Assignment 5 - Solutions

00103335: Deep Learning and Reinforcement Learning

Fall 2026

Problem 1. VC dimension

(a) Unions of intervals. For the lower bound, take the $2k$ points $1, 2, \ldots, 2k$. For any binary labeling, partition the positively labeled points into maximal consecutive runs. There are at most k such runs: s runs would require at least s positive points and $s-1$ separating negative points, so $2s-1 \le 2k$ and hence $s \le k$. For each run, take the closed interval from its first point to its last point. The union of these intervals has exactly the requested labels. If there are no positive points, use the empty set. Thus these $2k$ points are shattered.

For the upper bound, take any $2k+1$ distinct points and order them as $x_1 < \cdots < x_{2k+1}$. Label the odd-indexed points 1 and the even-indexed points 0. An interval containing two odd-indexed points would also contain a negatively labeled point between them. Consequently, each interval can contain at most one of the $k+1$ positive points. At least $k+1$ intervals would be needed, so this labeling cannot be realized by $\mathcal{I}_k$. Therefore

$$\boxed{\mathrm{VCdim}(\mathcal{I}_k) = 2k.}$$

(b) A one-parameter family. First consider the points $(0, 1, 3/2)$, in that order. The following choices realize all eight label vectors:

Label vector	000	001	010	011	100	101	110	111
α	5/2	5/4	1/4	3/4	$-1/4$	$-3/4$	0	-1

For example, $\alpha = -3/4$ gives $I_\alpha = [-3/4, 1/4] \cup [5/4, \infty)$ and hence the labels 101. This proves the lower bound 3.

For an arbitrary fixed sample point x_i, membership is equivalent to

$$x_i \in I_\alpha \quad \Longleftrightarrow \quad \alpha \in (-\infty, x_i - 2] \cup [x_i - 1, x_i].$$

Thus, as α increases, this coordinate's label follows $1 \to 0 \to 1 \to 0$, with exactly three changes. In particular, the labels at the three critical values are

$$\mathbf{1}_{I_{x_i-2}}(x_i) = 1, \qquad \mathbf{1}_{I_{x_i-1}}(x_i) = 1, \qquad \mathbf{1}_{I_{x_i}}(x_i) = 1.$$

The first and third changes occur immediately after their critical values; the middle change occurs at its critical value.

For m sample points, list all critical singleton points and all open intervals they cut out of $\mathbb{R}$, including the two unbounded intervals, in increasing order. The label vector is constant on each open interval. Along this list, each coordinate changes exactly three times, even if critical values from different coordinates coincide. Every change of the full label vector requires a change in at least one coordinate. Hence there are at most $3m$ changes of the full vector and at most $3m+1$ distinct label vectors. This argument includes the vectors attained exactly at critical values, which can differ from those on both adjacent intervals. For $m = 4$, the number of label vectors is therefore at most $13 < 16$. No four-point set is shattered, and

$$\boxed{\mathrm{VCdim}(\{I_\alpha : \alpha \in \mathbb{R}\}) = 3.}$$

(c) Closed balls. Lift $x \in \mathbb{R}^p$ to $\phi(x) = (x, \|x\|_2^2) \in \mathbb{R}^{p+1}$. The ball inequality is equivalent to an affine halfspace inequality:

$$\|x - c\|_2^2 \le r^2 \quad \Longleftrightarrow \quad (2c, -1)^\top \phi(x) + r^2 - \|c\|_2^2 \ge 0.$$

Thus every labeling realized by balls is realized by affine halfspaces on the lifted points. Affine halfspaces in $\mathbb{R}^{p+1}$ have VC dimension $p+2$, giving the desired upper bound. For completeness, the needed upper bound also follows directly from part (d), proved below: all functions

$$x \longmapsto r^2 - \|x - c\|_2^2$$

belong to the vector space spanned by $1, x_1, \ldots, x_p, \|x\|_2^2$, whose dimension is at most $p+2$. Their nonnegative level sets therefore have VC dimension at most $p+2$. In particular,

$$\boxed{\mathrm{VCdim}(\mathcal{B}_p) \le p + 2.}$$

(d) A vector space of functions. Fix any $r+1$ distinct points $x_1, \ldots, x_{r+1}$. The linear evaluation map

$$E : \mathcal{F} \longrightarrow \mathbb{R}^{r+1}, \qquad E(f) = (f(x_1), \ldots, f(x_{r+1}))$$

has rank at most r. Its image is therefore a proper linear subspace of $\mathbb{R}^{r+1}$. There is a nonzero vector $a = (a_1, \ldots, a_{r+1})$ orthogonal to this image, so

$$\sum_{i=1}^{r+1} a_i f(x_i) = 0 \qquad \text{for every } f \in \mathcal{F}.$$

Replacing a by $-a$ if necessary, assume that at least one a_i is positive. Request label 0 at every point with $a_i > 0$ and label 1 at every point with $a_i \le 0$. If some f realized these labels, the convention $\mathbf{1}_{\{f \ge 0\}}$ would imply

$$f(x_i) < 0 \quad (a_i > 0), \qquad f(x_i) \ge 0 \quad (a_i \le 0).$$

Every product $a_i f(x_i)$ would be nonpositive, and at least one would be strictly negative. Their sum would be negative, contradicting the displayed identity. Thus no $r+1$ points can be shattered, and

$$\boxed{\mathrm{VCdim}\big(\{\{x : f(x) \ge 0\} : f \in \mathcal{F}\}\big) \le r.}$$

The use of label 0 above is what gives a strict inequality despite the inclusion of zero in the positive class.

(e) Sine classifiers.

(i) Suppose $x \neq 0$. We show that the labeling $(1, 0, 0, 0)$ on $(x, 2x, 3x, 4x)$ is impossible. If it were realized, put $\theta = \omega x \pmod{2\pi}$ with $0 \le \theta < 2\pi$. The first two labels would require

$$\sin\theta \ge 0, \qquad \sin(2\theta) < 0.$$

The first condition gives $\theta \in [0, \pi]$, and the second then forces $\theta \in (\pi/2, \pi)$. On this interval, the third label $\sin(3\theta) < 0$ further forces

$$\frac{\pi}{2} < \theta < \frac{2\pi}{3}.$$

It follows that $2\pi < 4\theta < 8\pi/3$, where $\sin(4\theta) > 0$. This contradicts the fourth label 0. The intervals are strict because a sine value of zero receives label 1. The restricted domain has four points and is not shattered, so its VC dimension is at most 3.

For $x = 0$, the restricted domain is $A_0 = \{0\}$ and $h_\omega(0) = 1$ for every ω. The singleton cannot be shattered, whereas the empty set is shattered. Its VC dimension is therefore 0.

(ii) Fix $m \geq 1$ and any desired labels $y_1, \ldots, y_m$ on the distinct points $x_i = 2^{-i}$. Set

$$b_i = 1 - y_i \in \{0, 1\}, \qquad N = \sum_{i=1}^{m} b_i 2^{i-1}, \qquad \omega = 2\pi(N + \tfrac{1}{2}).$$

For each i, write $N = a_i 2^i + r_i$, where

$$r_i = N \bmod 2^i = \sum_{j=1}^{i} b_j 2^{j-1}, \qquad 0 \leq r_i < 2^i.$$

Consequently,

$$\omega x_i = 2\pi a_i + 2\pi \frac{r_i + 1/2}{2^i}.$$

If $b_i = 0$, then $r_i < 2^{i-1}$ and $0 < (r_i + 1/2)/2^i < 1/2$, so $\sin(\omega x_i) > 0$. If $b_i = 1$, then $r_i \geq 2^{i-1}$ and $1/2 < (r_i + 1/2)/2^i < 1$, so $\sin(\omega x_i) < 0$. In both cases $h_\omega(x_i) = y_i$. The added $1/2$ ensures that none of the sample points gives a zero sine value. Thus $\{2^{-i} : 1 \leq i \leq m\}$ is shattered for every m, proving

$$\boxed{\mathrm{VCdim}(\mathcal{H}_{\sin}) = \infty.}$$

Problem 2. Union of hypothesis classes

(a) For a fixed tuple $S = (x_1, \ldots, x_m)$, write

$$\mathcal{A}_S = \{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{A}\},$$

and define $\mathcal{B}_S$ and $\mathcal{C}_S$ in the same way. Since $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$,

$$\mathcal{C}_S = (\mathcal{A}_S) \cup (\mathcal{B}_S), \qquad |\mathcal{C}_S| \leq |\mathcal{A}_S| + |\mathcal{B}_S| \leq \Pi_{\mathcal{A}}(m) + \Pi_{\mathcal{B}}(m).$$

Taking the supremum over $S \in \mathcal{X}^m$ gives

$$\boxed{\Pi_{\mathcal{C}}(m) \leq \Pi_{\mathcal{A}}(m) + \Pi_{\mathcal{B}}(m).}$$

(b) Fix $m \geq d_A + d_B + 2$. By part (a), Sauer–Shelah's lemma, and binomial symmetry,

$$\begin{aligned}
\Pi_{\mathcal{C}}(m) &\leq \sum_{j=0}^{d_A} \binom{m}{j} + \sum_{j=0}^{d_B} \binom{m}{j} \\
&= \sum_{j=0}^{d_A} \binom{m}{j} + \sum_{j=m-d_B}^{m} \binom{m}{j} \\
&= 2^m - \sum_{j=d_A+1}^{m-d_B-1} \binom{m}{j}.
\end{aligned}$$

The last equality uses $m - d_B \geq d_A + 2$: the two included index ranges are disjoint, and the omitted middle range is nonempty. In particular, its term at $j = d_A + 1$ is strictly positive, so

$$\Pi_{\mathcal{C}}(m) \leq 2^m - \binom{m}{d_A + 1} < 2^m.$$

Hence no set of size $d_A + d_B + 2$ can be shattered. We conclude that

$$\boxed{\mathrm{VCdim}(\mathcal{A} \cup \mathcal{B}) \leq d_A + d_B + 1.}$$

Problem 3. Rademacher identities

Fix a sample $S = (x_1, \dots, x_m)$ and write

$$Z_\epsilon(h) := \frac{1}{m} \sum_{i=1}^{m} \epsilon_i h(x_i).$$

Throughout, $\widehat{\mathcal{R}}_S(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} Z_\epsilon(h)$ uses the convention without an absolute value inside the supremum.

(a) If $\alpha > 0$, taking the positive constant outside the supremum gives

$$\widehat{\mathcal{R}}_S(\alpha\mathcal{H}) = \alpha\, \widehat{\mathcal{R}}_S(\mathcal{H}).$$

If $\alpha < 0$, use the fact that $-\epsilon$ has the same law as ϵ:

$$\widehat{\mathcal{R}}_S(\alpha\mathcal{H}) = |\alpha|\, \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} Z_{-\epsilon}(h) = |\alpha|\, \widehat{\mathcal{R}}_S(\mathcal{H}).$$

For $\alpha = 0$, the class consists of the zero function, so the same identity holds. Averaging over $S \sim \mu^m$ proves

$$\boxed{\mathcal{R}_m(\alpha\mathcal{H}) = |\alpha| \mathcal{R}_m(\mathcal{H}).}$$

(b) For every fixed sign vector, the two functions can be chosen independently, so

$$\sup_{h \in \mathcal{H},\, h' \in \mathcal{H}'} \big(Z_\epsilon(h) + Z_\epsilon(h')\big) = \sup_{h \in \mathcal{H}} Z_\epsilon(h) + \sup_{h' \in \mathcal{H}'} Z_\epsilon(h').$$

This equality of suprema does not require maximizers: approximate each supremum arbitrarily closely and then choose the corresponding pair. Taking expectations in the signs gives

$$\widehat{\mathcal{R}}_S(\mathcal{H} + \mathcal{H}') = \widehat{\mathcal{R}}_S(\mathcal{H}) + \widehat{\mathcal{R}}_S(\mathcal{H}').$$

Averaging over the sample yields

$$\boxed{\mathcal{R}_m(\mathcal{H} + \mathcal{H}') = \mathcal{R}_m(\mathcal{H}) + \mathcal{R}_m(\mathcal{H}').}$$

(c) Put $\mathcal{D} = \mathcal{H} + (-\mathcal{H}')$ and $|\mathcal{D}| = \{x \mapsto |d(x)| : d \in \mathcal{D}\}$. The identity $\max\{a, b\} = (a + b + |a - b|)/2$ gives the class inclusion

$$\mathcal{M} \subseteq \tfrac{1}{2}(\mathcal{H} + \mathcal{H}' + |\mathcal{D}|).$$

Monotonicity under class inclusion and parts (a)–(b) therefore imply

$$\widehat{\mathcal{R}}_S(\mathcal{M}) \le \frac{1}{2} \left(\widehat{\mathcal{R}}_S(\mathcal{H}) + \widehat{\mathcal{R}}_S(\mathcal{H}') + \widehat{\mathcal{R}}_S(|\mathcal{D}|) \right).$$

The supplied contraction inequality with $\phi(t) = |t|$ gives

$$\widehat{\mathcal{R}}_S(|\mathcal{D}|) \le \widehat{\mathcal{R}}_S(\mathcal{D}) = \widehat{\mathcal{R}}_S(\mathcal{H}) + \widehat{\mathcal{R}}_S(-\mathcal{H}') = \widehat{\mathcal{R}}_S(\mathcal{H}) + \widehat{\mathcal{R}}_S(\mathcal{H}').$$

Consequently, the fixed-sample inequality is

$$\widehat{\mathcal{R}}_S(\mathcal{M}) \le \widehat{\mathcal{R}}_S(\mathcal{H}) + \widehat{\mathcal{R}}_S(\mathcal{H}').$$

Taking expectations over S proves

$$\boxed{\mathcal{R}_m(\mathcal{M}) \le \mathcal{R}_m(\mathcal{H}) + \mathcal{R}_m(\mathcal{H}').}$$

Problem 4. Regularized neural networks

All expectations below are over the independent Rademacher signs; the sample S is fixed.

(a) For fixed signs and hidden weights, define

$$a_j := \sum_{i=1}^{m} \epsilon_i \sigma(u_j^\top x_i).$$

The ℓ_1 constraint on the output weights gives

$$\sup_{\|w\|_1 \le \Lambda_0} \sum_{j=1}^{n_2} w_j a_j = \Lambda_0 \max_{1 \le j \le n_2} |a_j|.$$

Indeed, the upper bound follows from $|\sum_j w_j a_j| \le \|w\|_1 \max_j |a_j|$; equality follows by putting all output weight on a coordinate attaining the maximum, with its sign chosen to agree with a_j. If every a_j is zero or $\Lambda_0 = 0$, the equality holds with value zero.

Since $n_2 \ge 1$ and the hidden weights are independently constrained,

$$\sup_{\|u_j\|_2 \le \Lambda \ \forall j} \max_j \left| \sum_i \epsilon_i \sigma(u_j^\top x_i) \right| = \sup_{\|u\|_2 \le \Lambda} \left| \sum_i \epsilon_i \sigma(u^\top x_i) \right|.$$

The upper bound is immediate, and the reverse bound follows by assigning any admissible u to the first hidden unit. Optimizing over w first and taking expectations thus proves the exact identity

$$\boxed{\widehat{\mathcal{R}}_S(\mathcal{H}) = \frac{\Lambda_0}{m} \mathbb{E}_\epsilon \sup_{\|u\|_2 \le \Lambda} \left| \sum_{i=1}^{m} \epsilon_i \sigma(u^\top x_i) \right|.}$$

This also covers $\Lambda = 0$: then $u = 0$ and $\sigma(0) = 0$, so both sides vanish.

(b) Let $\mathcal{G} = \{x \mapsto u^\top x : \|u\|_2 \le \Lambda\}$. The supplied absolute-value contraction inequality and part (a) give

$$\widehat{\mathcal{R}}_S(\mathcal{H}) \le 2L\Lambda_0 \, \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_i \epsilon_i g(x_i) \right|.$$

The ball of admissible u is symmetric, so $\mathcal{G} = -\mathcal{G} = \mathcal{H}_{\text{lin}}$. In particular, for each sign vector,

$$\sup_{g \in \mathcal{G}} |Z_\epsilon(g)| = \sup_{g \in \mathcal{G}} Z_\epsilon(g).$$

This is where symmetry converts the absolute-value complexity to the convention used in the problem. Hence

$$\boxed{\widehat{\mathcal{R}}_S(\mathcal{H}) \le 2L\Lambda_0 \widehat{\mathcal{R}}_S(\mathcal{H}_{\text{lin}}).}$$

(c) Set $v_\epsilon = \sum_{i=1}^m \epsilon_i x_i$. Euclidean duality gives

$$\sup_{\|u\|_2 \le \Lambda, \ s \in \{-1,+1\}} s \, u^\top v_\epsilon = \Lambda \|v_\epsilon\|_2.$$

For $v_\epsilon \neq 0$ this is attained by $s = 1$ and $u = \Lambda v_\epsilon / \|v_\epsilon\|_2$; if $v_\epsilon = 0$, both sides are zero. Thus, also when $\Lambda = 0$,

$$\boxed{\widehat{\mathcal{R}}_S(\mathcal{H}_{\text{lin}}) = \frac{\Lambda}{m} \mathbb{E}_\epsilon \| \sum_{i=1}^{m} \epsilon_i x_i \|_2.}$$

(d) By Cauchy–Schwarz,

$$\mathbb{E}_\epsilon \|v_\epsilon\|_2 \le \sqrt{\mathbb{E}_\epsilon \|v_\epsilon\|_2^2}.$$

Independence, $\mathbb{E}\epsilon_i = 0$, and $\epsilon_i^2 = 1$ imply

$$\mathbb{E}_\epsilon \|v_\epsilon\|_2^2 = \sum_{i,k=1}^{m} \mathbb{E}[\epsilon_i \epsilon_k] \langle x_i, x_k \rangle = \sum_{i=1}^{m} \|x_i\|_2^2.$$

Substituting into part (c) proves

$$\boxed{\widehat{\mathcal{R}}_S(\mathcal{H}_{\text{lin}}) \le \frac{\Lambda}{m} \sqrt{\sum_{i=1}^{m} \|x_i\|_2^2}.}$$

(e) Combining parts (b) and (d), then using $\sum_i \|x_i\|_2^2 \le m r^2$, gives

$$\boxed{\widehat{\mathcal{R}}_S(\mathcal{H}) \le \frac{2L\Lambda_0\Lambda}{m} \sqrt{\sum_{i=1}^{m} \|x_i\|_2^2} \le \frac{2L\Lambda_0\Lambda r}{\sqrt{m}}.}$$

The bound is independent of the hidden-layer width n_2 for every $n_2 \ge 1$. The ℓ_1 constraint controls the total output weight, so optimizing over that weight reduces the complexity to one hidden unit, regardless of how many units are available.