DSC 190/291 · Assignment 4 Solutions

UCSD · Spring 2026

Part A: Sparse Linear Predictors and Model Selection (40 points)

This problem studies structural risk minimization in a setting where the complexity parameter is sparsity. The goal is to connect non-uniform learning bounds with the computational cost of searching over sparse predictors.

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Use the convention $\text{sign}(z) = +1$ for $z \geq 0$ and $\text{sign}(z) = -1$ for $z < 0$. For $1 \leq k \leq d$, define the class of k-sparse homogeneous linear classifiers

$$\mathcal{H}_k = \{x \to \text{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d, \|w\|_0 \leq k\}.$$

Let $\mathcal{H} = \bigcup_{k=1}^{d} \mathcal{H}_k$. Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$, and let $S = ((x_1, y_1), ..., (x_n, y_n)) \sim \mathcal{D}^n$. When a weight vector w is used as the argument of L_S or $L_{\mathcal{D}}$, it denotes the classifier $x \to \text{sign}(\langle w, x \rangle)$.

1. **VC dimension through supports.**

   Prove an upper bound of the form

   $$\text{VCdim}(\mathcal{H}_k) = O\left(k \log\left(e\frac{d}{k}\right)\right).$$

   Hint: write $\mathcal{H}_k$ as a union over the possible supports of w, then combine the growth-function bounds for those subclasses.

   Solution.

   For each $I \subseteq [d]$ with $|I| = k$, let $\mathcal{H}_I$ be the class of homogeneous halfspaces that use only coordinates in I. Since every support of size at most k is contained in some I of size exactly k,

   $$\mathcal{H}_k \subseteq \bigcup_{I \subseteq [d],\ |I| = k} \mathcal{H}_I.$$

   For a fixed I, the class $\mathcal{H}_I$ is a class of homogeneous halfspaces in $\mathbb{R}^k$, so

   $$\text{VCdim}(\mathcal{H}_I) \leq k.$$

   Therefore, by Sauer-Shelah, for every $m \geq k$,

   $$\Gamma_{\mathcal{H}_I}(m) \leq \left(e\frac{m}{k}\right)^k.$$

   There are $\binom{d}{k} \leq \left(e\frac{d}{k}\right)^k$ possible sets I, hence

   $$\Gamma_{\mathcal{H}_k}(m) \leq \binom{d}{k}\left(e\frac{m}{k}\right)^k \leq \left(e\frac{d}{k}\right)^k\left(e\frac{m}{k}\right)^k.$$

If $m < k$, the desired bound is immediate since $\log(e\frac{d}{k}) \geq 1$. Otherwise, if $\mathcal{H}_k$ shatters m points, then $2^m \leq \Gamma_{\mathcal{H}_k}(m)$, so

$$2^m \leq \left(e\frac{d}{k}\right)^k \left(e\frac{m}{k}\right)^k.$$

Let $u = \frac{m}{k}$ and $A = e\frac{d}{k}$. Then this implies $2^u \leq Aeu$. Since $A \geq e$, for a sufficiently large universal constant C, the inequality fails whenever $u > C \log A$. Hence

$$m \leq Ck \log\left(e\frac{d}{k}\right),$$

which proves

$$\mathrm{VCdim}(\mathcal{H}_k) = O\left(k \log\left(e\frac{d}{k}\right)\right).$$

2. **Penalty-based SRM.**

Choose a prior $p_k > 0$ over sparsity levels $k = 1, ..., d$ with $\sum_{k=1}^d p_k \leq 1$; for example, you may take $p_k \propto \frac{1}{k^2}$. Define an SRM rule of the form

$$\hat{h} \in \arg\min_{1 \leq k \leq d, h \in \mathcal{H}_k} \{L_S(h) + \mathrm{pen}(k, n, \delta)\}.$$

Derive a high-probability oracle inequality of the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{1 \leq k \leq d, h \in \mathcal{H}_k} \left\{ L_{\mathcal{D}}(h) + C\sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{1}{p_k}\right) + \log\left(\frac{1}{\delta}\right)}{n}} \right\}$$

for a universal constant C.

You may use the class-level SRM theorem from lecture, but you must instantiate every term in the bound and explain what statistical cost is paid for choosing the support and what cost is paid for choosing the sparsity level.

Solution.

Take, for example,

$$p_k = \frac{6}{\pi^2 k^2}$$

for $k = 1, ..., d$. Then $\sum_{k=1}^d p_k \leq 1$.

From the previous part,

$$\mathrm{VCdim}(\mathcal{H}_k) \leq C_0 k \log\left(e\frac{d}{k}\right)$$

for a universal constant C_0. Applying the class-level SRM theorem to the nested family $\mathcal{H} = \bigcup_{k=1}^d \mathcal{H}_k$ gives the following event with probability at least $1 - \delta$: for every k and every $h \in \mathcal{H}_k$,

$$|L_{\mathcal{D}}(h) - L_S(h)| \le C_1 \sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{1}{p_k}\right) + \log\left(\frac{1}{\delta}\right)}{n}}.$$

Define

$$\mathrm{pen}(k, n, \delta) = C_1 \sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{1}{p_k}\right) + \log\left(\frac{1}{\delta}\right)}{n}}.$$

On this event, for any k and any $h \in \mathcal{H}_k$,

$$\begin{aligned} L_{\mathcal{D}}\left(\hat{h}\right) &\le L_S\left(\hat{h}\right) + \mathrm{pen}\left(\hat{k}, n, \delta\right) \\ &\le L_S(h) + \mathrm{pen}(k, n, \delta) \\ &\le L_{\mathcal{D}}(h) + 2\,\mathrm{pen}(k, n, \delta), \end{aligned}$$

where $\hat{k}$ is any sparsity level for which $\hat{h} \in \mathcal{H}_{\hat{k}}$ and the middle inequality uses optimality of the SRM rule. Taking the infimum over k and $h \in \mathcal{H}_k$ and absorbing constants gives

$$L_{\mathcal{D}}\left(\hat{h}\right) \le \inf_{1 \le k \le d, h \in \mathcal{H}_k} \left\{ L_{\mathcal{D}}(h) + C \sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{1}{p_k}\right) + \log\left(\frac{1}{\delta}\right)}{n}} \right\}.$$

The term $k \log\left(e\frac{d}{k}\right)$ contains the price of searching over supports, through the factor $\log\binom{d}{k}$, together with the VC cost of fitting a k-dimensional halfspace once the support is fixed. The term $\log\left(\frac{1}{p_k}\right)$ is the additional price for selecting the sparsity level k itself.

3. **Comparison with dense halfspaces.**

   Suppose there exists $w^* \in \mathbb{R}^d$ with $\|w^*\|_0 \le s$ and $L_{\mathcal{D}}(w^*) \le \eta$. Use the SRM bound above to state a sample size sufficient to guarantee

   $$L_{\mathcal{D}}\left(\hat{h}\right) \le \eta + \varepsilon$$

   with probability at least $1 - \delta$.

   Then compare this with the sample size obtained by learning over all homogeneous halfspaces in $\mathbb{R}^d$. In what regime is the sparse bound smaller?

Solution.

Apply the oracle bound with $k = s$ and $h = w^*$. With probability at least $1 - \delta$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \le \eta + C \sqrt{\frac{s \log\left(e\frac{d}{s}\right) + \log\left(\frac{1}{p_s}\right) + \log\left(\frac{1}{\delta}\right)}{n}}.$$

Therefore it is enough to take

$$n \ge C^2 \frac{s \log\left(e\frac{d}{s}\right) + \log\left(\frac{1}{p_s}\right) + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}.$$

For the concrete choice $p_s \propto \frac{1}{s^2}$, this is

$$n = O\left(\frac{s\log\left(e\frac{d}{s}\right) + \log s + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}\right).$$

If we instead learn over all homogeneous halfspaces in $\mathbb{R}^d$, the VC dimension is d, so the corresponding agnostic sample size is

$$n = O\left(\frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}\right).$$

The sparse bound is smaller when

$$s\log\left(e\frac{d}{s}\right) + \log\left(\frac{1}{p_s}\right) \ll d.$$

With $p_s \propto \frac{1}{s^2}$, this means $s\log\left(e\frac{d}{s}\right) + \log s \ll d$, for example when the target uses far fewer than d coordinates.

4. **Validation as model selection.**

   Consider the following validation rule. Assume n is even, and split the sample into two equal parts S_1 and S_2. For each $k = 1, ..., d$, let

   $$w_k \in \arg\min_{\|w\|_0 \le k} L_{S_1}(w).$$

   Then choose the output $w_{\hat{k}}$ with

   $$\hat{k} \in \arg\min_{1 \le k \le d} L_{S_2}(w_k).$$

   Prove a validation-style oracle bound of the form

   $$L_{\mathcal{D}}(w_{\hat{k}}) \le \inf_{1 \le k \le d, w:\|w\|_0 \le k} \left\{ L_{\mathcal{D}}(w) + C\sqrt{\frac{k\log\left(e\frac{d}{k}\right) + \log\left(\frac{d}{\delta}\right)}{n}} \right\}$$

   for a universal constant C.

   Hint: after conditioning on S_1, the validation step is a finite-class selection problem over the candidates $w_1, ..., w_d$.

   Then compare this validation rule to the penalty-based SRM rule above: what samples are used to fit predictors, what samples are used to choose the complexity level, and what statistical price is paid for that separation?

Solution.

Since $|S_1| = |S_2| = \frac{n}{2}$, first apply the VC uniform-convergence bound to the training half S_1, and union bound over $k = 1, ..., d$. With probability at least $1 - \frac{\delta}{2}$, for every k and every w with $\|w\|_0 \le k$,

$$|L_{\mathcal{D}}(w) - L_{S_1}(w)| \le C_1 \sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{d}{\delta}\right)}{\frac{n}{2}}}.$$

On this event, since w_k minimizes L_{S_1} over $\mathcal{H}_k$, for every w with $\|w\|_0 \le k$,

$$\begin{aligned} L_{\mathcal{D}}(w_k) &\le L_{S_1}(w_k) + \varepsilon_k \\ &\le L_{S_1}(w) + \varepsilon_k \\ &\le L_{\mathcal{D}}(w) + 2\varepsilon_k, \end{aligned}$$

where

$$\varepsilon_k = C_1 \sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{d}{\delta}\right)}{\frac{n}{2}}}.$$

Now condition on S_1. The candidates $w_1, ..., w_d$ are fixed relative to the validation sample S_2. By Hoeffding's inequality and a union bound over these d candidates, with probability at least $1 - \frac{\delta}{2}$, for every k,

$$|L_{\mathcal{D}}(w_k) - L_{S_2}(w_k)| \le C_2 \sqrt{\frac{\log\left(\frac{d}{\delta}\right)}{\frac{n}{2}}}.$$

On this event, for any k,

$$\begin{aligned} L_{\mathcal{D}}(w_{\hat{k}}) &\le L_{S_2}(w_{\hat{k}}) + \alpha \\ &\le L_{S_2}(w_k) + \alpha \\ &\le L_{\mathcal{D}}(w_k) + 2\alpha, \end{aligned}$$

where $\alpha = C_2 \sqrt{\frac{\log\left(\frac{d}{\delta}\right)}{\frac{n}{2}}}$. Combining the two displays, we get

$$L_{\mathcal{D}}(w_{\hat{k}}) \le \inf_{w : \|w\|_0 \le k} \left\{ L_{\mathcal{D}}(w) + C \sqrt{\frac{k \log\left(e\frac{d}{k}\right) + \log\left(\frac{d}{\delta}\right)}{n}} \right\}$$

for every k. Taking the infimum over k proves the desired oracle inequality.

The validation rule uses S_1 to fit one predictor for each sparsity level, and uses the independent sample S_2 only to choose among the d fitted candidates. This separation makes the model-selection step a finite-class problem with a $\log d$ price, but it also means that only half the data are used to fit each w_k. The penalty-based SRM rule uses the full sample for both fitting and model selection, and pays for selecting k through the prior term $\log\left(\frac{1}{p_k}\right)$.

5. **Computation.**

   Now restrict attention to the realizable case. Suppose the sample is consistent with some k-sparse homogeneous halfspace. Describe the brute-force algorithm that enumerates supports $I \subseteq [d]$ with $|I| = k$ and solves a halfspace feasibility problem on each support. Estimate its runtime as a function of d, k, and n. In which regimes of k is this polynomial

in d? Explain how this illustrates the Week 4 distinction between sample complexity and computational complexity.

Solution.

The brute-force algorithm is:

1. Enumerate all supports $I \subseteq [d]$ with $|I| = k$.
2. For each support I, restrict every example x_i to the coordinates in I.
3. Solve the homogeneous halfspace feasibility problem in $\mathbb{R}^k$: find w_I such that
$$\langle w_I, x_{i,I} \rangle \geq 0$$
for positive examples and
$$\langle w_I, x_{i,I} \rangle < 0$$
for negative examples.
4. If a feasible w_I is found, extend it by zeros outside I and output the resulting k-sparse classifier.

The feasibility problem on a fixed support is a linear feasibility problem in k variables with n constraints, so it can be solved in time $\text{poly}(n, k)$, up to the usual bit-complexity dependence. The number of supports is
$$\binom{d}{k} \leq \left(e\frac{d}{k}\right)^k.$$

Thus the total runtime is
$$\binom{d}{k} \text{poly}(n, k) \leq \left(e\frac{d}{k}\right)^k \text{poly}(n, k).$$

This is polynomial in d for fixed k, with degree depending on k. If k grows with d, the enumeration factor is generally not polynomial in d; for example $k = O(\log d)$ gives a quasi-polynomial-type enumeration cost.

The statistical bound can be much smaller than the dense d-dimensional bound when the target is sparse, since it depends on $k \log\left(e\frac{d}{k}\right)$. The brute-force algorithm shows that a small statistical complexity term does not automatically give an efficient algorithm: finding the sparse support may require a combinatorial search.

Part B: PAC-Bayes for Thresholds (45 points)

PAC-Bayes bounds use $\text{KL}(Q \,\|\, P)$ as the complexity term. This problem compares that term with VC dimension for thresholds on a finite ordered domain.

Throughout, let
$$\mathcal{X}_N = \{1, 2, ..., N\}, \quad \mathcal{Y} = \{0, 1\},$$
and define the threshold class

$$\mathcal{H}_N = \{h_t : t \in \{1, ..., N+1\}\},$$

where

$$h_t(x) = \mathbf{1}[x \geq t].$$

Thus h_1 labels every point by 1, and h_{N+1} labels every point by 0.

Use the PAC-Bayes bound from lecture: For any distribution $\mathcal{D}$ over $\mathcal{X}_N \times \mathcal{Y}$ and any prior P, with probability at least $1-\delta$ over $S = ((x_1, y_1), ..., (x_n, y_n)) \sim \mathcal{D}^n$, simultaneously for all posteriors Q,

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\left(2\frac{n}{\delta}\right)}{2(n-1)}}.$$

Here

$$L_{\mathcal{D}}(Q) = \mathbb{E}_{h\sim Q}[L_{\mathcal{D}}(h)], \quad L_S(Q) = \mathbb{E}_{h\sim Q}[L_S(h)].$$

We work in the realizable setting.

1. **VC dimension and point-posterior PAC-Bayes.**

   Prove that $\mathrm{VCdim}(\mathcal{H}_N) = 1$. Let P be the uniform prior over $\mathcal{H}_N$, let $A(S)$ be any deterministic ERM rule that returns a consistent threshold $h_{\hat{t}}$, and define $Q_S = \delta_{h_{\hat{t}}}$. Compute $\mathrm{KL}(Q_S \parallel P)$, plug it into the PAC-Bayes theorem, and write the resulting bound on $L_{\mathcal{D}}(Q_S)$. Finally, explain in two or three sentences why this PAC-Bayes certificate scales with $\log(N+1)$ even though a VC-style realizable guarantee for thresholds should not scale with $\log N$.

   Solution.

   The class shatters any one point x: h_1 labels it by 1, while h_{N+1} labels it by 0. Therefore $\mathrm{VCdim}(\mathcal{H}_N) \geq 1$.

   It cannot shatter two ordered points $x < z$. The labeling $(1, 0)$ on (x, z) is impossible, because if $h_t(x) = 1$, then $x \geq t$, and since $z > x$, also $z \geq t$, so $h_t(z) = 1$. Hence $\mathrm{VCdim}(\mathcal{H}_N) = 1$.

   Under the uniform prior, $P(h_t) = \frac{1}{N+1}$ for every t. For the point posterior $Q_S = \delta_{h_{\hat{t}}}$,

   $$\mathrm{KL}(Q_S \parallel P) = \log\left(\frac{1}{P(h_{\hat{t}})}\right) = \log(N+1).$$

   Since $A(S)$ returns a consistent threshold, $L_S(Q_S) = 0$. Although Q_S depends on S, the PAC-Bayes theorem holds simultaneously for all posteriors, so it applies to this data-dependent posterior. The theorem gives, with probability at least $1-\delta$,

   $$L_{\mathcal{D}}(Q_S) \leq \sqrt{\frac{\log(N+1) + \log\left(2\frac{n}{\delta}\right)}{2(n-1)}}.$$

   The $\log(N+1)$ term appears because a point posterior must identify one threshold among $N+1$ possibilities under the uniform prior. This certificate pays for naming

the selected threshold. A VC-style realizable guarantee uses the fact that the ordered threshold class has VC dimension 1, so its sample complexity need not grow with $\log N$.

2. **Version-space posterior.**

   Define $a(S) = \max\{x_i : y_i = 0\}$, with $a(S) = 0$ if there are no negative examples, and define $b(S) = \min\{x_i : y_i = 1\}$, with $b(S) = N + 1$ if there are no positive examples. Prove that the set of thresholds consistent with S is

   $$V(S) = \{t : a(S) < t \leq b(S)\}$$

   and hence $|V(S)| = b(S) - a(S)$.

   Let Q_V be the uniform posterior over $\{h_t : t \in V(S)\}$. For the uniform prior P over $\mathcal{H}_N$, compute $\mathrm{KL}(Q_V \parallel P)$ exactly. Prove that $L_S(Q_V) = 0$, and explain why Q_V can give a better PAC-Bayes bound than a point posterior when $|V(S)|$ is large.

   Solution.

   A threshold h_t is consistent with a negative example $(x_i, 0)$ exactly when $x_i < t$. Thus consistency with all negative examples is equivalent to $a(S) < t$. Similarly, h_t is consistent with a positive example $(x_i, 1)$ exactly when $x_i \geq t$, so consistency with all positive examples is equivalent to $t \leq b(S)$. Therefore

   $$V(S) = \{t : a(S) < t \leq b(S)\}.$$

   Since t ranges over integers, this set has size $|V(S)| = b(S) - a(S)$.

   For Q_V uniform over $V(S)$ and P uniform over all $N + 1$ thresholds,

   $$\mathrm{KL}(Q_V \parallel P) = \sum_{t \in V(S)} \frac{1}{|V(S)|} \log\left(\frac{\frac{1}{|V(S)|}}{\frac{1}{N+1}}\right) = \log\left(\frac{N+1}{|V(S)|}\right).$$

   Every threshold in $V(S)$ is consistent with S, so each has empirical error zero. Therefore

   $$L_S(Q_V) = \mathbb{E}_{h \sim Q_V}[L_S(h)] = 0.$$

   Compared with a point posterior, whose KL is $\log(N + 1)$, the version-space posterior has KL $\log(N + 1) - \log|V(S)|$. When many thresholds remain consistent with the sample, spreading the posterior over all of them lowers the PAC-Bayes complexity term without increasing empirical error.

3. **Spreading helps KL, but can hurt true risk.**

   Let P be the uniform prior over $\mathcal{H}_N$. Let $\mathcal{D}$ be the realizable distribution whose marginal on $\mathcal{X}_N$ is uniform and whose labels are generated by h_τ. For any nonempty set $W \subseteq \{1, ..., N + 1\}$, let Q_W be uniform over $\{h_t : t \in W\}$. Prove that

   $$L_{\mathcal{D}}(Q_W) = \frac{1}{N|W|} \sum_{t \in W} |t - \tau|.$$

   Construct a concrete example with $N \geq 20$, a true threshold τ, and a realizable sample S such that $|V(S)|$ is large,

$$\mathrm{KL}\big(Q_{V(S)} \,\|\, P\big) < \mathrm{KL}\big(\delta_{h_\tau} \,\|\, P\big),$$

but

$$L_{\mathcal{D}}\big(Q_{V(S)}\big) > L_{\mathcal{D}}\big(\delta_{h_\tau}\big).$$

Explain why this does not contradict PAC-Bayes.

Solution.

Fix a threshold t. If $t < \tau$, then h_t and h_τ differ exactly on $x = t, t+1, ..., \tau - 1$, which has size $\tau - t$. If $t > \tau$, they differ exactly on $x = \tau, \tau + 1, ..., t - 1$, which has size $t - \tau$. If $t = \tau$, they agree everywhere. Since the marginal on $\mathcal{X}_N$ is uniform,

$$L_{\mathcal{D}}(h_t) = \frac{|t - \tau|}{N}.$$

Averaging over $t \sim Q_W$ gives

$$L_{\mathcal{D}}(Q_W) = \frac{1}{|W|} \sum_{t \in W} L_{\mathcal{D}}(h_t) = \frac{1}{N|W|} \sum_{t \in W} |t - \tau|.$$

For a concrete example, take $N = 20$ and $\tau = 10$. Let the sample contain the two labeled examples $S = ((1, 0), (20, 1))$. This sample is realizable by h_{10}. Here $a(S) = 1$, $b(S) = 20$, and $V(S) = \{2, 3, ..., 20\}$, so $|V(S)| = 19$.

The uniform prior has $P(h_t) = \frac{1}{21}$ for every t. Thus

$$\mathrm{KL}\big(Q_{V(S)} \,\|\, P\big) = \log\left(\frac{21}{19}\right) < \log(21) = \mathrm{KL}\big(\delta_{h_{10}} \,\|\, P\big).$$

But

$$L_{\mathcal{D}}\big(\delta_{h_{10}}\big) = 0,$$

while

$$L_{\mathcal{D}}\big(Q_{V(S)}\big) = \frac{1}{20 \cdot 19} \sum_{t=2}^{20} |t - 10| = \frac{36 + 55}{380} = \frac{91}{380} > 0.$$

This does not contradict PAC-Bayes. The theorem gives an upper bound on the true risk of each posterior; it does not say that the posterior with smaller KL must have smaller true risk. In this example the version-space posterior has zero empirical error and smaller KL, but it places mass on thresholds that are consistent with the small sample and still make mistakes under the population distribution.

4. **Every fixed prior can be attacked.**

   Let P be any prior over $\mathcal{H}_N$, fixed before seeing the sample. Assume $N \geq 3$.

   Prove that there exists an internal threshold $\tau \in \{2, ..., N\}$ such that

   $$P(h_\tau) \leq \frac{1}{N - 1}.$$

   For this τ, define a realizable distribution $\mathcal{D}_\tau$ by

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_\tau}[x=\tau-1,y=0]=\frac{1}{2},\quad \mathbb{P}_{(x,y)\sim\mathcal{D}_\tau}[x=\tau,y=1]=\frac{1}{2}.$$

Prove that with probability at least $1-2^{1-n}$ over $S\sim\mathcal{D}_\tau^n$, the sample contains both support points.

On this event, prove that $V(S)=\{\tau\}$. Conclude that any posterior Q with $L_S(Q)=0$ must be $Q=\delta_{h_\tau}$. Prove that, on this event,

$$\mathrm{KL}(Q\parallel P)\geq\log(N-1)$$

for every zero-empirical-error posterior Q.

Explain what this lower bound does and does not show. In particular, why is it a limitation of zero-empirical-error PAC-Bayes certificates with a fixed prior, rather than a lower bound on the sample complexity of learning thresholds?

Solution.

There are $N-1$ internal thresholds $h_2,...,h_N$. Since the total prior mass over all thresholds is at most 1,

$$\sum_{\tau=2}^{N}P(h_\tau)\leq 1.$$

Therefore at least one internal threshold $\tau\in\{2,...,N\}$ satisfies $P(h_\tau)\leq\frac{1}{N-1}$.

Now fix such a τ and consider $\mathcal{D}_\tau$. The only two support points are $(\tau-1,0)$ and $(\tau,1)$, each with probability $\frac{1}{2}$. The probability that the sample misses $(\tau-1,0)$ is 2^{-n}, and the probability that it misses $(\tau,1)$ is also 2^{-n}. By a union bound, the probability that it misses at least one of the two support points is at most $2\cdot 2^{-n}=2^{1-n}$. This distribution is realizable by h_τ, since $h_{\tau(\tau-1)}=0$ and $h_{\tau(\tau)}=1$. Hence with probability at least $1-2^{1-n}$, both support points appear in S.

On this event, $a(S)=\tau-1$ and $b(S)=\tau$, so

$$V(S)=\{t:\tau-1<t\leq\tau\}=\{\tau\}.$$

If $L_S(Q)=0$, then $\mathbb{E}_{h\sim Q}[L_S(h)]=0$. Since empirical losses are nonnegative, Q must put all its mass on zero-empirical-error thresholds. The only such threshold is h_τ, so $Q=\delta_{h_\tau}$.

Therefore, for every zero-empirical-error posterior Q on this event,

$$\mathrm{KL}(Q\parallel P)=\mathrm{KL}\Big(\delta_{h_\tau}\parallel P\Big)=\log\left(\frac{1}{P(h_\tau)}\right)\geq\log(N-1).$$

If $P(h_\tau)=0$, then the KL is infinite, so the same lower bound holds.

This lower bound shows that for every fixed prior, there is a realizable threshold problem on which any zero-empirical-error PAC-Bayes certificate must pay a $\log N$ complexity term once the sample identifies a low-prior threshold. It does not show that thresholds require $\log N$ samples to learn. VC theory gives dimension 1, and for the two-point distribution above, seeing both support points already identifies the correct threshold and yields zero true risk. The lower bound is therefore about this particular

style of PAC-Bayes certificate, with a fixed prior and zero empirical error, not about the intrinsic sample complexity of the threshold class.