Homework 3 Sample Solutions

Problem 1: Lower bounds for hypothesis selection

Take D to be uniform on $\{-1, 1\}^k \subseteq \mathbb{R}^k$, and use the k classifiers $f_j(x) = \mathbf{1}_{\{x_j \geq 0\}}$, i.e., whether the i-th coordinate is positive or negative. For $i \neq j$, if $X \sim D$, then X_i and X_j are independent unbiased bits, so

$$\Pr[f_i(X) \neq f_j(X)] = \tfrac{1}{2}.$$

Thus the construction has constant separation: it satisfies the required condition with $\varepsilon = 1/2$ (or any smaller ε).

Let J be the unknown, true index, and suppose we consider the case where it is chosen uniformly at random from $\{1, \ldots, k\}$. Conditioned on the inputs $X_1, \ldots, X_n$, the only information about J is the label string $(f_J(X_1), \ldots, f_J(X_n))$. There are at most 2^n such strings, and for each string the learner produces one output. If several different choices of J give the same string, that output can be correct for at most one of them. It follows that

$$\Pr[\text{recover } J \mid X_1, \ldots, X_n] \leq \frac{2^n}{k}.$$

Setting the right hand side to some small constant and solving for n yields the desired bound. Note that this argument covers every algorithm, including randomized ones.

Problem 2: VC dimension

(a) Use the standard basis $S = \{e_1, \ldots, e_d\}$. For any desired labels $b_1, \ldots, b_d \in \{0, 1\}$, set $v_i = 2b_i - 1$. Then $\langle v, e_i \rangle = v_i$ is 1 when $b_i = 1$ and -1 when $b_i = 0$. Thus every labeling of S is realized.

(b) Any $d + 1$ points $x_1, \ldots, x_{d+1} \in \mathbb{R}^d$ must be linearly dependent. Thus, there exist some parameters $a_1, \ldots, a_{d+1} \in \mathbb{R}^d$ which are not all zero so that $\sum_i a_i x_i = 0$. Flip all the coefficients if needed so that at least one $a_i > 0$. Give label 0 to points with $a_i > 0$, label 1 to points with $a_i < 0$, and either label to points with $a_i = 0$.

If some v realized those labels, then $\langle v, x_i \rangle < 0$ for $a_i > 0$ and $\langle v, x_i \rangle \geq 0$ for $a_i < 0$. Consequently every $a_i \langle v, x_i \rangle$ is nonpositive and at least one is strictly negative, so in particular, $\sum_{a_i} \langle v, x_i \rangle \neq 0$. But on the other hand, we also have

$$\sum_i a_i \langle v, x_i \rangle = \left\langle v, \sum_i a_i x_i \right\rangle = 0 \,,$$

which is a contradiction. Together, (a) and (b) give $\boxed{\text{VCdim}(A) = d}$.

(c) Since $q(x) > 0$, the sign of $p(x)/q(x)$ is the sign of $p(x)$. Let $\phi(x)$ list all monomials of total degree at most k, including the constant 1. There are $m = \binom{d+k}{k}$ such monomials, and thus every classifier in A_k is a homogeneous linear threshold in $\mathbb{R}^m$ applied to $\phi(x)$. If A_k shattered $m + 1$ inputs, these thresholds would shatter their feature vectors, contrary to (b) in dimension m. Hence

$$\boxed{\text{VCdim}(A_k) \leq \binom{d+k}{k}.}$$

Problem 3: Regularization and measurement error

(a) Setting the gradient of the objective to zero gives $(X^\top X + \lambda I)\theta_\lambda = X^\top y$. Since $\lambda > 0$, the matrix on the left is positive definite, so the unique minimizer is

$$\boxed{\theta_\lambda = (X^\top X + \lambda I)^{-1} X^\top y.}$$

(b) Let $X_{\cdot j}$ denote column j of X and put $\eta_j = X_{\cdot j}^\top \xi$. The assumptions give $\mathbb{E}\eta_j = 0$ and $\mathrm{Var}(\eta_j) = \sigma^2 \|X_{\cdot j}\|_2^2 = \sigma^2 s_j$. Using $y = X\theta^* + \xi$ and the diagonal form of $X^\top X$ gives

$$(\theta_\lambda)_j = \frac{s_j \theta_j^* + \eta_j}{s_j + \lambda}.$$

Therefore

$$\boxed{\mathbb{E}[(\theta_\lambda)_j] - \theta_j^* = -\frac{\lambda}{s_j + \lambda}\theta_j^*, \qquad \mathrm{Var}((\theta_\lambda)_j) = \frac{\sigma^2 s_j}{(s_j + \lambda)^2}.}$$

For $s_j > 0$, increasing λ shrinks the expected estimate toward zero: the bias magnitude increases toward $|\theta_j^*|$ while the variance decreases toward zero. If $s_j = 0$, column j is zero and $(\theta_\lambda)_j = 0$ for every $\lambda > 0$.

(c) The expected squared error of each coordinate is its squared bias plus its variance. Summing gives

$$\boxed{\mathrm{MSE}(\lambda) = \sum_{j=1}^{d} \frac{\lambda^2 (\theta_j^*)^2 + \sigma^2 s_j}{(s_j + \lambda)^2}.}$$

No independence between the coordinates of θ_λ is needed for this sum.

(d) Take $n = d = 1$, $X = [1]$, $\theta^* = 1$, and $\sigma^2 = 1$. The unregularized estimator is $y = 1 + \xi$ and has MSE 1. Taking $\lambda = 1$ instead gives

$$\boxed{\mathrm{MSE}(1) = \frac{1 + 1}{(1 + 1)^2} = \tfrac{1}{2} < 1.}$$

Here the variance falls from 1 to 1/4, more than offsetting the squared bias of 1/4.