Problem 1.1

Denote $\mathcal{H}_\circ$ as the hypothesis class defined in the problem. Since we assume that $\mathcal{H}_\circ$ is realizable, there is a perfect h_* that with $\text{err}_\mathcal{D}(h_*) = 0$. Denote h_{ERM} as the hypothesis which achieves 0 error-rate on $\mathcal{S}$. First assume there is at least one positive label. We choose h_{ERM} by finding the smallest radius such that every positive label is contained inside (the boundary is allowed as well). We must have that h_* contains h_{ERM}, otherwise h_* has non-zero error. Let the annulus B between h_{ERM} and h_* have probability mass ϵ. This area is where h_* and h_{ERM} disagree: for $b \in B$, $h_*(b) = 1$, $h_{ERM}(b) = 0$, and is therefore the generalization error. Then note that if there were $x_i \in B$, h_{ERM} would be larger. Therefore, $\mathbf{P}\{x_i \notin B\} \leq 1 - \epsilon$ since B has probability mass ϵ. Thus the probability that no $x_i \in \mathcal{S}$ is $\in B$ is $(1-\epsilon)^{|\mathcal{S}|}$.

Then $\mathbf{P}\{\text{err}_\mathcal{D}(h_{ERM}) > \epsilon\} \leq (1-\epsilon)^{|\mathcal{S}|} \leq e^{-\epsilon|\mathcal{S}|} = \delta$, and we choose $|S| = m_{\mathcal{H}_\circ}(\epsilon, \delta) = \frac{1}{\epsilon}\ln\left(\frac{1}{\delta}\right)$.

If there are no positive labels, then choose h_{ERM} so that it has a radius r_{ERM} not containing a negative label such that $r_{ERM} + \epsilon_0 = R_{\text{smallest negative label}}$. Then we have a two sided bound for h_*, since h_* can be inside h_{ERM} if there are no positive labels in the sample. Denote the annulus of radii $r_{ERM} < r < r_{ERM} + \epsilon_0$ as B_+, and the annulus of radius $r_{h_*} < r < r_{ERM}$ as B_-. Let the area of B_+ be ϵ_+, and the area of B_- be ϵ_-, and let $\epsilon = \max(\epsilon_+, \epsilon_-)$.

Then $\mathbf{P}\{\text{err}_\mathcal{D}(h_{ERM}) > \epsilon\} \leq 2(1-\epsilon)^{|\mathcal{S}|} \leq 2e^{-\epsilon|\mathcal{S}|} = \delta$, and we choose $|S| = \frac{1}{\epsilon}\ln\left(\frac{2}{\delta}\right)$.

Then we choose the maximum between the two cases as our lower bound, $|S| = \frac{1}{\epsilon}\ln\left(\frac{2}{\delta}\right)$.

Since this is not necessarily a positive integer, and the number of samples is an integer, we know that the minimum number of required samples is bounded above by $\lceil \frac{1}{\epsilon}\ln\left(\frac{2}{\delta}\right) \rceil$, and thus $\mathcal{H}_\circ$ is PAC-learnable since $m_{\mathcal{H}_\circ}(\epsilon, \delta)$ is polynomial in all required quantities and obtains the desired error bound.

Problem 1.2

Let $\mathcal{D}$ be the original distribution over concept $h \in \mathcal{C}$ ($(x, y) \in \mathcal{X} \times \mathcal{Y}$ have zero mass in $\mathcal{D}$ unless they satisfy the concept h), and let $\mathcal{D}'$ be the distribution under the uniform noise. We are given that concept class $\mathcal{C}$ is agnostically learnable. Therefore, with sample complexity $\text{poly}\left(\frac{1}{\epsilon}, \ln\left(\frac{1}{\delta}\right)\right)$, some algorithm $\mathcal{A}$ returns a hypothesis h_A such that

$$\text{err}_{\mathcal{D}'}(h_A) \le \min_{h^* \in \mathcal{C}} \text{err}_{\mathcal{D}'}(h^*) + \epsilon$$

Now consider that

$$\begin{aligned}\min_{h^* \in \mathcal{C}} \text{err}_{\mathcal{D}'}(h^*) &= \min_{h^* \in \mathcal{C}} \left\{(1 - \epsilon_0) * \text{err}_{\mathcal{D}}(h^*) + \frac{\epsilon_0}{2} * 1 + \frac{\epsilon_0}{2} * 0\right\} \\ &= \frac{\epsilon_0}{2} + (1 - \epsilon_0) * \min_{h^* \in \mathcal{C}} \text{err}_{\mathcal{D}}(h^*)\end{aligned} \tag{1}$$

since any hypothesis has a $\epsilon_0/2$ chance of getting the wrong answer in the noisy case and thus takes on $\epsilon_0/2$ error in expectation. Then, since we know that our original concept $h \in \mathcal{C}$, we know that

$$\min_{h^* \in \mathcal{C}} \text{err}_{\mathcal{D}}(h^*) = 0$$

and thus

$$\min_{h^* \in \mathcal{C}} \text{err}_{\mathcal{D}'}(h^*) = \frac{\epsilon_0}{2}$$

Thus, we conclude

$$\text{err}_{\mathcal{D}'}(h_A) \le \frac{\epsilon_0}{2} + \epsilon$$

as desired.

Problem 1.3

We have x_i are i.i.d Bernoulli random variables. We have $X = \sum_{i=1}^{k} x_i$.

First we have that $X \geq t$ implies $e^{\lambda X} \geq e^{\lambda t}$ for $\lambda \geq 0$, which holds since e^z is monotone increasing over its whole domain and since λ does not change the signs of the exponent terms since $\lambda \geq 0$. The other direction also holds for the same reason. Therefore, $\mathbf{P}\{X \geq t\} = \mathbf{P}\{e^{\lambda X} \geq e^{\lambda t}\} = \mathbf{P}\{e^{-\lambda t} \prod_{i=1}^{k} e^{\lambda x_i} \geq 1\}$. The probability that this event occurs is bounded above by $\mathbf{E}[e^{-\lambda t} \prod_{i=1}^{k} e^{\lambda x_i}]$ by Markov's inequality, since by taking exponents we made all values non-negative. Then since x_i are independent, and therefore $e^{\lambda x_i}$ are independent, $\mathbf{E}[e^{-\lambda t} \prod_{i=1}^{k} e^{\lambda x_i}] = e^{-\lambda t} \prod_{i=1}^{k} \mathbf{E}[e^{\lambda x_i}]$. Then $x_i = 1$ or -1, each with probability $\frac{1}{2}$, so $\mathbf{E}[e^{\lambda x_i}] = \frac{1}{2}\left(e^{\lambda} + e^{-\lambda}\right)$. Now we bound this quantity using Taylor Expansion $e^z = \sum_{i=0}^{\infty} \frac{z^i}{i!}$. Summing the Taylor expansions of e^{λ} and $e^{-\lambda}$, the odd terms cancel and the even terms are doubled. Diving by half, we get $\frac{1}{2}\left(e^{\lambda} + e^{-\lambda}\right) = \sum_{i=0}^{\infty} \frac{\lambda^{2i}}{(2i)!}$. Then we note that by Taylor expansion, $e^{\frac{\lambda^2}{2}} = \sum_{i=0}^{\infty} \frac{\lambda^{2i}}{2^i i!}$. Therefore we only need to show $(2i)! \geq 2^i i! \rightarrow \frac{(2i)!}{i!} \geq 2^i$ to get our upper bound. Consider that $(2i) * (2i-1) * \cdots * (i+1)$ has i terms, and each term is greater than or equal to 2. Therefore the inequality follows, and we have $\mathbf{P}\{X \geq t\} \leq e^{-\lambda t} e^{\frac{k\lambda^2}{2}}$. Then since $t > 0$, we can set $\lambda = \frac{t}{k}$ to get $\mathbf{P}\{X \geq t\} \leq e^{-\frac{t^2}{k}} e^{\frac{kt^2}{2k^2}} = e^{-\frac{t^2}{2k}}$, as desired.

Problem 1.4

(a) We prove that if a concept f is PAC-learnable by $\mathcal{H}$, then $f \in \mathcal{H}$ for finite domain $\mathcal{X}$.

Since f is PAC-learnable by $\mathcal{H}$, there is an algorithm $\mathcal{A}$ that returns a hypothesis h such that $\mathrm{err}_{\mathcal{D}}(h) < \epsilon$, where sample complexity $m_{\mathcal{H}}(\epsilon, \delta) = \mathrm{poly}\left(\frac{1}{\epsilon}, \ln\left(\frac{1}{\delta}\right), \ln\left(|\mathcal{H}|\right)\right)$. Since $\mathcal{X}$ is finite, and the number of samples cannot exceed the domain, we have that $|\mathcal{X}|$ is $\mathrm{poly}\left(\frac{1}{\epsilon}\right)$.

Therefore $|\mathcal{X}| \sim \left(\frac{1}{\epsilon}\right)^c$ and $\epsilon \leq \left(\frac{1}{|\mathcal{X}|}\right)^{\frac{1}{c}} \leq \frac{1}{|\mathcal{X}|}$ for $c \geq 1$. Since $\mathrm{err}_{\mathcal{D}}(h) < \epsilon \leq \frac{1}{|\mathcal{X}|}$, $\mathrm{err}_{\mathcal{D}}(h)$ must be 0, since the smallest non-zero error probability for finite domain $|\mathcal{X}|$ is $\frac{1}{|\mathcal{X}|}$. Therefore, $h = f$ and we have $f \in \mathcal{H}$ as desired.

(b) We provide a counter-example (f, $\mathcal{H}$, $\mathcal{X}$) to show that this theorem does not hold in general if $\mathcal{X}$ has infinite cardinality. Let $\mathcal{X} = \mathbb{R}$, and let $\mathcal{H} = \mathcal{H}_+$ be the hypothesis class of positive half-lines as defined in Lecture 2 (Definition 3.2), and define concept $f : \mathcal{X} \to \{0, 1\}$ as the 0 constant function; i.e. $f(x) = 0 \ \forall x \in \mathbb{R}$.

We have that $f \notin \mathcal{H}_+$ since the constant function does not change its value and the positive half-line function necessarily does, therefore implying non-zero generalization error for all $h \in \mathcal{H}$. We now show that this concept is in fact PAC-learnable by $\mathcal{H}_+$. Let us choose the hypothesis h for a given sample set S as

$$h_r(x) = \begin{cases} 1 & : x \geq r \\ 0 & : x < r \end{cases} \text{ where } r = 1 + \max_{(x_i, y_i) \in \mathcal{S}}(x_i)$$

Then let the probability mass of $\{x > r\}$ be $\epsilon > 0$. Note that this is the generalization error, since we should like to choose r at ∞ (but there is no such perfect classifier). The probability that a sample lands in this bad set is ϵ, therefore we have $\mathbf{P}\{\mathrm{err}_{\mathcal{D}}(h_r) > \epsilon\} \leq (1 - \epsilon)^{|\mathcal{S}|} \leq e^{-\epsilon|\mathcal{S}|} = \delta$, and we get that for $m_{\mathcal{H}}(\epsilon, \delta) \geq \frac{1}{\epsilon}\ln\left(\frac{1}{\delta}\right)$, we have $\mathrm{err}_{\mathcal{D}}(h_r) < \epsilon$ with probability $1 - \delta$, demonstrating that f is PAC-learnable even though $f \notin \mathcal{H}$.

Problem 1.5

We essentially repeat the proof of No Free Lunch from the book.

We will choose $|\mathcal{X}| \geq \frac{m}{2\epsilon}$ for some $\epsilon > 0$ to prove the statement. Then we can let m be any number smaller than $|\mathcal{X}| * 2\epsilon$ representing a training set size. Let C be the subset of $\mathcal{X}$ of size $m/2\epsilon$. Note there are $T = 2^{\frac{m}{2\epsilon}}$ possible functions from $C \rightarrow \{0, 1\}$. Denote these functions by $f_1, \cdots, f_T$. For each function, let distribution $\mathcal{D}_i$ over $\mathcal{X} \times \{0, 1\}$ be uniform (probability $1/|C|$) for point-labels (x, y) such that $f_i(x) = y$, and 0 everywhere else. By this definition, $\text{err}_{\mathcal{D}_i}(f_i) = 0$.

Now, we will prove that for every algorithm A that receives m examples from $C \times \{0, 1\}$ returns a function $h_A : C \rightarrow \{0, 1\}$ such that

$$\max_{i \in [T]} \mathbf{E}_{S \sim \mathcal{D}_i^m}[\text{err}_{\mathcal{D}_i}(h_A)] \geq \frac{1}{2} - \epsilon$$

This result demonstrates that for every algorithm receiving m samples there is a $(f_{i^*}, \mathcal{D}_{i^*})$ over which $\text{err}_{\mathcal{D}_{i^*}}(f_{i^*}) = 0$ and additionally for which

$$\mathbf{E}_{S \sim \mathcal{D}_{i^*}^m}[\text{err}_{\mathcal{D}_{i^*}}(h_A)] \geq \frac{1}{2} - \epsilon$$

as desired.

Now we prove it. Note there are $k = (m/2\epsilon)^m$ possible sequences of m examples from C. Denote these sequences by $S_1, \cdots, S_k$. We use the index j to denote which sequence, and the index i to denote the fact that we drew from $(f_i, \mathcal{D}_i)$. Indexing over both we get all possible training sets $\{S_j^i\}_{i \in [T], j \in [k]}$. We have

$$\begin{aligned}
\max_{i\in[T]} \mathbf{E}_{S\sim\mathcal{D}_i^m}[\mathrm{err}_{\mathcal{D}_i}(h_A)] &= \max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} \mathrm{err}_{\mathcal{D}_i}(h_A(S_j^i)) \\
&\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} \mathrm{err}_{\mathcal{D}_i}(h_A(S_j^i)) \\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} \mathrm{err}_{\mathcal{D}_i}(h_A(S_j^i)) \\
&\geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} \mathrm{err}_{\mathcal{D}_i}(h_A(S_j^i))
\end{aligned} \tag{2}$$

Now fix $j \in [k]$ as the minimum in the above, denote $S_j = (x_1, \cdots, x_m)$ and let us consider the samples $v_1, \cdots, v_p$ which do not appear in S_j. Since $|S_j| = m/2\epsilon$, we have that $p = m/2\epsilon - m = \left(\frac{1}{2\epsilon} - 1\right) m \geq m$. Therefore, for every $h : C \to \{0, 1\}$, since the error on a subset is $\leq$ the error on the whole set,

$$\begin{aligned}
\mathrm{err}_{\mathcal{D}_i}(h) &= \frac{2\epsilon}{m}\sum_{x\in C} \mathbf{1}\{h(x) \neq f_i(x)\} \\
&\geq \frac{2\epsilon}{m}\sum_{r=1}^{p} \mathbf{1}\{h(v_r) \neq f_i(v_r)\} \\
&= \frac{2\epsilon\left(\frac{1}{2\epsilon} - 1\right)}{p}\sum_{r=1}^{p} \mathbf{1}\{h(v_r) \neq f_i(v_r)\} \\
&= \frac{1-2\epsilon}{p}\sum_{r=1}^{p} \mathbf{1}\{h(v_r) \neq f_i(v_r)\}
\end{aligned} \tag{3}$$

Plugging this back in to what we had before, we get

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}\mathrm{err}_{\mathcal{D}_i}(h_A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1-2\epsilon}{p}\sum_{r=1}^{p}\mathbf{1}\{h(v_r)\neq f_i(v_r)\} \\
&= (1-2\epsilon)*\frac{1}{p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\{h_A(S_j^i)(v_r)\neq f_i(v_r)\} \qquad (4)\\
&\geq (1-2\epsilon)*\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\{h_A(S_j^i)(v_r)\neq f_i(v_r)\}
\end{aligned}$$

Fix $r \in [p]$ as the minimum. Now, we can partition the functions $f_1, \cdots, f_T$ into $T/2$ disjoint pairs where for pair of concepts (f_a, f_b) we have that they ONLY differ on the point v_r. That is $f_a(v_r) \neq f_b(v_r)$, and only at that point. Since $v_r \notin S_j^a$ and $v_r \notin S_j^b$ by definition, we have that $S_j^a = S_j^b$; i.e. the training sets will be the same. Furthermore, $h_A(S_j^a) = h_A(S_j^b)$ and if we denote pairs by $(f_a, f_b)_i$, we can just call this $h_A(S_j^i)$. Thus, we can rewrite

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\{h_A(S_j^i)(v_r)\neq f_i(v_r)\} &= \frac{1}{T}\sum_{i=1}^{T/2}\mathbf{1}\{h_A(S_j^i)(v_r)\neq f_{a_i}(v_r)\}+\mathbf{1}\{h_A(S_j^i)(v_r)\neq f_{b_i}(v_r)\} \\
&= \frac{1}{T}\sum_{i=1}^{T/2}1 \\
&= \frac{1}{T}*\frac{T}{2}=\frac{1}{2}
\end{aligned} \qquad (5)$$

since exactly one of the indicator functions in each sum term was 1. Thus, we can conclude that

$$\begin{aligned}
(1-2\epsilon)*\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\{h_A(S_j^i)(v_r)\neq f_i(v_r)\} &= (1-2\epsilon)*(1/2) \\
&= \frac{1}{2}-\epsilon
\end{aligned} \qquad (6)$$

as desired.

Problem 1.6

(a) We demonstrate that we cannot shatter the labelling $(0, 0, 1, 0)$.

Consider a sin curve with half-period $k \in \mathbb{R}^+$ (this corresponds to taking $\omega = \pi/k$) (WLOG consider $\omega > 0$, since $-\sin(x) = \sin(-x)$ and we'll just flip all labels).

Now consider that we're trying to use this sin curve to classify

$(x, 0); (2x, 0); (3x, 1); (4x, 0)$. Also note that x will only be below the sin curve on intervals of the form $[2m * k, (2m + 1) * k)$ for $m \in \mathbb{Z}$.

Suppose that $x = (n + \delta) * k$, where n is an even integer and $\delta \in [0, 1)$ is the fractional part. Note that n must be even so that x is below the sin curve.

Now consider $2x = 2(n+\delta)*k = (2n+2\delta)*k$. We have $2n$ is even. Since $2\delta < 2$, $\lfloor 2\delta \rfloor = 0$ or 1. We require it to be 0 to maintain evenness so that $2x$ is under the sin curve as well (we want to get the classification $(2x, 0)$). Therefore, $2\delta < 1$ and $\delta < 1/2$.

Now consider $3x = (3n + 3\delta) * k$. Again since we assumed n is even, $3n$ is also even. This time, $3\delta < 3/2$, we want $\lfloor 3\delta \rfloor = 1$ instead of 0 since even + odd = odd and since we want $3x$ to be above the sin curve. Therefore, we must have $3\delta > 1$, or $\delta > 1/3$.

Thus, we have established that to classify $(x, 0); (2x, 0); (3x, 1)$, we need

$$1/3 < \delta < 1/2$$

Finally, to shatter $(0, 0, 1, 0)$, we need $\lfloor 4n + 4\delta \rfloor$ to be even. $4n$ is even, thus 4δ must be even as well. We have from before that

$$4/3 < 4\delta < 4/2 = 2$$

implying that 4δ is necessarily odd, a contradiction. Thus, we cannot shatter $((x, 0), (2x, 0), (3x, 1), (4x, 0)$ with the family $\{t \to \sin(\omega t)\}$.

(b) We prove that we can shatter any configuration of points $S_m = \{2^{-i} : i \in [1, ..., m]\}$ with $\sin(\omega x)$ for any integer $m > 1$. This directly demonstrates that the VC-dimension of the class is infinite.

Let $x_i = 2^{-i}$, where i goes from 1 to some positive integer $m > 1$. Choose an arbitrary classification of these m points $y \in \{0,1\}^m$, denoting the classification of x_i by y_i. Then, define

$$\omega = \pi \left(\sum_{i=1}^{m} y_i 2^{i-1} \right)$$

Note that our classifier is the sign of the sin function. We will show that $\text{sgn}(\sin(\omega x_i)) = 1 - 2y_i$ in all cases. Since y was arbitrary, we have demonstrated that we can shatter S_m for all $m > 1$.

Let us calculate the classification of x_j, $1 \le j \le m$. Then,

$$\begin{aligned} \sin(\omega x_j) &= \sin(\omega 2^{-j}) \\ &= \sin\left(\pi \left(\sum_{i=1}^{m} y_i 2^{i-1} 2^{-j} \right) \right) \\ &= \sin\left(\pi y_i + \pi \sum_{i > j+1} y_i 2^{i-(j+1)} + \pi \sum_{i < j+1} y_i 2^{i-(j+1)} \right) \end{aligned} \tag{7}$$

Note that the last sum in the sin function is necessarily an even factor of π. Using the identity $\sin(x + 2\pi) = \sin(x)$, we can throw that term out. Thus we get

$$\begin{aligned} \sin(\omega x_j) &= \sin\left(\pi y_i + \pi \sum_{i > j+1} y_i 2^{i-(j+1)} \right) \\ &= \sin\left(\alpha\pi + \pi y_i \right) \end{aligned} \tag{8}$$

defining the remaining sum term coefficient of π to be α.

Now consider that

$$\begin{aligned}
\alpha = \sum_{i>j+1} y_i 2^{i-(j+1)} &\leq \sum_{i=j+2}^{m} \frac{1}{2^{(j+1)-i}} \\
&= \sum_{i=1}^{m-(j+1)} \left(\frac{1}{2}\right)^i \\
&< \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = 1
\end{aligned} \tag{9}$$

Now we consider both cases $y_j = 0$ and $y_j = 1$. If $y_j = 0$, then since $0 < \alpha < 1$ we have that the argument is in the first or second quadrant, meaning sin is positive which corresponds with $1 - 2 * 1 = -1$. If $y_j = 1$, then we add an extra term of π to the argument of sin, and since $\sin(x + \pi) = -\sin(x)$, the sign changes to positive which corresponds to $1 - 2 * 0 = 1$, as desired.

Thus we have demonstrated that we can completely shatter any S_m for all $m > 1$.