

1 Generalization Error Bounds and PAC Learning

Although the concept of consistency of a learning algorithm is very important, it only measures how the expectation of a random variable $R(\hat{h}_n)$ converges to the optimal Bayes risk asymptotically. However, it does not say how fast this convergence is, and neither does it tell us how this random variable is distributed. In particular, we are interested in probability bounds of the generalization error, such as the following: “with probability at least $1 - \delta$, the risk $R(\hat{h}_n)$ is bounded by some quantity”.

Recall that the excess risk can be decomposed into approximation error and estimation error, i.e.

$$R(\hat{h}_n) - R^* = \underbrace{\left(\inf_{h \in \mathcal{H}} R(h) - R^* \right)}_{\text{approximation error}} + \underbrace{\left(R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \right)}_{\text{estimation error}}.$$

The approximation error is deterministic and mainly caused by two possible reasons: (1) the restriction of using the function class \mathcal{H} ; (2) if $\inf_{h \in \mathcal{H}} R(h)$ in the above equation were replaced by the minimum risk achievable by the learning algorithm with infinite amount of data, then it can also be caused by the systematic bias of the learning algorithm. Such an error is often not controllable as we do not know the underlying distribution $P_{X,Y}$. On the other hand, the estimation error depends on the sample size, the function class \mathcal{H} and the learning algorithm which we have control over. We would like to obtain probability bounds for the estimation error.

Specifically, if we use ERM to obtain our predictor $\hat{h}_n = \arg \min_{h \in \mathcal{H}} \hat{R}_n(h)$, and assume that $\inf_{h \in \mathcal{H}} R(h) = R(h^*)$ for some $h^* \in \mathcal{H}$, then we have

$$\begin{aligned} R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) &= R(\hat{h}_n) - R(h^*) \\ &\leq R(\hat{h}_n) - R(h^*) + \hat{R}_n(h^*) - \hat{R}_n(\hat{h}_n) \\ &= (R(\hat{h}_n) - \hat{R}_n(\hat{h}_n)) - (R(h^*) - \hat{R}_n(h^*)) \\ &\leq 2 \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_n(h) \right|. \end{aligned}$$

Thus if we can obtain uniform bound of $\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)|$ then the approximation error can be bounded. Thus again justifies the usage of the ERM method.

Intuitively, for any $h \in \mathcal{H}$, $\hat{R}_n(h)$ is a random variable which follows (in fact $n\hat{R}_n(h)$) a Binomial distribution with mean $R(h)$. Or we could think of it as the average of a series of random variables. Thus we should be able to bound the difference between an average of a set of random variables and their mean. The uniform bound, however, will depend crucially on how large/complex the hypothesis space \mathcal{H} is.

The *probably approximately correct* (PAC) learning model typically states as follows: we say that \hat{h}_n is ϵ -accurate with probability $1 - \delta$, if

$$P \left(R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) > \epsilon \right) < \delta.$$

In other words, we have $R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \leq \epsilon$ with probability at least $(1 - \delta)$.

2 Concentration Inequalities

Concentration inequalities will be used to measure how fast the empirical risk converges to the true risk. We start with some loose but simple ones and then get to more useful results.

Theorem 3-1 (Markov Inequality). For any nonnegative random variable X and $\epsilon > 0$,

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

PROOF. We have

$$\mathbb{E}[X] \geq \mathbb{E}[I(X \geq \epsilon)X] \geq \epsilon \mathbb{E}[I(X \geq \epsilon)] = \epsilon P(X \geq \epsilon)$$

and thus $P(X \geq \epsilon) \leq \mathbb{E}[X]/\epsilon$. \square

Theorem 3-2 (Chernoff Inequality) For any random variable X and $\epsilon > 0$,

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t\epsilon)}$$

and thus

$$P(X \geq \epsilon) \leq \inf_{t>0} \frac{\mathbb{E}[\exp(tX)]}{\exp(t\epsilon)}.$$

PROOF. For any $t > 0$, since $\exp(tx)$ is a nonnegative monotone increasing function in x , we have

$$P(X \geq \epsilon) = P(\exp(tX) \geq \exp(t\epsilon)) \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t\epsilon)}.$$

\square

Theorem 3-3. (Chebyshev Inequality) For any random variable X and $\epsilon > 0$,

$$P(|X - \mathbb{E}[X]| > \epsilon) \leq \frac{\mathbb{V}[X]}{\epsilon^2}.$$

PROOF. Apply Markov Inequality with random variable $Y = |X - \mathbb{E}[X]|$. \square

Both Markov and Chebyshev bounds are polynomial in $1/\epsilon$ and often we need bounds which can converge to zero exponentially fast. In fact, the Chebyshev inequality can be quite poor. Consider the following example: Let binary iid random variables $X_1, \dots, X_n \in \{0, 1\}$ and $p = P(X_i = 1)$. Then we have $\sigma^2 := \mathbb{V}[X_i] = p(1-p)$. Define $S_n = \sum_{i=1}^n X_i$ and we have $\mathbb{E}[S_n] = np$ and $\mathbb{V}[S_n] = np(1-p) = n\sigma^2$. From Chebyshev inequality by using $\tilde{\epsilon} = n\epsilon$, we have

$$P\left(\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \geq \tilde{\epsilon}\right) \leq \frac{\sigma^2}{n\tilde{\epsilon}^2}.$$

Thus the tail probability goes to zero at a rate of n^{-1} . But from the *central limit theorem* (CLT) we have

$$\sqrt{n}\left(\frac{1}{n}S_n - \frac{1}{n}\mathbb{E}[S_n]\right) \rightarrow_d \mathcal{N}(0, \sigma^2).$$

In other words, we have

$$P\left(\sqrt{\frac{n}{\sigma^2}}\left(\frac{S_n}{n} - p\right) \geq y\right) \rightarrow 1 - \Phi(y) = \int_y^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx \leq \frac{1}{\sqrt{2\pi}} \frac{\exp(-y^2/2)}{y}.$$

So

$$P\left(\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n} \geq \tilde{\epsilon}\right) = P\left(\sqrt{\frac{n}{\sigma^2}}\left(\frac{S_n}{n} - p\right) \geq \sqrt{\frac{n}{\sigma^2}}\tilde{\epsilon}\right) \approx \exp\left(-\frac{n\tilde{\epsilon}^2}{2p(1-p)}\right)$$

which decreases exponentially fast as a function of n . So the Chebyshev inequality does poorly in this case and we need something better.

The Hoeffding's inequality studies the concentration on the sum of independent random variables and gives an exponential tail bound. Given random variables X_1, \dots, X_n which are independent, and let $S_n = \sum_{i=1}^n X_i$. By Chernoff bound we have

$$\begin{aligned} P(S_n - \mathbb{E}[S_n] \geq \epsilon) &= P(\exp(t(S_n - \mathbb{E}[S_n])) \geq \exp(t\epsilon)) \\ &\leq \exp(-t\epsilon)\mathbb{E}[\exp(t(S_n - \mathbb{E}[S_n]))] \\ &= \exp(-t\epsilon)\mathbb{E}\left[\exp\left(t\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)\right] \\ &= \exp(-t\epsilon)\prod_{i=1}^n \mathbb{E}[\exp(t(X_i - \mathbb{E}[X_i]))]. \end{aligned}$$

The following lemma shows some property of a bounded random variable with mean zero.

Lemma 3-4. *If random variable X has mean zero, i.e. $\mathbb{E}[X] = 0$, and is bounded in $[a, b]$, then for any $s > 0$,*

$$\mathbb{E}[\exp(sX)] \leq \exp(s^2(b-a)^2/8).$$

PROOF.

By convexity of exponential function and Jensen's inequality and the fact $a \leq X \leq b$,

$$\exp(sX) \leq \frac{X-a}{b-a} \exp(sb) + \frac{b-X}{b-a} \exp(sa).$$

Taking expectation on both sides, and utilizing the fact that $\mathbb{E}[X] = 0$, we have

$$\begin{aligned} \mathbb{E}[\exp(sX)] &\leq \frac{b \exp(sa) - a \exp(sb)}{b-a} \\ &= [1 - \lambda + \lambda \exp(s(b-a))] \exp(-\lambda s(b-a)) \end{aligned}$$

where $\lambda = -\frac{a}{b-a}$. Now let $u = s(b-a)$ and define

$$\phi(u) := -\lambda u + \log(1 - \lambda + \lambda \exp(u)),$$

then the above inequality becomes

$$\mathbb{E}[\exp(sX)] \leq \exp(\phi(u)).$$

Now we need to find an upper bound on $\exp(\phi(u))$. Using Taylor's expansion we have

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\xi)$$

for some $\xi \in [0, u]$. It is easy to verify that $\phi(0) = 0$ and $\phi'(0) = 0$. And we have

$$\begin{aligned} \phi''(u) &= \frac{\lambda \exp(u)}{1 - \lambda + \lambda \exp(u)} - \frac{(\lambda \exp(u))^2}{(1 - \lambda + \lambda \exp(u))^2} \\ &= \frac{\lambda \exp(u)}{1 - \lambda + \lambda \exp(u)} \left(1 - \frac{\lambda \exp(u)}{1 - \lambda + \lambda \exp(u)}\right) \\ &\leq \frac{1}{4}. \end{aligned}$$

So we have $\phi(u) \leq u^2/8$, and therefore

$$\mathbb{E}[\exp(sX)] \leq \exp(s^2(b-a)^2/8).$$

□

Now we are ready to present the Hoeffding's inequality.

Theorem 3-5 (Hoeffding Inequality) Let X_1, \dots, X_n be independent bounded random variables such that $X_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$, we have:

1. $P(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$
2. $P(S_n - \mathbb{E}[S_n] \leq -\epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$
3. $P(|S_n - \mathbb{E}[S_n]| \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$.

PROOF. By the above derivation and Lemma 3-4 we have

$$\begin{aligned} P(S_n - \mathbb{E}[S_n] \geq \epsilon) &\leq \exp(-t\epsilon) \prod_{i=1}^n \mathbb{E}[\exp(t(X_i - \mathbb{E}[X_i]))] \\ &\leq \exp(-t\epsilon) \exp\left(\sum_{i=1}^n \frac{t^2(b_i - a_i)^2}{8}\right) \end{aligned}$$

Now choose $t = \frac{4\epsilon}{\sum_{i=1}^n (b_i - a_i)^2}$ we have

$$P(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Similarly we can prove the other two claims.

□

If we apply the Hoeffding inequality to the average of a series of bernoulli random variables X_1, \dots, X_n , we have

$$P(S_n/n - p \geq \epsilon) \leq \exp(-2n\epsilon^2)$$

since $b_i - a_i = 1$. This is exactly what the CLT indicates when $p = 1/2$. The following is a straightforward application of the Hoeffding inequality:

Corollary 3-6 Assume that $\mathcal{H} = \{h_1, \dots, h_m\}$. Then for all $\epsilon > 0$,

$$P\left(\sup_{h \in \mathcal{H}} \left| \hat{R}_n(h) - R(h) \right| \geq \epsilon\right) \leq 2m \exp(-2n\epsilon^2),$$

for any distribution $P_{X,Y}$, where $R(h) = \mathbb{E}_{X,Y}[I(Y \neq h(X))]$ and $\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n I(y_i \neq h(x_i))$.

Finally we will introduce the McDiarmid inequality which generalizes the Hoeffding's inequality to some function of iid random variables. Some restrictions are needed in order to get exponential bounds.

Theorem 3-6 (McDiarmid Inequality/Bounded Differences) Suppose random variables $X_1, \dots, X_n \in \mathcal{X}$ are independent, f is a mapping from \mathcal{X}^n to \mathbb{R} . If for any i and any $x_1, \dots, x_n, x'_i \in \mathcal{X}$, f satisfies

$$|f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i.$$

Then for all $\epsilon > 0$,

$$\begin{aligned} P(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq \epsilon) &\leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \\ P(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \leq -\epsilon) &\leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \end{aligned}$$

PROOF. The proof utilizes a martingale sequence. Define

$$V_i = \mathbb{E}[f|X_1, \dots, X_i] - \mathbb{E}[f|X_1, \dots, X_{i-1}],$$

and note that $\mathbb{E}[V_i] = 0$ and

$$\sum_{i=1}^n V_i = \mathbb{E}[f|X_1, \dots, X_n] - \mathbb{E}[f] = f(X_1, \dots, X_n) - \mathbb{E}[f].$$

Define upper and lower bounds as

$$\begin{aligned} L_i &= \inf_x \mathbb{E}[f|X_1, \dots, X_{i-1}, x] - \mathbb{E}[f|X_1, \dots, X_{i-1}] \\ U_i &= \sup_x \mathbb{E}[f|X_1, \dots, X_{i-1}, x] - \mathbb{E}[f|X_1, \dots, X_{i-1}] \end{aligned}$$

and note that $L_i \leq V_i \leq U_i$. Furthermore, we have

$$U_i - L_i = \sup_{X_i} \sup_{X'_i} (\mathbb{E}[f|X_1, \dots, X_i] - \mathbb{E}[X_1, \dots, X'_i]) \leq c_i.$$

And we have $\mathbb{E}[V_i|X_1, \dots, X_{i-1}] = 0$. Similar to the proof of Hoeffding's inequality, we have

$$P(f - \mathbb{E}[f] \geq \epsilon) \leq \inf_{t>0} \exp(-t\epsilon) \mathbb{E} \left[\prod_{i=1}^n \exp(tV_i) \right].$$

And we have

$$\begin{aligned} \mathbb{E} \left[\prod_{i=1}^n \exp(tV_i) \right] &= \mathbb{E} \left[\mathbb{E} \left[\exp(tV_n) \prod_{i=1}^{n-1} \exp(tV_i) \mid X_1, \dots, X_{n-1} \right] \right] \\ &= \mathbb{E} \left[\prod_{i=1}^{n-1} \exp(tV_i) \mathbb{E} [\exp(tV_n) \mid X_1, \dots, X_{n-1}] \right] \\ &\leq \mathbb{E} \left[\prod_{i=1}^{n-1} \exp(tV_i) \right] \exp(t^2 c_n^2 / 8) \\ &\vdots \\ &\leq \exp \left(\frac{t^2 \sum_{i=1}^n c_i^2}{8} \right). \end{aligned}$$

Setting $t = \frac{4\epsilon}{\sum_{i=1}^n c_i^2}$ we obtain the claimed results.

□

Example. Consider

$$f(X_1, \dots, X_n) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(X)] - \frac{1}{n} \sum_{i=1}^n g(X_i) \right|.$$

If all $g : \mathcal{X} \mapsto [a, b]$ then we have $c_i = (b - a)/n$. Thus we have

$$P \left(\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(X)] - \frac{1}{n} \sum_{i=1}^n g(X_i) \right| - \mathbb{E} \left[\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(X)] - \frac{1}{n} \sum_{i=1}^n g(X_i) \right| \right] \geq \epsilon \right) \leq \exp \left(-\frac{2n\epsilon^2}{(b-a)^2} \right).$$

As a final note, the bounds we obtained are the worst case scenario since we did not utilize the variance information. We could obtain bounds if the variance were known, such as Bernstein's inequality.