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 $|S| = m$:

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

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:

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

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

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 $\mathcal{H}$.

Think $\mathcal{C}$ may be some polynomial function, but $\mathcal{H}$ is what the model we propose to apply, say linear.

1.2.1 bound amount of over-fitting

We are interested to compute:

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

1.2.2 Version space

1. **definition: Version space (VS)**

$$\mathrm{VS}_{\mathcal{H},S} \equiv \{\forall h \in \mathcal{H} \quad | \; \hat{R}_S(h) = 0\} \tag{5}$$

2. **definition: ϵ-exhausted version space**

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

$$\{\forall h \in \mathrm{VS}_{\mathcal{H},S} \quad | \; R(h) \leq \epsilon\} \tag{6}$$

   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 $VS_{\mathcal{H},S}$ is **not** *ϵ-exhausted is at most:*

$$|\mathcal{H}| \exp^{-\epsilon m} \tag{7}$$

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 $\mathrm{VS}_{\mathcal{H},S}$ (clearly, by including h_{bad}, $\mathrm{VS}_{\mathcal{H},S}$ is **not** ϵ-exhausted), then, it must classify all m data in S correctly:

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

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

$$\implies \mathbf{Pr}\Big(\exists_{h\in\mathcal{H}} : \big(\hat{R}_S(h) = 0\big)\Big) \leq |\mathcal{H}| \exp^{-\epsilon m} \tag{9}$$

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

1.2.4 what does tell you about m?

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

let $|S| = m$:

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

Say we fix ϵ, then if one desires to have a very small chance (i.e., set δ to be very small) that the $\text{VS}_{\mathcal{H},S}$ 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):

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

using the second definition for Bernoulli random variable:

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

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

2 PAC Bayes

2.1 Big picture

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

The generalization error bound:

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

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

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

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:

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

2. Empirical Risk over Q:

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

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* $S \sim \mathcal{D}$*:*

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

$\mathcal{C}\big(\hat{R}_S(Q)\|R(Q)\big)$ 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., $m \to \infty$ 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:

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

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:

$$\begin{aligned} \mathbf{Pr}(X \geq a) &\leq \frac{\mathbb{E}[X]}{a} \\ \mathbf{Pr}(f(S) \geq a) &\leq \frac{\mathbb{E}[f(S)]}{a} \quad \text{let } f(S) \equiv X \end{aligned} \tag{22}$$

letting:

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

substitute:

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

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

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

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

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.,

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

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

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

note that the above is just standard variational Bayes, if we have:

$$\begin{aligned} P(h) \to p(z) \quad Q(h) \to q(z|x) \quad g(h) \to p(x|z) \\ \implies \log \mathbb{E}_{z\sim p(z)}[p(x|z)] \geq -\mathrm{KL}(q(z|x)\|p(z)) + \mathbb{E}_{z\sim q(z|x)}[\log(p(x|z)] \end{aligned} \tag{29}$$

2.5.1 back to proof

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

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

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

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

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

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

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

The term $\mathcal{C}\big(\hat{R}_S(h),R(h)\big)$ 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:

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

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:

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

$$
\begin{aligned}
\text{the term } \hat{R}_S(h) &= \frac{1}{m}\sum_{(x,y)\in S} l(h;(x,y)) \quad \text{here, sample } S \text{ is given/fixed} \\
&= \frac{\text{number of times } l(h;(x,y))=1}{m} \\
&\in \left\{0,\frac{1}{m},\frac{2}{m},\ldots,1\right\} \quad \text{of course, each of their probabilities is different}
\end{aligned} \tag{35}
$$

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:

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

assume that under the same hypothesis h:

$$
\begin{aligned}
\mathbf{Pr}_{(x,y)\sim\mathcal{D}}(l(h;(x,y))=1) &= \mathbf{Pr}_{(x,y)\sim\mathcal{D}}(h(x)\neq y) \\
&= R(h)
\end{aligned} \tag{37}
$$

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

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

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

note that unlike:

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

however,

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

by substitution:

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

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

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

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

obviously, $\mathrm{KL}\big(\mathrm{Ber}(\hat{R}_S(Q))\|\mathrm{Ber}(R(Q))\big)$ 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 $\frac{m+1}{\delta} \to \frac{2\sqrt{n}}{\delta}$

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

$$\mathrm{KL}\big(\mathrm{Ber}(\hat{R}_S(Q))\|\mathrm{Ber}(R(Q)\big) \leq \frac{\mathrm{KL}(Q\|P) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n} \tag{44}$$

need to check its literature later

2.8 Pinsker's inequality converts KL back to its argument

2.8.1 tighter venison:

$$\begin{aligned} \mathrm{KL}(\hat{p}\|p) &\geq \frac{(p-\hat{p})^2}{2p} \\ \implies p - \hat{p} &\leq \sqrt{2p\mathrm{KL}(\mathrm{Ber}(\hat{p})\|\mathrm{Ber}(p))} \\ \implies R(Q) - \hat{R}_S(Q) &\leq \sqrt{2R(Q)\mathrm{KL}\big(\mathrm{Ber}(\hat{R}_S(Q))\|\mathrm{Ber}(R(Q)\big)} \quad \hat{p} \to \hat{R}_S(Q) \quad p \to R(Q) \end{aligned} \tag{45}$$

substitute: Eq.(45)

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

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

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

substitution from Eq.(45):

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