

Machine Learning Theory Lecture 5: PAC Bayesian Learning

Richard Xu

October 12, 2021

1 Warm-up: PAC Learning

1.1 Big picture

let size of data to be $|S| = m$:

$$\begin{aligned} \Pr(\exists_{h \in \mathcal{H}} : (\hat{R}_S(h) = 0) \cap (R(h) > \epsilon)) &\leq |\mathcal{H}| \exp^{-\epsilon m} \leq \delta \\ \implies m &\geq \frac{\log(|\mathcal{H}|)}{\epsilon} + \frac{\log(1/\delta)}{\epsilon} \end{aligned} \quad (1)$$

1.2 Some definition from usual PAC learning text

Firstly, instead of writing $\{x_i, y_i\}_{i=1}^m$, we can write it as:

$$\{x_i, c(x_i)\}_{i=1}^m \quad (2)$$

$$\begin{cases} y = c(x) & \in \mathcal{C} \quad \text{concept set} \\ \hat{y} = h(x) & \in \mathcal{H} \quad \text{hypothesis set} \end{cases} \quad (3)$$

Concept set is all set of “latent” functions that maps each x_i perfectly with y_i . I used $(x, y) \sim \mathcal{D}$ in all writings.

Of course, we cannot observe \mathcal{C} and c may not be a member of \mathcal{H} .

1.2.1 bound amount of over-fitting

We are interested to compute:

$$\Pr(\underbrace{R(h) - \hat{R}_S(h)}_{\text{amount of over-fitting}} > \epsilon) \leq \delta(\epsilon) \quad (4)$$

1.2.2 Version space

1. definition: Version space (VS)

$$\text{VS}_{\mathcal{H},S} \equiv \{\forall h \in \mathcal{H} \mid \hat{R}_S(h) = 0\} \quad (5)$$

2. definition: ϵ -exhausted version space

$\text{VS}_{\mathcal{H},S}$ is ϵ -exhausted iff:

$$\{\forall h \in \text{VS}_{\mathcal{H},S} \mid R(h) \leq \epsilon\} \quad (6)$$

meaning that for all hypothesis h in version space (zero training error), h has less than ϵ testing error (low error)

Theorem 1 *If hypothesis space \mathcal{H} is finite and S is a sequence of $m \geq 1$ i.i.d random examples of target concept c , then for any $0 \leq \epsilon \leq 1$:*

*Probability that version space $\text{VS}_{\mathcal{H},S}$ is **not** ϵ -exhausted is at most:*

$$|\mathcal{H}| \exp^{-\epsilon m} \quad (7)$$

1.2.3 proof

Start from just one $h_{\text{bad}} \in \mathcal{H}$ that assumes to be a bad classifier with error rate $\geq \epsilon$, i.e., $R(h_{\text{bad}}) \equiv \mathbb{E}_{(x,y) \sim \mathcal{D}}[R(h_1)] > \epsilon$. In order for h_1 to be element of the version space $\text{VS}_{\mathcal{H},S}$ (clearly, by including h_{bad} , $\text{VS}_{\mathcal{H},S}$ is **not** ϵ -exhausted), then, it must classify all m data in S correctly:

$$\begin{aligned} \Pr(\hat{R}_S(h_{\text{bad}}) = 0) &= \Pr(h_{\text{bad}}(x_1) = y_1 \cap \dots \cap h_{\text{bad}}(x_m) = y_m) \\ &= \Pr(\hat{R}_{x_1, y_1}(h_{\text{bad}}) = 0 \cap \dots \cap \hat{R}_{x_m, y_m}(h_{\text{bad}}) = 0) \\ &\leq (1 - \epsilon)^m \\ &\leq \exp^{-\epsilon m} \end{aligned} \quad (8)$$

using well known fact: $1 + x \leq \left(1 + \frac{x}{2}\right)^2 \leq \dots \leq \left(1 + \frac{x}{n}\right)^n \xrightarrow{n \rightarrow \infty} \exp^x$

$$\implies \Pr(\exists h \in \mathcal{H} : (\hat{R}_S(h) = 0)) \leq |\mathcal{H}| \exp^{-\epsilon m} \quad (9)$$

union bound, since any h_i makes it **not** ϵ -exhausted

1.2.4 what does tell you about m ?

$$\begin{aligned}
\text{let } |\mathcal{H}| \exp^{-\epsilon m} &\leq \delta \\
\implies -\epsilon m &\leq \log(\delta) - \log(|\mathcal{H}|) \\
\implies m &\geq \frac{\log(|\mathcal{H}|)}{\epsilon} - \frac{\log(\delta)}{\epsilon} \\
&= \frac{\log(|\mathcal{H}|)}{\epsilon} + \frac{\log(1/\delta)}{\epsilon}
\end{aligned} \tag{10}$$

let $|S| = m$:

$$\Pr(\exists_{h \in \mathcal{H}} : (\hat{R}_S(h) = 0) \cap (R(h) > \epsilon)) \leq |\mathcal{H}| \exp^{-\epsilon m} \leq \delta \tag{11}$$

Say we fix ϵ , then if one desires to have a very small chance (i.e., set δ to be very small) that the $VS_{\mathcal{H}, S}$ is **not** ϵ -exhausted, i.e., the set has good generalization (zero training error, test error to be less than ϵ), then one must feed in a very large m

1.3 Outside of version space

Hoeffding Inequality (mean version):

$$\begin{aligned}
\Pr(\bar{X} - \mu \geq \epsilon) &\leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\Pr(\mu - \bar{X} \geq \epsilon) &\leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)
\end{aligned} \tag{12}$$

using the second definition for Bernoulli random variable:

$$\begin{aligned}
\Pr(R(h) - \hat{R}_S(h) \geq \epsilon) &\leq \exp\left(-\frac{2m^2\epsilon^2}{\sum_{i=1}^m 1^2}\right) \\
&= \exp(-2m\epsilon^2) \\
\implies \Pr(\exists_{h \in \mathcal{H}} : R(h) - \hat{R}_S(h) \geq \epsilon) &\leq |\mathcal{H}| \exp(-2m\epsilon^2)
\end{aligned} \tag{13}$$

$$\begin{aligned}
\text{Let } |\mathcal{H}| \exp(-2m\epsilon^2) &\leq \delta \\
\implies -2m\epsilon^2 &\leq \log(\delta) - \log(|\mathcal{H}|) \\
\implies m &\geq \frac{\log(|\mathcal{H}|)}{2\epsilon^2} - \frac{\log(\delta)}{2\epsilon^2} \\
&= \frac{\log(|\mathcal{H}|)}{2\epsilon^2} + \frac{\log(1/\delta)}{2\epsilon^2}
\end{aligned} \tag{14}$$

2 PAC Bayes

2.1 Big picture

$$\mathcal{C}(\hat{R}_S(Q) \| R(Q)) \leq \frac{\text{KL}(Q \| P) + \log \left[\frac{1}{\delta} \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m \mathcal{C}(\hat{R}_S(h), R(h))} \right] \right]}{m} \quad (15)$$

The generalization error bound:

$$R(Q) \leq \hat{R}_S(Q) + \sqrt{\frac{\text{KL}(Q \| Q^0) + \log \left(\frac{2\sqrt{n}}{\delta} \right)}{2n}} \quad (16)$$

2.2 definition

1. \mathcal{X} input space
2. $\mathcal{Y} \in \{+1, -1\}$
3. \mathcal{D} be “true distribution” of input-output pair defined on $\mathcal{X} \times \mathcal{Y}$, such that one may

$$(x, y) \sim \mathcal{D} \quad (17)$$

4. $S^m \sim \mathcal{D}$ be the sampled data pair $\in \mathcal{X} \times \mathcal{Y}$, i.e., training data

2.2.1 $Q(h)$

let Q be a distribution defined over \mathcal{H} :

1. Expected risk over Q :

$$R(Q) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{E}_{h \sim Q} [l(h; (x, y))] \quad (18)$$

2. Empirical Risk over Q :

$$\hat{R}_S(Q) = \frac{1}{m} \sum_{(x,y) \in S} \mathbb{E}_{h \sim Q} [l(h; (x, y))] \quad (19)$$

3. in classification, typical $l(h; (x, y)) = \mathbb{1}_{h(x) \neq y}$
4. without the red bits, they are just ordinary empirical and expected risks
5. probability distributions occur in two places:
 - (a) Q encodes hypotheses
 - (b) \mathcal{D} describes randomness in the real-world

2.3 Theorem to bound PAC-Bayes

Theorem 2 with probability at least $1 - \delta$ over $S \sim \mathcal{D}$:

$$\underbrace{\mathcal{C}(\hat{R}_S(Q) \| R(Q))}_{\text{consistency for } Q} \leq \frac{\overbrace{KL(Q \| P)}^{\text{how similar is } Q \text{ and } P} + \log \left[\frac{1}{\delta} \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m \mathcal{C}(\hat{R}_S(h), R(h))} \right] \right]}{m} \quad (20)$$

$\mathcal{C}(\hat{R}_S(Q) \| R(Q))$ can be thought of how consistent is the performance between hypothesis from Q on both training($\hat{R}_S(Q)$) and testing($R(Q)$) data-set

2.3.1 notes on Theorem 2

1. **difference** between loss distribution using *sampled data* and *population/test data* is (i.e., consistency for $h \in Q$) is bounded by:

- (a) consistency for $h \in P$
- (b) similarity between P and Q ,

the bound is true $\forall Q$ defined over \mathcal{H} , i.e., posterior distribution on \mathcal{H}

2. for example, when consistency for $h \in P$ is good, and P and posterior Q are similar, then the consistency for $h \in Q$ is also good
3. when large amount of data is used, i.e., $m \rightarrow \infty$ the difference between $\hat{R}_S(Q)$ and $R(Q)$ is negligible
4. that Q need not be a Bayesian posterior, it can be **any** distribution

2.4 Proof

we use intermediate term:

$$f(S) = \mathbb{E}_{h \sim P} \exp^{m \mathcal{C}(\hat{R}_S(h), R(h))} \quad (21)$$

since $f(S)$ is a non-negative random variable (function of S), as $\exp(\cdot) > 0$, using Markov's inequality:

2.4.1 Markov's inequality

If X is a non-negative random variable and $a > 0$, then:

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad (22)$$

$$\Pr(f(S) \geq a) \leq \frac{\mathbb{E}[f(S)]}{a} \quad \text{let } f(S) \equiv X$$

letting:

$$\begin{aligned} \delta = \frac{\mathbb{E}[f(S)]}{a} &\implies a = \frac{\mathbb{E}[f(S)]}{\delta} \\ \implies \Pr\left(f(S) \geq \frac{\mathbb{E}[f(S)]}{\delta}\right) &\leq \delta \end{aligned} \quad (23)$$

substitute:

$$\begin{aligned} f(S) &\equiv \mathbb{E}_{h \sim P} \left[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))} \right] \\ \mathbb{E}[f(S)] &\equiv \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))} \right] \end{aligned} \quad (24)$$

$$\Pr\left(\mathbb{E}_{h \sim P} \left[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))} \right] > \frac{1}{\delta} \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))} \right]\right) \leq \delta \quad (25)$$

since $\log(\cdot)$ is monotonically increasing, it won't change inequality sign:

$$\Pr\left(\log\left(\mathbb{E}_{h \sim P} \left[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))} \right]\right) \leq \log\left(\frac{1}{\delta} \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))} \right]\right)\right) \geq 1 - \delta \quad (26)$$

2.5 find the lower bound of $\log(\mathbb{E}_{h \sim P}[f(h)])$

let \mathcal{H}_Q be support of Q i.e.,

$$h \in \mathcal{H}_Q \implies Q(h) > 0 \quad (27)$$

For any $g : \mathcal{H} \rightarrow \mathbb{R}$ and Q and P , we have:

$$\begin{aligned} \mathbb{E}_{h \sim P}[g(h)] &= \int_{\mathcal{H}} g(h) P(h) dh \\ &= \underbrace{\int_{\mathcal{H}_Q} g(h) P(h) dh}_{Q \text{ support}} + \underbrace{\int_{\mathcal{H} \setminus \mathcal{H}_Q} g(h) P(h) dh}_{\text{no } Q \text{ support}} \\ &= \int_{\mathcal{H}_Q} g(h) \frac{P(h)}{Q(h)} Q(h) dh + \int_{\mathcal{H} \setminus \mathcal{H}_Q} g(h) P(h) dh \quad \text{introduce } Q(h) \text{ to where it has support} \\ &\geq \mathbb{E}_{h \sim Q} \left[\frac{P(h)}{Q(h)} g(h) \right] \\ \implies \log \mathbb{E}_{h \sim P}[g(h)] &\geq \log \left[\mathbb{E}_{h \sim Q} \left[\frac{p(h)}{Q(h)} g(h) \right] \right] \\ &\geq \mathbb{E}_{h \sim Q} \left[\log \left[\frac{P(h)}{Q(h)} g(h) \right] \right] \\ &= \mathbb{E}_{h \sim Q} \left[\log \left[\frac{P(h)}{Q(h)} \right] \right] + \mathbb{E}_{h \sim Q} [\log [g(h)]] \\ &= -\text{KL}(Q \| P) + \mathbb{E}_{h \sim Q} [\log [g(h)]] \end{aligned} \quad (28)$$

2.5.1 back to proof

substitute $g(h) = \exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}$:

$$\begin{aligned} \log(\mathbb{E}_{h \sim P}[g(h)]) &\geq -\text{KL}(Q\|P) + \mathbb{E}_{h \sim Q}[\log[g(h)]] \\ \implies \log(\mathbb{E}_{h \sim P}[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}]) &\geq -\text{KL}(Q\|P) + \mathbb{E}_{h \sim Q}[\log[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}]] \\ &= -\text{KL}(Q\|P) + m\mathbb{E}_{h \sim Q}[\mathcal{C}(\hat{R}_S(h), R(h))] \end{aligned} \quad (30)$$

inequality automatically applies to the lower bound with **higher** probability:

$$\begin{aligned} \Pr(\log(\mathbb{E}_{h \sim P}[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}]) &\leq \log(\frac{1}{\delta}\mathbb{E}_{S \sim \mathcal{D}}\mathbb{E}_{h \sim P}[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}])) \geq 1 - \delta \\ \implies \Pr(-\text{KL}(Q\|P) + m\mathbb{E}_{h \sim Q}[\mathcal{C}(\hat{R}_S(h), R(h))] &\leq \log(\frac{1}{\delta}\mathbb{E}_{S \sim \mathcal{D}}\mathbb{E}_{h \sim P}[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}])) \geq 1 - \delta \\ \implies \Pr(\mathbb{E}_{h \sim Q}[\mathcal{C}(\hat{R}_S(h), R(h))] &\leq \frac{1}{m}\{\text{KL}(Q\|P) + \log[\frac{1}{\delta}\mathbb{E}_{S \sim \mathcal{D}}\mathbb{E}_{h \sim P}[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}]]\}) \geq 1 - \delta \end{aligned} \quad (31)$$

therefore, with probability of at least $1 - \delta$ and $\forall Q$ on \mathcal{H} :

$$\mathcal{C}(\hat{R}_S(Q), R(Q)) \leq \frac{\text{KL}(Q\|P) + \log[\frac{1}{\delta}\mathbb{E}_{S \sim \mathcal{D}}\mathbb{E}_{h \sim P}[\exp^{m\mathcal{C}(\hat{R}_S(h), R(h))}]]}{m} \quad (32)$$

2.6 example of: $\mathcal{C}(\hat{R}_S(h), R(h))$

The term $\mathcal{C}(\hat{R}_S(h), R(h))$ is measuring the consistency between $R(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[h(x,y)]$ and its discrete sample approximation $\hat{R}_S = \frac{1}{n} \sum_{i=1}^n h(x_i, y_i)$ one of the most used choice of:

$$\begin{aligned} \mathcal{C}(\hat{R}_S(h), R(h)) &\equiv \text{KL}(\text{Bernoulli}(\hat{R}_S(h)), \text{Bernoulli}(R(h))) \\ &= \sum_{x \in \{0,1\}} (p(\hat{R}_S(h) = x)) \log\left(\frac{p(\hat{R}_S(h) = x)}{p(R(h) = x)}\right) \\ &= (p(\hat{R}_S(h) = 1)) \log\left(\frac{p(\hat{R}_S(h) = 1)}{p(R(h) = 1)}\right) + (p(\hat{R}_S(h) = 0)) \log\left(\frac{p(\hat{R}_S(h) = 0)}{p(R(h) = 0)}\right) \\ &= \hat{R}_S(h) \log\left(\frac{\hat{R}_S(h)}{R(h)}\right) + (1 - \hat{R}_S(h)) \log\left(\frac{1 - \hat{R}_S(h)}{1 - R(h)}\right) \end{aligned} \quad (33)$$

which say, instead of measure directly the difference between $\hat{R}_S(h)$, $R(h)$, we measure the KL between Bernoulli distribution using $\hat{R}_S(h)$, $R(h)$ as parameters. By substitution:

$$\begin{aligned}
& \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m \mathcal{C}(\hat{R}_S(h), R(h))} \right] \\
&= \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m \left[\hat{R}_S(h) \log \left(\frac{\hat{R}_S(h)}{R(h)} \right) + (1 - \hat{R}_S(h)) \log \left(\frac{1 - \hat{R}_S(h)}{1 - R(h)} \right) \right]} \right] \\
&= \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\left(\frac{\hat{R}_S(h)}{R(h)} \right)^{m \hat{R}_S(h)} \left(\frac{1 - \hat{R}_S(h)}{1 - R(h)} \right)^{m(1 - \hat{R}_S(h))} \right] \\
&= \mathbb{E}_{h \sim P} \underbrace{\mathbb{E}_{S \sim \mathcal{D}} \left[\left(\frac{\hat{R}_S(h)}{R(h)} \right)^{m \hat{R}_S(h)} + \left(\frac{1 - \hat{R}_S(h)}{1 - R(h)} \right)^{m(1 - \hat{R}_S(h))} \right]}_{\text{swap two expectations}} \quad (34)
\end{aligned}$$

the term $\hat{R}_S(h) = \frac{1}{m} \sum_{(x,y) \in S} l(h; (x,y))$ here, sample S is given/fixed

$$\begin{aligned}
&= \frac{\text{number of times } l(h; (x,y)) = 1}{m} \\
&\in \left\{ 0, \frac{1}{m}, \frac{2}{m}, \dots, 1 \right\} \quad \text{of course, each of their probabilities is different} \quad (35)
\end{aligned}$$

2.6.1 consider only: $\mathbb{E}_{S \sim \mathcal{D}} \left[\left(\frac{\hat{R}_S(h)}{R(h)} \right)^{m \hat{R}_S(h)} + \left(\frac{1 - \hat{R}_S(h)}{1 - R(h)} \right)^{m(1 - \hat{R}_S(h))} \right]$

instead of summing all combination of $\sum_{S \sim \mathcal{D}}$, we change the expectation variables to be $\hat{R}_S(h)$, i.e., $\mathbb{E}_{\hat{R}_S(h) \sim \text{Binomial}(m, R(h))}[\cdot]$. Instead of taking expectation over $S \sim \mathcal{D}$, we only have finite number of different $\hat{R}_S(h)$ values:

$$= \sum_{k=0}^m \underbrace{\binom{m}{k} R(h)^k (1 - R(h))^{m-k}}_{p(\hat{R}_S(h) = \frac{k}{m})} \underbrace{\left(\frac{\frac{k}{m}}{R(h)} \right)^{m \frac{k}{m}} \left(\frac{1 - \frac{k}{m}}{1 - R(h)} \right)^{m(1 - \frac{k}{m})}}_{p(f(\hat{R}_S(h)) \mid \hat{R}_S(h) = \frac{k}{m})} \quad (36)$$

$$\begin{aligned}
&= \sum_{k=0}^m \binom{m}{k} R(h)^k (1 - R(h))^{m-k} \left(\frac{\frac{k}{m}}{R(h)} \right)^k \left(\frac{1 - \frac{k}{m}}{1 - R(h)} \right)^{m-k} \\
&= \sum_{k=0}^m \binom{m}{k} \left(\frac{k}{m} \right)^k \left(1 - \frac{k}{m} \right)^{m-k} \quad (38)
\end{aligned}$$

that's fantastic, as it contains no $\hat{R}_S(h)$ nor $R(h)$, i.e., no h

$$\mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m \mathcal{C}(\hat{R}_S(h), R(h))} \right] = \sum_{k=0}^m \underbrace{\binom{m}{k} \left(\frac{k}{m} \right)^k \left(1 - \frac{k}{m} \right)^{m-k}}_{\leq 1} \leq m + 1 \quad (39)$$

note that unlike:

$$\sum_{k=0}^m \binom{m}{k} p^k (1-p)^{m-k} = 1 \quad (40)$$

however,

$$\sum_{k=0}^m \binom{m}{k} \underbrace{\left(\frac{k}{m} \right)^k}_{\text{variable}} \left(1 - \frac{k}{m} \right)^{m-k} \neq 1 \quad (41)$$

by substitution:

$$\begin{aligned} \mathcal{C}(\hat{R}_S(Q), R(Q)) &\leq \frac{\text{KL}(Q \| P) + \log \left[\frac{1}{\delta} \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{h \sim P} \left[\exp^{m \mathcal{C}(\hat{R}_S(h), R(h))} \right] \right]}{m} \\ &= \frac{\text{KL}(Q \| P) + \log \left(\frac{m+1}{\delta} \right)}{m} \end{aligned} \quad (42)$$

2.7 lower bound of $\mathcal{C}(\hat{R}_S(Q), R(Q))$

When consider risk function to be $\mathcal{C}(\hat{R}_S(h), R(h)) \equiv \text{KL}(\text{Bernoulli}(\hat{R}_S(h)), \text{Bernoulli}(R(h)))$, Eq.(42) gives:

$$\begin{aligned} \mathcal{C}(\hat{R}_S(Q), R(Q)) &\equiv \text{KL}(\text{Ber}(\hat{R}_S(Q)) \| \text{Ber}(R(Q))) \\ &\leq \frac{\text{KL}(Q \| P) + \log \left(\frac{m+1}{\delta} \right)}{m} \end{aligned} \quad (43)$$

obviously, $\text{KL}(\text{Ber}(\hat{R}_S(Q)) \| \text{Ber}(R(Q)))$ are not useful. we can not disentangle between $\hat{R}_S(Q)$ and $R(Q)$.

We hope to bring in its lower bound in terms of $R(Q) - \hat{R}_S(Q)$, then we can just leave $R(Q)$ alone in the LHS. Therefore, anything on the RHS becoes the upper-bound of $R(Q)$ we can **minimize**

2.7.1 tighter bound $\frac{m+1}{\delta} \rightarrow \frac{2\sqrt{n}}{\delta}$

with a tighter bound, we can have $\frac{m+1}{\delta} \rightarrow \frac{2\sqrt{n}}{\delta}$:

$$\text{KL}(\text{Ber}(\hat{R}_S(Q)) \parallel \text{Ber}(R(Q))) \leq \frac{\text{KL}(Q \parallel P) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n} \quad (44)$$

2.8 Pinsker's inequality converts KL back to its argument

2.8.1 tighter version:

$$\begin{aligned} \text{KL}(\hat{p} \parallel p) &\geq \frac{(p - \hat{p})^2}{2p} \\ \implies p - \hat{p} &\leq \sqrt{2p \text{KL}(\text{Ber}(\hat{p}) \parallel \text{Ber}(p))} \\ \implies R(Q) - \hat{R}_S(Q) &\leq \sqrt{2R(Q) \text{KL}(\text{Ber}(\hat{R}_S(Q)) \parallel \text{Ber}(R(Q)))} \quad \hat{p} \rightarrow \hat{R}_S(Q) \quad p \rightarrow R(Q) \end{aligned} \quad (45)$$

substitute: Eq.(45)

$$\begin{aligned} R(Q) - \hat{R}_S(Q) &\leq \sqrt{2R(Q) \text{KL}(\text{Ber}(\hat{R}_S(Q)) \parallel \text{Ber}(R(Q)))} \leq \sqrt{2R(Q) \frac{\text{KL}(Q \parallel Q^0) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} \\ \implies R(Q) &\leq \hat{R}_S(Q) + \sqrt{2R(Q) \frac{\text{KL}(Q \parallel Q^0) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} \quad \text{remove the middle inequality} \end{aligned} \quad (46)$$

Bound is tight when population risk $R(Q)$ is smaller, because of $\sqrt{R(Q)}$. This form of expression is only for showcase such tight bound, it is however not useful in practice. You can **not** express in the form where $R(Q)$ **appears** in the right.

2.8.2 loose version:

We need another form of looser version of "Pinsker's inequality" that does not require to have $R(Q)$ on the RHS:

$$\begin{aligned} \text{KL}(\text{Ber}(\hat{p}) \parallel \text{Ber}(p)) &\geq 2(p - \hat{p})^2 \\ \implies \sqrt{\text{KL}(\text{Ber}(\hat{p}) \parallel \text{Ber}(p))} &\geq \sqrt{2}(p - \hat{p}) \\ \implies \sqrt{2}p &\leq \sqrt{\text{KL}(\text{Ber}(\hat{p}) \parallel \text{Ber}(p))} + \sqrt{2}\hat{p} \\ \implies p &\leq \sqrt{\frac{\text{KL}(\text{Ber}(\hat{p}) \parallel \text{Ber}(p))}{2}} + \hat{p} \end{aligned} \quad (47)$$

substitution from Eq.(45):

$$\begin{aligned}
\Rightarrow R(Q) &\leq \hat{R}_S(Q) + \sqrt{\frac{\text{KL}(\text{Ber}(\hat{R}_S(Q)) \parallel \text{Ber}(R(Q)))}{2}} \\
&\leq \hat{R}_S(Q) + \sqrt{\frac{\text{KL}(Q \parallel Q^0) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}} \\
&\stackrel{\text{substitute}}{\leq} \text{KL}(\text{Ber}(\hat{R}_S(Q)) \parallel \text{Ber}(R(Q))) \leq \frac{\text{KL}(Q \parallel P) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}
\end{aligned}
\tag{48}$$