3.4-3.5 Generalization bounds via uniform convergence and Concentration inequalities

Theorem 4 is based on the assumptions of finite hypothesis and realizability, now we want to break free of these restrictive assumptions.

$$L(\hat{h}) - L(h^*) = \underbrace{[L(\hat{h}) - \hat{L}(\hat{h})]}_{\text{Concentration}} + \underbrace{[\hat{L}(\hat{h}) - \hat{L}(h^*)]}_{\leq 0} + \underbrace{[\hat{L}(h^*) - L(h^*)]}_{\text{Concentration}}$$

Note: (i) h^* is non-random, so the third term is simple.

(ii) $\hat{h}$ is r.v. w.r.t. training examples, so the first term is not a sum of i.i.d. r.v..

The contrapositive can be write as:

$$\underbrace{P\{L(\hat{h}) - L(h^*) \geq \varepsilon\}}_{\text{excess risk}} \leq P\{\underbrace{\sup_{h \in \mathcal{H}} |L(h) - \hat{L}(h)| \geq \frac{\varepsilon}{2}}_{\text{uniform convergence}}\}$$

3.5 Concentration inequalites

-Mean estimation

Let $X_1, \cdots, X_n$ be i.i.d. real-valued r.v. with mean $\mu := E[X_i]$, define $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. The question is: How does $\hat{\mu}_n$ relate to μ?

-Types of statements

(i) Consistency: by the law of large number,

$$\hat{\mu}_n - \mu \xrightarrow{P} 0$$

(ii) Asymptotic normality: Letting $Var[X_i] = \sigma^2$, by CLT:

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

(iii) Tail bounds: Ideally, we want such a statement:

$$P\{|\hat{\mu}_n - \mu| \geq \varepsilon\} \leq \text{SomeFunction}(n, \varepsilon) = \delta$$

Based on (ii), we prefer an exponential decay.

Typical technique:

[Theorem 5] Markov's inequality

Let $Z > 0$ be a random variable, then

$$P[Z \geq t] \leq \frac{E[Z]}{t}$$

Remarks: (i) We can apply $Z = (X - \mu)^2$ and $t = \varepsilon^2$ to obtain
Second moment
Chebyshev's inequality:

$$P\{|X-\mu| \geq \varepsilon\} \leq \frac{Var[X]}{\varepsilon^2}$$

Applying it to $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (X_i are i.i.d.), then $Var[\hat{\mu}_n] = \frac{Var[X_1]}{n}$, which decays at a rate $O(\frac{1}{n})$.

To get stronger bounds, we need to apply Markov's inequality on higher order moments. In particular, we consider all moments by $Z = e^{tX}$, where t is a free paramter to optimize the bound.

[Definition 6] moment generating function.

For a r.v. X, the moment generating function (MGF) of X is:

$$M_X(t) := E[e^{tX}]$$

Note: (i) $M_X(t) = 1 + tE[X] + \frac{t^2}{2}E[X^2] + \cdots$

(ii) $\frac{d^k M_X(t)}{dt^k}\Big|_{t=0} = E[X^k]$

(iii) If X_1 and X_2 are independent r.v., then

$$M_{X_1+X_2} = M_{X_1} M_{X_2}$$

Applying Markov's inequality to $Z = e^{tX}$:

$$P\{X \geq \varepsilon\} \leq e^{-t\varepsilon} M_X(t) \quad \text{for all } t > 0 \quad \cdots (172)$$

For $X = \hat{\mu}_n$, by computing $P[\hat{\mu}_n \geq \varepsilon] = P[X_1 + \cdots + X_n \geq n\varepsilon]$,

$$P[\hat{\mu}_n \geq \varepsilon] \leq (e^{-t\varepsilon} M_{X_1}(t))^n$$

We will work with X s.t. $M_{X_1}(t) < \infty$ for all $t > 0$.

[Example 5] MGF of Gaussian variables.

Let $X \sim N(0, \sigma^2)$, Then $M_X(t) = e^{\sigma^2 t^2/2}$.

This is because:

$$M_X(t) = E[e^{tX}] = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} + tx\right) dx$$
$$= \int (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\sigma^2 t)^2 - \sigma^4 t^2}{2\sigma^2}\right) dx$$
$$= \exp\left(\frac{\sigma^2 t^2}{2}\right) \qquad \square$$

[Lemma 3] Tail bound for Gaussian variables.

$$P[X \geq \varepsilon] \leq \inf_t \exp\left\{\frac{\sigma^2 t^2}{2} - t\varepsilon\right\}$$

which is the corollary of (172). Setting $t = \varepsilon/\sigma^2$, we have

$$P[X \geq \varepsilon] \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

[Definition 7] sub-Gaussian

A mean-zero r.v. X is sub-Guassian with parameter σ^2 if:

$$M_X(t) \leq \exp\left(\frac{\sigma^2 t^2}{2}\right)$$

It follows immediately that $P[X \geq \varepsilon] \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$ ······(180)

Bounded random variables (Hoeffding's lemma):

If $a \leq X \leq b$ with prob. 1 and $E[X] = 0$, then X is sub-Guassian with $\sigma^2 = (b-a)^2/4$.

Pf: $e^{tx} \leq \frac{x-a}{b-a} e^{tb} + \frac{b-x}{b-a} e^{ta}$ (convexity of e^{hx})

$$\Rightarrow E(e^{tx}) \leq -\frac{a}{b-a} e^{tb} + \frac{b}{b-a} e^{ta}$$
$$= p e^{(1-p)y} + (1-p) e^{-py}$$

$$= e^{-py}(1-p+pe^{y})$$
$$=: e^{f(y)}$$

where $p = -\frac{a}{b-a}$, $y = (b-a)t$,

$$f(y) = -py + \ln(1-p+pe^{y}), \quad f(0)=0$$

$$f'(y) = -p + \frac{pe^{y}}{1-p+pe^{y}} = -p + \frac{p}{p+(1-p)e^{-y}}, \quad f'(0)=0$$

$$f''(y) = \frac{p(1-p)e^{-y}}{(p+(1-p)e^{-y})^2} = \frac{p(1-p)}{p^2e^{y}+(1-p)^2e^{-y}+2p(1-p)} \le \frac{1}{4}$$

By Taylor expension, we have

$$f(y) = f(0) + f'(0)y + \frac{1}{2}f''(\theta y)y^2 \le \frac{1}{8}y^2$$

Then we have

$$E[e^{tX}] \le e^{\frac{1}{8}y^2} = e^{\frac{1}{8}(b-a)^2t^2} = e^{\frac{\sigma^2t^2}{2}}$$

where $\sigma^2 = \frac{1}{4}(b-a)^2$ $\square$

Properties:

1. Sum: X_1, X_2 independent sub-Guassian r.v. with σ_1^2 and σ_2^2, then X_1+X_2 sub-Guassian with $\sigma_1^2+\sigma_2^2$.

2. Multiplication by a constant: If X_1 sub-Guassian with σ^2, then for any $c>0$, cX sub-Guassian with $c^2\sigma^2$.

[Theorem 6] (Hoeffding's inequality)

Let $X_1, \cdots, X_n$ be independent r.v., $a_i \le X_i \le b_i$,

Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, Then

$$P[\hat{\mu}_n \ge E[\hat{\mu}_n] + \varepsilon] \le \exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

Pf: $\hat{\mu}_n - E_n[\hat{\mu}_n]$ is sub-Guassian with parameter $\frac{1}{n^2}\sum_{i=1}^{n}\frac{(b_i-a_i)^2}{4}$

Then by (180) we have

$$P[\hat{\mu}_n - E_n[\hat{\mu}_n] \ge \varepsilon] \le \exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}$$ $\square$