or even within the same line.

Note that

$$\text{ramp}(h, \mathbf{x}, y) = [\![1 - y\langle \mathbf{w}, \mathbf{x}\rangle]\!] = 1 - [\![y\langle \mathbf{w}, \mathbf{x}\rangle]\!] \tag{11}$$

Shifting by a constant and negating do not change the covering number of a function class. Therefore, $\mathcal{N}(\epsilon, \text{RAMP} \circ \mathcal{H}_{2,0}, L_2(S))$ is equal to the covering number of $\{(\mathbf{x}, y) \mapsto [\![y\langle \mathbf{w}_a + \mathbf{w}_b, \mathbf{x}\rangle/]\!] | \|\mathbf{w}_a\|_2 \leq B, \|\mathbf{w}_b\|_0 \leq k\}$.

Define

$$\mathcal{F} = \{\mathbf{x} \mapsto y\langle \mathbf{w}_a, \mathbf{x}\rangle | \|\mathbf{w}_a\|_2 \leq B\}. \tag{12}$$

Let $\mathcal{G} : \mathbb{R}^{\mathbb{R}^d} \mapsto 2^{\mathbb{R}^{\mathbb{R}^d}}$ be the mapping defined by:

$$\mathcal{G}(f) = \{\mathbf{x} \mapsto [\![f(\mathbf{x}) + y\langle \mathbf{w}_b, \mathbf{x}\rangle]\!] - f(\mathbf{x}) | \|\mathbf{w}_b\|_0 \leq k\}. \tag{13}$$

From Lemma 2, $\mathcal{G}$ is 1-Lipschitz with respect to the Hausdorff distance. Clearly, $\mathcal{F}_{\mathcal{G}} = \{f + g | f \in \mathcal{F}, g \in \mathcal{G}(f)\} = \text{RAMP} \circ \mathcal{H}_{2,0}$. Thus, from Lemma 1, it holds that

$$\begin{aligned}&\mathcal{N}(\epsilon, \text{RAMP} \circ \mathcal{H}_{2,0}, L_2(S)) \leq \\ &\quad \mathcal{N}(\epsilon/3, \mathcal{F}, L_2(S)) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\epsilon/3, \mathcal{G}(f), L_2(S)).\end{aligned} \tag{14}$$

We now proceed to bound the two covering numbers on the right-hand side. First, consider $\mathcal{N}(\epsilon/3, \mathcal{G}(f), L_2(S))$. From Lemma 3, the pseudo-dimension of $\mathcal{G}(f)$ is the same as the pseudo-dimension of $\{\mathbf{x} \mapsto y\langle \mathbf{w}_b, \mathbf{x}\rangle | \|\mathbf{w}_b\|_0 \leq k\}$, which is given by Lemma 4. The L_2 covering number of $\mathcal{G}(f)$ may then be bounded by its pseudo-dimension as follows (Bartlett, 2006):

$$\mathcal{N}(\epsilon/3, \mathcal{G}(f), L_2(S)) \leq 2\left(\frac{36e}{\epsilon^2}\right)^{k \log d}. \tag{15}$$

Second, consider $\mathcal{N}(\epsilon/3, \mathcal{F}, L_2(S))$. From Sudakov's minoration theorem (Sudakov, 1971; Ledoux and Talagrand, 1991),

$$\ln \mathcal{N}(\epsilon/3, \mathcal{F}, L_2(S)) \leq \frac{C}{m\epsilon^2} \mathbb{E}_s^2[\sup_{f \in \mathcal{F}} \sum_{i=1}^m s_i f(\mathbf{x}_i)], \tag{16}$$

where s_i are independent standard normal variables. The right-hand side can be bounded as follows:

$$\begin{aligned}\mathbb{E}[\sup_{f \in \mathcal{F}} \sum_{i=1}^m s_i f(\mathbf{x}_i)] &= \mathbb{E}_s\left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq B} y\langle \mathbf{w}, \sum_{i=1}^m s_i \mathbf{x}_i\rangle\right] \\ &\leq B\mathbb{E}_s\left[\sqrt{\|\sum_{i=1}^m s_i \mathbf{x}_i]\|_2^2}\right] \\ &\leq B\sqrt{\mathbb{E}_s\left[\|\sum_{i=1}^m s_i \mathbf{x}_i\|_2^2\right]} \\ &= B\sqrt{\mathbb{E}_s\left[\|\sum_{i=1}^m \mathbf{x}_i\|_2^2\right]} \\ &\leq B\sqrt{m},\end{aligned}$$

where we used Jensen's inequality. Therefore, we have

$$\ln \mathcal{N}(\epsilon/3, \mathcal{F}, L_2(S)) \leq \frac{CB^2}{\epsilon^2}. \tag{17}$$

Substituting (15) and (17) in (14) and adjusting constants, we get

$$\begin{aligned}&\ln \mathcal{N}(\epsilon, \text{RAMP} \circ \mathcal{H}_{2,0}, L_2(S)) \leq \\ &\qquad C_1\left(1 + k \log d \ln \frac{C_2}{\epsilon} + \frac{B^2}{\epsilon^2}\right).\end{aligned} \tag{18}$$

We can now bound the Rademacher complexity of $\text{RAMP} \circ \mathcal{H}_{2,0}$ by its L_2 covering numbers. From Mendelson (2002), it holds that, for any monotone sequence $\{\epsilon_i\}$ decreasing to 0 such that $\epsilon_0 = 1$,

$$\begin{aligned}&\sqrt{m}\mathcal{R}(\text{RAMP} \circ \mathcal{H}_{2,0}, S) \\ \leq\ & C_1 \sum_{i=1}^N \epsilon_{i-1}\sqrt{\ln \mathcal{N}(\epsilon, \text{RAMP} \circ \mathcal{H}_{2,0}, L_2(S))} + 2\epsilon_N\sqrt{m} \\ \leq\ & C_1 \sum_{i=1}^N \epsilon_{i-1}\sqrt{1 + k \log d \ln \frac{C_2}{\epsilon} + \frac{B^2}{\epsilon^2}} + 2\epsilon_N\sqrt{m} \\ \leq\ & C_1 \sum_{i=1}^N \epsilon_{i-1}\left(1 + \sqrt{k \log d \ln \frac{C_2}{\epsilon}} + \frac{B}{\epsilon}\right) + 2\epsilon_N\sqrt{m},\end{aligned}$$

where we substituted (18). Let $\epsilon_i = 2^{-i}$. We obtain

$$\begin{aligned}&\sqrt{m}\mathcal{R}(\text{RAMP} \circ \mathcal{H}_{2,0}, S) \leq \\ &\qquad C\left(1 + \sqrt{k \log d} + NB\right) + 2^{-N+1}\sqrt{m}.\end{aligned} \tag{19}$$

Setting $N = \log(2m)$, we have

$$\mathcal{R}(\text{RAMP} \circ \mathcal{H}_{2,0}, S) \leq \frac{C}{\sqrt{m}}\left(1 + \sqrt{k \log d} + B\log(2m)\right). \tag{20}$$

Taking expectation over both sides yields

$$\begin{aligned}\mathcal{R}(\text{RAMP} \circ \mathcal{H}_{2,0}, \mathcal{D}) &\leq \frac{C}{\sqrt{m}}\left(1 + \sqrt{k \log d} + B\log(2m)\right) \\ &\leq \sqrt{\frac{\mathcal{O}(k \log d + B^2 log^2(2m))}{m}}.\end{aligned} \tag{21}$$

□

We can prove similarly the following theorem for the hypothesis class $\mathcal{H}_{1,0}$.

Theorem 2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Assume that all samples are such that $\|\mathbf{x}_i\|_2 \leq 1$. Then,*

$$\mathcal{R}(RAMP \circ \mathcal{H}_{1,0}, \mathcal{D}) \leq \sqrt{\frac{O(k \log d + B^2 \log d \log^2(m))}{m}}. \tag{22}$$

The proof is similar to that of Theorem 1 and is thus omitted here. In this case, (17) becomes

$$\ln \mathcal{N}(\epsilon/3, \mathcal{F}, L_2(S)) \leq \frac{CB^2 \log d}{\epsilon^2}, \tag{23}$$

using the following lemma, adapted from Lemma 19 in Bartlett and Mendelson (2002).

Lemma 5. *Let $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\|_1$. Define*

$$\mathcal{F}_1 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle | \|\mathbf{w}\|_1 \leq B\}. \tag{24}$$

Then, we have

$$\mathcal{G}(\mathcal{F}, S) \leq CB\sqrt{\frac{\log d}{m}}, \tag{25}$$

for some $C > 0$, where $\mathcal{G}(\mathcal{F}, S)$ is the empirical Gaussian complexity, defined below (1).

References

Peter Bartlett. Lecture notes. *https://people.eecs.berkeley.edu/ bartlett/courses/281b-sp06/lecture25.ps*, 2006.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Springer, 1991.

Shahar Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE transactions on Information Theory*, 48(1):251–263, 2002.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

S. Sabato, N. Srebro, and Naftali Tishby. Distribution-dependent sample complexity of large margin learning. *The Journal of Machine Learning Research*, 14 (1):2119–2149, 2013.

Vladimir N. Sudakov. Gaussian random processes and solid angle measures in Hilbert space. *Doklady Akademii Nauk SSSR*, 197(1):43, 1971.