EE378C: Homework #3 Solutions

Due on Wednesday, May 19, 2021

Please hand in your homework via Gradescope before 11:59 PM. Typing your solution using LaTeX is highly recommended.

1. The concept of orthogonal polynomials is useful not only in proving lower bounds, but also in constructing and analyzing unbiased estimators. Consider the same setting as Lecture 9: let $(P_\theta)_{\theta\in[\theta_0-\varepsilon,\theta_0+\varepsilon]}$ be a one-dimensional family of probability distributions with the following local expansion:

$$\frac{P_{\theta_0+u}(x)}{P_{\theta_0}(x)} = \sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{u^m}{m!}, \qquad \forall |u| \le \varepsilon, x \in \mathcal{X}.$$

In addition, we assume that the quantity $\sum_{x\in\mathcal{X}} P_{\theta_0+u}(x)P_{\theta_0+v}(x)/P_{\theta_0}(x)$ depends only on θ_0 and uv, for all $u, v \in [-\varepsilon, \varepsilon]$. Lecture 9 shows that $\{p_m(x;\theta_0)\}_{m\ge 0}$ are orthogonal in the sense that

$$\mathbb{E}_{X\sim P_{\theta_0}}[p_m(X;\theta_0)p_n(X;\theta_0)] = A_m(\theta_0)\cdot \mathbb{1}(m=n)$$

for some constants $\{A_m(\theta_0)\}_{m\ge 0}$.

(a) Show that for $X \sim P_{\theta_0+u}$, the estimator $p_m(X;\theta_0)$ is an unbiased estimator of u^m up to scaling. In other words, show that for all $u \in [-\varepsilon, \varepsilon]$,

$$\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)] = c_m u^m$$

for some constant c_m independent of u. Find the expression of c_m using $A_m(\theta_0)$.

(b) We can also say something about the second moment $\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)^2]$, with some additional properties of p_m. Suppose we could show that

$$p_m(x;\theta_0) = \sum_{\ell=0}^{m} b(m,\ell,\theta_0,u)\cdot p_\ell(x;\theta_0+u), \qquad \forall |u| \le \varepsilon, x \in \mathcal{X},$$

and we assume that both the local expansion and the orthogonality condition could be extended from θ_0 to any $\theta_0 + u$. Show that

$$\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)^2] = \sum_{\ell=0}^{m} b(m,\ell,\theta_0,u)^2\cdot A_\ell(\theta_0+u).$$

(c) Next we use a specific example to see how we could find the expression of $b(m,\ell,\theta_0,u)$. Suppose that $\mathcal{X} = \mathbb{N}, P_\theta = \mathsf{Poi}(\theta)$, so that for all $t \ge 0$, it holds that

$$\sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{\theta_0^m t^m}{m!} = \frac{P_{\theta_0(1+t)}(x)}{P_{\theta_0}(x)} = e^{-\theta_0 t}(1+t)^x = e^{ut}\cdot e^{-(\theta_0+u)t}(1+t)^x$$

$$= e^{ut} \cdot \frac{P_{(\theta_0+u)(1+t)}(x)}{P_{\theta_0+u}(x)} = e^{ut} \cdot \sum_{\ell=0}^{\infty} p_\ell(x;\theta_0+u)\frac{(\theta_0+u)^\ell t^\ell}{\ell!}.$$

Use this chain of equalities to argue that in the Poisson model, we have

$$b(m,\ell,\theta_0,u) = \binom{m}{\ell}\frac{(\theta_0+u)^\ell u^{m-\ell}}{\theta_0^m}.$$

Solution:

(a) Note that by assumption, the quantity

$$\begin{aligned}\sum_{x\in\mathcal{X}} \frac{P_{\theta_0+u}(x)P_{\theta_0+v}(x)}{P_{\theta_0}(x)} &= \sum_{x\in\mathcal{X}} P_{\theta_0+u}(x) \cdot \sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{v^m}{m!} \\ &= \sum_{m=0}^{\infty} \mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)] \cdot \frac{v^m}{m!}\end{aligned}$$

only depends on θ_0 and uv. Since the expectation $\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)]$ depends only on θ_0 and u, it must be the case that $\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)] = c_m v^m$ for some constant c_m independent of u. For the expression of c_m, recall that another way to write the above quantity is

$$\begin{aligned}\sum_{x\in\mathcal{X}} \frac{P_{\theta_0+u}(x)P_{\theta_0+v}(x)}{P_{\theta_0}(x)} &= \mathbb{E}_{X\sim P_{\theta_0}}\left[\frac{P_{\theta_0+u}(X)}{P_{\theta_0}(X)} \cdot \frac{P_{\theta_0+v}(X)}{P_{\theta_0}(X)}\right] \\ &= \sum_{m,n\geq 0} \mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)] \cdot \frac{u^m v^n}{m!n!} \\ &= \sum_{m=0}^{\infty} A_m(\theta_0) \cdot \frac{(uv)^m}{(m!)^2}.\end{aligned}$$

Comparing the above expressions, we conclude that $c_m = A_m(\theta_0)/m!$.

(b) The orthogonality condition of $p_m(x;\theta_0+u)$ says that

$$\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0+u)p_n(X;\theta_0+u)] = A_m(\theta_0+u)\cdot\mathbb{1}(m=n).$$

Consequently,

$$\begin{aligned}&\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)^2] \\ &= \sum_{\ell,\ell'=0}^{m} b(m,\ell,\theta_0,u)b(m,\ell',\theta_0,u)\cdot\mathbb{E}_{X\sim P_{\theta_0+u}}[p_\ell(X;\theta_0+u)p_{\ell'}(X;\theta_0+u)] \\ &= \sum_{\ell=0}^{m} b(m,\ell,\theta_0,u)^2 \cdot A_\ell(\theta_0+u).\end{aligned}$$

(c) Using $e^{ut} = \sum_{n\geq 0}(ut)^n/n!$ and comparing the coefficients of t^m, we have

$$p_m(x;\theta_0)\frac{\theta_0^m}{m!} = \sum_{\ell=0}^{m} \frac{u^{m-\ell}}{(m-\ell)!} \cdot \frac{(\theta_0+u)^\ell}{\ell!} p_\ell(x;\theta_0+u).$$

A simple rearrangement gives the claimed expression of $b(m,\ell,\theta_0,u)$.

2. This problem fills in the missing steps of the tree-based inequality in Lecture 10. Let $T = ([m], E)$ be an undirected graph with vertex set $[m]$ and edge set E, and be a tree in the sense that T is both connected and acyclic.

(a) Show that for any real numbers $x_1, \cdots, x_m$, it holds that

$$\sum_{i=1}^{m} x_i - \max_{i\in[m]} x_i \geq \sum_{(i,j)\in E} \min\{x_i, x_j\}.$$

(b) Use the result in Part (a), show that for probability distributions $P_1, \cdots, P_m$,

$$\min_{\Psi} \frac{1}{m}\sum_{i=1}^{m} P_i(\Psi \neq i) \geq \frac{1}{m}\sum_{(i,j)\in E} (1 - \|P_i - P_j\|_{\mathrm{TV}}),$$

where the minimum is over all possible tests $\Psi : \mathcal{X} \to [m]$.

(c) Evaluate the terms on both sides of (b) under $P_i = \mathcal{N}(i\Delta, 1)$ and a line tree with edge set $E = \{(1,2),(2,3),\cdots,(m-1,m)\}$, and show that they are equal.

Solution:

(a) Without loss of generality we assume that $x_1 \leq x_2 \leq \cdots \leq x_m$. We have

$$\begin{aligned}
\sum_{i=1}^{m} x_i - \max_{i\in[m]} x_i = \sum_{i=1}^{m-1} x_i &= (m-1)x_1 + \sum_{k=2}^{m-1}(m-k)(x_k - x_{k-1}) \\
&\overset{\text{(a)}}{\geq} (m-1)x_1 + \sum_{k=2}^{m-1}(x_k - x_{k-1}) \sum_{(i,j)\in E} \mathbb{1}(\min\{x_i,x_j\} \geq x_k) \\
&\overset{\text{(b)}}{=} \sum_{(i,j)\in E} \left(x_1 + \sum_{k=2}^{m-1}(x_k - x_{k-1}) \cdot \mathbb{1}(\min\{x_i,x_j\} \geq x_k) \right) \\
&= \sum_{(i,j)\in E} \min\{x_i, x_j\},
\end{aligned}$$

where (a) follows from the fact that restricting the tree T to the vertices $\{k, k+1, \cdots, m\}$ is still acyclic and thus

$$\sum_{(i,j)\in E} \mathbb{1}(\min\{x_i,x_j\} \geq x_k) = |\{(i,j) \in E : i \geq k, j \geq k\}| \leq m-k,$$

and (b) is due to $|E| = m-1$ for any tree on m vertices.

(b) In Lecture 10 we have shown that

$$\min_{\Psi} \frac{1}{m}\sum_{i=1}^{m} P_i(\Psi \neq i) = 1 - \frac{1}{m}\sum_{x\in\mathcal{X}} \max_{i\in[m]} P_i(x) = \frac{1}{m}\sum_{x\in\mathcal{X}} \left(\sum_{i=1}^{m} P_i(x) - \max_{i\in[m]} P_i(x) \right).$$

Now the claimed inequality follows from Part (a) and the identity

$$\sum_{x\in\mathcal{X}} \min\{P_i(x), P_j(x)\} = 1 - \|P_i - P_j\|_{\mathrm{TV}}.$$

(c) To evaluate the optimal test error, note that $P_1(x)$ attains the maximum over all $P_i(x)$ iff $x \le 3\Delta/2$, and $P_2(x)$ attains the maximum iff $3\Delta/2 \le x \le 5\Delta/2$, and so on. Consequently, summing over the respective regions gives

$$\begin{aligned}
\min_{\Psi} \frac{1}{m}\sum_{i=1}^{m} P_i(\Psi \neq i) &= 1 - \frac{1}{m}\sum_{x\in\mathcal{X}} \max_{i\in[m]} P_i(x) \\
&= 1 - \frac{1}{m}\left(2\cdot\Phi(\Delta/2) + (m-2)\cdot(2\Phi(\Delta/2)-1)\right) \\
&= \frac{2(m-1)}{m}\left(1-\Phi(\Delta/2)\right),
\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$. For the tree-based quantity, note that

$$\|P_i - P_{i+1}\|_{\mathrm{TV}} = 2\Phi(\Delta/2) - 1, \quad \forall i \in [m-1],$$

therefore the RHS quantity is the same as the LHS.

3. In this problem, we aim to show that under the p-dimensional Gaussian location model $X \sim \mathcal{N}(\theta, I_p)$ with $\|\theta\|_0 \le 1$, the minimax risk of denoising the 1-sparse vector in high dimensions satisfies

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^p:\|\theta\|_0\le 1} \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] \ge (2 - o(1))\log p, \qquad p \to \infty.$$

In fact, the same upper bound (with constant $2 + o(1)$) could also be achieved.

(a) We first establish an inequality which is similar to the generalized Fano's inequality, but involves the χ^2-divergence rather than the KL divergence in the definition of mutual information. Similar to Lecture 10, consider a generic prior π on Θ and loss function $L(\theta, a)$. Show that for every $\Delta > 0$,

$$\inf_{T} \mathbb{E}_{\theta\sim\pi}\mathbb{P}_\theta[L(\theta, T(X)) > \Delta] \ge 1 - p_\Delta - \sqrt{p_\Delta(1-p_\Delta) I_{\chi^2}(\theta; X)},$$

where the quantity $p_\Delta \triangleq \max_{a\in\mathcal{A}} \pi(\{\theta\in\Theta : L(\theta, a) \le \Delta\})$ is defined in Lecture 10, and $I_{\chi^2}(\theta; X)$ is the χ^2-mutual information defined as

$$I_{\chi^2}(\theta; X) \triangleq \inf_{Q_X} \mathbb{E}_{\theta\sim\pi}[\chi^2(P_{X|\theta}, Q_X)].$$

Hint: consider the kernel $K : \Theta\times\mathcal{X} \to \{0,1\}$ which sends (θ, x) to $\mathbb{1}(L(\theta, T(x)) > \Delta)$, apply a data-processing inequality specialized to χ^2-divergence. It is acceptable to prove a weaker result, provided that your result is sufficient to handle Part (c).

(b) Consider the uniform distribution $\pi = \mathsf{Unif}(\{\tau e_1, \cdots, \tau e_p\})$, where $e_1, \cdots, e_p$ are canonical vectors of $\mathbb{R}^p$, and $\tau > 0$ is a parameter to be determined. We apply the above method with $L(\theta, a) = \|\theta - a\|_2^2$ and $\Delta = (1-\delta)\tau^2$. Show that $p_\Delta \le 1/(p\delta)$. *Note: this shows that the best separation parameter here is $(1-o(1))\tau^2$, as opposed to $\tau^2/2$ in the original Fano's method.*

(c) Choosing $\tau = \sqrt{(2-\varepsilon)\log p}$, prove the claimed minimax lower bound using the above method and choosing $\varepsilon, \delta \to 0$ appropriately.

(d) Based on the minimax lower bound for 1-sparse vectors, argue that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s} \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] \ge (2 - o(1)) s \log(p/s), \qquad p/s \to \infty.$$

Solution:

(a) Consider the kernel K in the hint. Under this kernel, the distribution $\pi(\theta) \times Q_X$ is mapped to $\mathrm{Bern}(p)$, and the distribution $\pi(\theta) \times P_{X|\theta}$ is mapped to $\mathrm{Bern}(q)$:

$$p = \mathbb{E}_{\theta\sim\pi}\mathbb{P}_{X\sim Q_X}[L(\theta, T(X)) > \Delta] \ge 1 - p_\Delta, \qquad q = \mathbb{E}_{\theta\sim\pi}\mathbb{P}_\theta[L(\theta, T(X)) > \Delta].$$

Then for any distribution Q_X on $\mathcal{X}$, we have

$$\begin{aligned}
\mathbb{E}[\chi^2(P_{X|\theta}, Q_X)] &\overset{\text{(a)}}{=} \chi^2(\pi(\theta) \times P_{X|\theta}, \pi(\theta) \times Q_X) \\
&\overset{\text{(b)}}{\ge} \chi^2(\mathrm{Bern}(q), \mathrm{Bern}(p)) \\
&\overset{\text{(c)}}{\ge} \chi^2(\mathrm{Bern}(q), \mathrm{Bern}(1 - p_\Delta)) \\
&= \frac{(1 - p_\Delta - q)^2}{p_\Delta(1 - p_\Delta)}
\end{aligned}$$

where (a) follows from simple algebra, (b) is due to data-processing, (c) uses the fact that $p \mapsto \chi^2(\mathrm{Bern}(q), \mathrm{Bern}(p))$ is increasing on $[1 - p_\Delta, 1]$ if $q \le 1 - p_\Delta$. Now choosing Q_X to attain the χ^2-mutual information completes the proof.

(b) It suffices to show that there can be at most $1/\delta$ points in $\{\tau e_1, \cdots, \tau e_p\}$ which can be Δ-close to a single point $a \in \mathbb{R}^p$ under L. Assume by contradiction that there could be $m > 1/\delta$ points, and w.l.o.g. we may assume that $\|\tau e_i - a\|_2^2 \le \Delta$ for all $i \in [m]$. However,

$$\begin{aligned}
m\Delta \ge \sum_{i=1}^m \|\tau e_i - a\|_2^2 &\ge \sum_{i=1}^m \left((\tau - a_i)^2 + (m-1)a_i^2\right) \\
&= \sum_{i=1}^m \left(m a_i^2 - 2\tau a_i + \tau^2\right) \overset{\text{(a)}}{\ge} (m-1)\tau^2,
\end{aligned}$$

where (a) follows from the minimization of the quadratic form. If $m > 1/\delta$, we have $(m-1)\tau^2 > m(1-\delta)\tau^2 = m\Delta$, which is a contradiction.

(c) Based on the χ^2 computation in HW2 Q5, we have

$$I_{\chi^2}(\theta; X) \le \mathbb{E}_{\theta\sim\pi}[\chi^2(\mathcal{N}(\theta, I_p), \mathcal{N}(0, I_p))] = \frac{\exp(\tau^2) - 1}{p} < p^{1-\varepsilon}.$$

Consequently, by the inequality in (a) and $L \ge \Delta \cdot \mathbb{1}(L > \Delta)$, we have

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^p:\|\theta\|_0\le 1} \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] \ge (1-\delta)(2-\varepsilon)\log p \cdot \left(1 - \frac{1}{p\delta} - \sqrt{\frac{1}{p^\varepsilon \delta}\left(1 - \frac{1}{p\delta}\right)}\right).$$

Now the claimed result follows from choosing $\delta = \omega(p^{-\varepsilon}) \cap o(1)$ and $\varepsilon \to 0$.

(d) Since $p/s \to \infty$, w.l.o.g we assume that p/s is an integer. Consider the following block-wise product prior: partition $[p]$ into s blocks each of size p/s, and assign the prior π in (b) on p/s elements to each block independently. Under this product prior, the sparsity level is always s, and the Bayes risk is precisely s times each individual Bayes risk, which by (c) is at least $(2 - o(1))\log(p/s)$, as desired.

4. This problem is a continuation of the learning theory example covered in Lecture 11. We have a function class $\mathcal{F}$ with VC dimension d, and n training data $(x_1, y_1), \cdots, (x_n, y_n)$ drawn from an unknown joint distribution P_{XY}, with $\mathcal{Y} = \{0, 1\}$. Define the following class $\mathcal{P}(\mathcal{F}, \varepsilon)$ of joint distributions where the best classifier has an error at most ε:

$$\mathcal{P}(\mathcal{F}, \varepsilon) = \left\{P_{XY} : \inf_{f^\star\in\mathcal{F}} P_{XY}(Y \ne f^\star(X)) \le \varepsilon\right\}.$$

So $\varepsilon = 0$ corresponds to the well-specified case, and $\varepsilon = 1$ corresponds to the misspecified case. Define the minimax excess risk $R^\star(\mathcal{F}, \varepsilon)$ over $\mathcal{P}(\mathcal{F}, \varepsilon)$ as

$$R^\star(\mathcal{F}, \varepsilon) = \inf_{\widehat{f}} \sup_{P_{XY}\in\mathcal{P}(\mathcal{F},\varepsilon)} \mathbb{E}\left[P_{XY}(Y \ne \widehat{f}(X)) - \inf_{f^\star\in\mathcal{F}} P_{XY}(Y \ne f^\star(X))\right].$$

Show that for all $\varepsilon \in [0, 1]$,

$$R^\star(\mathcal{F}, \varepsilon) = \Omega\left(\min\left\{\sqrt{\frac{d}{n}\cdot\varepsilon} + \frac{d}{n}, 1\right\}\right).$$

Note: the empirical risk minimization (ERM) approach attains the above lower bound possibly up to logarithmic factors. A possible hint is to try out the marginal distribution construction in the optimistic case and the conditional distribution construction in the pessimistic case, with appropriate parameters tailored for this problem.

Solution: As in Lecture 11, fix $x_1, \cdots, x_d \in \mathcal{X}$ and $f_v \in \mathcal{F}$ such that $f_v(x_i) = v_i$ for all $v \in \{\pm 1\}^d$ and $i \in [d]$. For each $u \in \{\pm 1\}^{d-1}$, consider the following joint distribution:

$$P_X(\{x_i\}) = \frac{\varepsilon}{d-1}, \quad i \in [d-1], \qquad P_X(\{x_d\}) = 1 - \varepsilon,$$

$$P_u(Y = u_i \mid X = x_i) = \frac{1}{2} + \delta, \quad i \in [d-1], \qquad P_u(Y = 1 \mid X = x_d) = 1.$$

Clearly, under each distribution $P_{XY,u}$, the classifier $f_{(u,1)} \in \mathcal{F}$ incurs an expected error $\varepsilon(1/2-\delta) < \varepsilon$. Moreover, for the loss function

$$L(P_{XY}, \widehat{f}) = P_{XY}(Y \neq \widehat{f}(X)) - \inf_{f^\star \in \mathcal{F}} P_{XY}(Y \neq f^\star(X)),$$

it holds that

$$L(P_{XY,u}, \widehat{f}) + L(P_{XY,u'}, \widehat{f}) \geq \frac{\varepsilon\delta}{d-1} \cdot d_{\mathrm{H}}(u, u'),$$

thus the separation condition of Assouad's lemma holds with $\Delta = \frac{\varepsilon\delta}{d-1}$. Finally, for $\delta \in (0, 1/4)$, the KL divergence between two neighboring distributions is

$$\max_{d_{\mathrm{H}}(u,u')=1} D_{\mathrm{KL}}(P_{XY,u}^{\otimes n} \| P_{XY,u'}^{\otimes n}) = \frac{n\varepsilon}{d-1} D_{\mathrm{KL}}(\mathrm{Bern}(1/2+\delta) \| \mathrm{Bern}(1/2-\delta)) \leq \frac{16n\varepsilon\delta^2}{d-1}.$$

Consequently, Assouad's lemma gives that

$$R^\star(\mathcal{F}, \varepsilon) \geq \frac{\varepsilon\delta}{4} \exp\left(-\frac{16n\varepsilon\delta^2}{d-1}\right).$$

If $\varepsilon \geq d/n$, we may choose $\delta = \sqrt{(d-1)/(16n\varepsilon)} \in (0, 1/4)$ to conclude that $R^\star(\mathcal{F}, \varepsilon) = \Omega(\sqrt{d\varepsilon/n})$. If $\varepsilon < d/n$, the target lower bound follows from the well-specified case

$$R^\star(\mathcal{F}, \varepsilon) \geq R^\star(\mathcal{F}, 0) = \Omega\left(\min\left\{\frac{d}{n}, 1\right\}\right).$$

5. In this problem we construct explicit packings in some examples, and therefore prove lower bounds on the packing number $M(A, d, \varepsilon)$. Recall that $M(A, d, \varepsilon)$ denotes the maximum number m of points $x_1, \cdots, x_m$ such that $d(x_i, x_j) \geq \varepsilon$ for every $i \neq j \in [d]$.

   (a) Suppose that $A = \{\pm 1\}^n$ is the binary hypercube, and $d = d_{\mathrm{H}}$ is the Hamming distance. Show the following Gilbert–Varshamov bound: for $\varepsilon \in [0, 1/2]$,

   $$M(\{\pm 1\}^n, d_{\mathrm{H}}, n\varepsilon) \geq 2^{n(1-h_2(\varepsilon))},$$

   where $h_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.
   Hint: you may use the following result: for every $x \in \{\pm 1\}^n$ and $\varepsilon \in [0, 1/2]$,

   $$|\{y \in \{0,1\}^n : d_{\mathrm{H}}(x, y) \leq n\varepsilon\}| \leq 2^{nh_2(\varepsilon)}.$$

   (b) Suppose that A is the set of all non-decreasing functions $f : [0,1] \to [0,1]$, and d is the L_2 norm between functions. Show that for $\varepsilon \in [0, 1]$, there exist universal constants $c_1, c_2 > 0$ such that

   $$\log M(A, L_2([0,1]), c_1\varepsilon) \geq \frac{c_2}{\varepsilon}.$$

(c) Suppose that A is the set of all convex functions $f : [0,1] \to [0,1]$, and d is the L_2 norm between functions. Show that for $\varepsilon \in [0,1]$, there exist universal constants $c_1, c_2 > 0$ such that

$$\log M(A, L_2([0,1]), c_1\varepsilon) \geq \frac{c_2}{\sqrt{\varepsilon}}.$$

Hint: for (b) and (c), you may try to break into several small intervals, find two possible function constructions in each interval, and concatenate them. The result in (a) might also be useful.

Solution:

(a) The maximal packing property ensures that every point of $\{\pm 1\}^n$ belongs to a ball with center in the packing and radius $n\varepsilon$. By the volume argument,

$$M(\{\pm 1\}^n, d_{\mathrm{H}}, n\varepsilon) \geq \frac{|\{\pm 1\}^n|}{\max_{x \in \{\pm 1\}^n} |\{y \in \{0,1\}^n : d_{\mathrm{H}}(x,y) \leq n\varepsilon\}|} \geq 2^{n(1-h_2(\varepsilon))}.$$

(b) W.l.o.g assume that $1/\varepsilon$ is an integer. By Part (a), we may find some $B \subseteq \{\pm 1\}^{1/\varepsilon}$ which is a $1/(4\varepsilon)$-packing of $\{\pm 1\}^{1/\varepsilon}$ under the Hamming distance, and $\log |B| \geq c_2/\varepsilon$ with $c_2 = (1 - h_2(1/4)) \log 2 > 0$. For $v \in B$, associate a function $f_v \in A$ which is the concatenation of $f_{1,v_1}, f_{2,v_2}, \cdots, f_{1/\varepsilon, v_{1/\varepsilon}}$, where f_{i,v_i} is defined on $[(i-1)\varepsilon, i\varepsilon)$:

$$f_{i,+}(x) = x, \qquad f_{i,-}(x) = (i-1)\varepsilon.$$

One may verify that every f_v is non-decreasing and maps $[0,1]$ to $[0,1]$. Moreover, for distinct $v, v' \in B$, we have

$$\|f_v - f_{v'}\|_2 = \sqrt{d_{\mathrm{H}}(v,v') \cdot \frac{\varepsilon^3}{3}} \geq \frac{\varepsilon}{\sqrt{12}}.$$

Consequently, we may choose $c_1 = 1/\sqrt{12}$ and $c_2 = (1 - h_2(1/4)) \log 2$.

(c) Similar to Part (b), we assume that $1/\sqrt{\varepsilon}$ is an integer, and $B \subseteq \{\pm 1\}^{1/\sqrt{\varepsilon}}$ is a $1/(4\sqrt{\varepsilon})$-packing of $\{\pm 1\}^{1/\sqrt{\varepsilon}}$ under the Hamming distance, and $\log |B| \geq c_2/\sqrt{\varepsilon}$. For every $v \in B$, associate a function $f_v \in A$ concatenates $f_{1,v_1}, \cdots, f_{1/\sqrt{\varepsilon}, v_{1/\sqrt{\varepsilon}}}$, with f_{i,v_i} defined on $[(i-1)\sqrt{\varepsilon}, i\sqrt{\varepsilon}]$:

$$f_{i,+}(x) = x^2, \qquad f_{i,-}(x) = (2i-1)\sqrt{\varepsilon}x - i(i-1)\varepsilon.$$

Note that $f_{i,-}(x)$ is simply the affine function connecting the endpoints of $f_{i,+}(x)$ at the boundaries of the interval. Since each $f_v(x)$ could be written as a pointwise maximum of some class of convex functions, we conclude that $f_v \in A$ for every $v \in \{\pm 1\}^{1/\sqrt{\varepsilon}}$. Moreover, for distinct $v, v' \in B$, we have

$$\|f_v - f_{v'}\|_2 = \sqrt{d_{\mathrm{H}}(v,v') \cdot \frac{(\sqrt{\varepsilon})^5}{30}} \geq \frac{\varepsilon}{\sqrt{120}},$$

establishing the claim.