Notes on PAC Bayes

Renjie Liao
Department of Computer Science
University of Toronto
rjliao@cs.toronto.edu
8th July, 2020

Abstract

In this note, we review the PAC-Bayesian approaches following [4, 3].

1 Prerequisites

We follow the literature to use KL^+ to denote the KL divergence between two Bernoulli distributions p and q as,

$$\mathrm{KL}^+(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

Recall the definition of strongly convex functions as follows.

Definition 1.1. A twice differentiable function f is called strongly convex with parameter $m > 0$ if for any x in the domain, we have

$$\nabla^2 f(x) \succeq mI.$$

Lemma 1.1. For any $p, q \in [0, 1]$, we have

$$\mathrm{KL}^+(p\|q) \geq 2(p-q)^2,$$

Proof. Let $f(p) = \mathrm{KL}^+(p\|q) - 2(p-q)^2$, we have

$$\begin{aligned} f'(p) &= \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{q}{1-q}\right) - 4(p-q) \\ f''(p) &= \frac{1}{p(1-p)} - 4 \end{aligned}$$

Since $p(1-p)$ achieves its maximum $1/4$ with $p = 1/2$, we have $f''(p) \geq 0, \forall p \in [0, 1]$. Note that $f'(q) = 0$. Hence, $f(p)$ decreases when $p \leq q$, increases when $p \geq q$, and achieves the minimum value $f(q) = 0$. □

Lemma 1.2. For any $p, q \in [0, 1]$ with $p \leq q$, we have

$$\mathrm{KL}^+(p\|q) \geq \frac{(p-q)^2}{2q},$$

Proof. Let $f(p) = \mathrm{KL}^+(p\|q) - \frac{(p-q)^2}{2q}$, we have

$$\begin{aligned} f'(p) &= \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{q}{1-q}\right) - \frac{p-q}{q} \\ f''(p) &= \frac{1}{p(1-p)} - \frac{1}{q} = \frac{q-p+p^2}{qp(1-p)} \end{aligned}$$

Preprint. Under review.

Since $p \leq q$, then we have $f''(p) \geq 0$. Note that $f'(q) = 0$. Hence, $f(p)$ decreases when $p \leq q$. Therefore $f(p) \geq f(q) = 0$ which proves the claim. $\square$

Lemma 1.3. $\mathrm{KL}^+(p\|q)$ is 4-strongly convex w.r.t. argument p and convex w.r.t. argument q.

Proof. Similar to the proof of Lemma 1.1, denoting $f(p) = \mathrm{KL}^+(p\|q)$, we have

$$\begin{aligned} f'(p) &= \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{q}{1-q}\right) \\ f''(p) &= \frac{1}{p(1-p)} \geq 4. \end{aligned}$$

Therefore, according to the definition 1.1, $f(p)$ is 4-strongly convex.

Denoting $g(q) = \mathrm{KL}^+(p\|q)$, we have

$$\begin{aligned} g'(q) &= -\frac{p}{q} + \frac{1-p}{1-q} \\ g''(q) &= \frac{p}{q^2} + \frac{1-p}{(1-q)^2} > 0. \end{aligned}$$

Therefore $g(q)$ is convex. $\square$

Lemma 1.4. (Hoeffding's Lemma [2]) For bounded random variables $X_1, \cdots, X_m$ where $X_i \in [0,1]$ and is i.i.d. with mean μ, let $\bar{X}_m = \frac{1}{m}\sum_{i=1}^m X_i$, then for any $\epsilon > 0$ with $\mu + \epsilon < 1$ and $\mu - \epsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(\bar{X}_m \geq \mu + \epsilon\right) &\leq e^{-m\,\mathrm{KL}^+(\mu+\epsilon\|\mu)} \leq e^{-2m\epsilon^2} \\ \mathbb{P}\left(\bar{X}_m \leq \mu - \epsilon\right) &\leq e^{-m\,\mathrm{KL}^+(\mu-\epsilon\|\mu)} \leq e^{-2m\epsilon^2} \end{aligned}$$

Proof. Let $t > 0$, $S_m = \sum_{i=1}^m X_i$, $p = \mu + \epsilon$, $\bar{p} = 1 - p$, and $\bar{\mu} = 1 - \mu$, we have

$$\begin{aligned} \mathbb{P}\left(\bar{X}_m - \mu \geq \epsilon\right) &= \mathbb{P}\left(S_m - m\mu \geq m\epsilon\right) \\ &= \mathbb{P}\left(e^{tS_m} \geq e^{tm(\mu+\epsilon)}\right) \\ &\leq \frac{\mathbb{E}\left[e^{tS_m}\right]}{e^{tm(\mu+\epsilon)}} \qquad \text{(Markov Inequality)} \\ &= \frac{\mathbb{E}\left[e^{tX_1}\right]^m}{e^{tm(\mu+\epsilon)}} \\ &\leq \frac{\mathbb{E}\left[X_1 e^t + (1-X_1)e^0\right]^m}{e^{tm(\mu+\epsilon)}} \qquad \text{(Convexity of } e^{tX}\text{)} \\ &= \frac{(\mu e^t + 1 - \mu)^m}{e^{tm(\mu+\epsilon)}} \\ &= \left(\frac{\mu e^t + 1 - \mu}{e^{t(\mu+\epsilon)}}\right)^m \\ &= \left(\mu e^{t\bar{p}} + \bar{\mu} e^{-tp}\right)^m \end{aligned} \tag{1}$$

Let $f(t) = m \ln g(t) = m \ln\left(\mu e^{t\bar{p}} + \bar{\mu} e^{-tp}\right)$, we have

$$\begin{aligned} f'(t) &= m\frac{g'(t)}{g(t)} \\ f''(t) &= m\frac{m\left(g(t)g''(t) - g'(t)^2\right)}{g(t)^2} \\ g'(t) &= \mu\bar{p}e^{t\bar{p}} - \bar{\mu}pe^{-tp} \\ g''(t) &= \mu\bar{p}^2e^{t\bar{p}} + \bar{\mu}p^2e^{-tp} \\ g(t)g''(t) - g'(t)^2 &= \mu\bar{\mu}e^{t(\bar{p}-p)}(\bar{p}-p)^2. \end{aligned}$$

Since $f''(t) \geq 0$, f achieves its minimum while $e^t = \frac{\bar{\mu}p}{\mu\bar{p}}$,

$$\begin{aligned}
\min_t \ f(t) &= \min_t \ m \ln\left(\mu e^{t(1-p)} + \bar{\mu}e^{-tp}\right) \\
&= \min_t \ m \ln\left(e^{-tp}\left(\mu e^t + \bar{\mu}\right)\right) \\
&= m \ln\left(\left(\frac{\mu\bar{p}}{\bar{\mu}p}\right)^p \left(\frac{\bar{\mu}p}{\bar{p}} + \bar{\mu}\right)\right) \\
&= m \ln\left(\frac{\mu^p \bar{p}^{(p-1)}}{\bar{\mu}^{(p-1)}p^p}\right) \\
&= m \ln\left(\frac{\mu^p \bar{\mu}^{\bar{p}}}{\bar{p}^{\bar{p}}p^p}\right) \qquad (2)
\end{aligned}$$

Combine Eq. (1) and Eq. (2), we have

$$\begin{aligned}
\mathbb{P}\left(\bar{X}_m - \mu \geq \epsilon\right) &\leq \min_t \ e^{f(t)} \\
&= \left(\frac{\mu^p \bar{\mu}^{\bar{p}}}{\bar{p}^{\bar{p}}p^p}\right)^m \\
&= e^{-mp\ln\left(\frac{p}{\mu}\right) - m\bar{p}\ln\left(\frac{\bar{p}}{\bar{\mu}}\right)} \\
&= e^{-m\,\mathrm{KL}^+(\mu+\epsilon\|\mu)} \\
&\leq e^{-2m\epsilon^2} \qquad \text{(Lemma 1.1)}
\end{aligned}$$

Similarly, we can prove the case for the other side. □

Remark. This proof is based on the one from [8].

Lemma 1.5. For non-negative continuous random variables X, we have

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq \nu)\mathrm{d}\nu.$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty X\mathbb{P}(X)\mathrm{d}X \\
&= \int_0^\infty \int_0^X \mathbf{1}\mathrm{d}\nu\mathbb{P}(X)\mathrm{d}X \\
&= \int_0^\infty \int_0^X \mathbb{P}(X)\mathrm{d}\nu\mathrm{d}X \\
&= \int_0^\infty \int_\nu^\infty \mathbb{P}(X)\mathrm{d}X\mathrm{d}\nu \qquad \text{(region of the integral is the same)} \\
&= \int_0^\infty \mathbb{P}(X \geq \nu)\mathrm{d}\nu
\end{aligned}$$

□

Remark. This lemma can also be obtained via **Integration by Parts**. Specifically, let $f(\nu) = \int_\nu^\infty \mathbb{P}(X)\mathrm{d}X = \mathbb{P}(X \geq \nu)$ and $g(\nu) = \int_0^\nu \mathbf{1}\mathrm{d}X = \nu$, we have

$$\begin{aligned}
\int_0^\infty \mathbb{P}(X \geq \nu)\mathrm{d}\nu = \int_0^\infty f(\nu)g'(\nu)\mathrm{d}\nu &= f(\nu)g(\nu)|_0^\infty - \int_0^\infty f'(\nu)g(\nu)\mathrm{d}\nu \\
&= -\int_0^\infty -\mathbb{P}(\nu)\nu\mathrm{d}\nu \\
&= \mathbb{E}[\nu] = \mathbb{E}[X]
\end{aligned}$$

Similar result holds for discrete nonnegative random variables.

Lemma 1.6. [2-side] Let X be a random variable satisfying $\mathbb{P}(X \geq \epsilon) \leq e^{-2m\epsilon^2}$ and $\mathbb{P}(X \leq -\epsilon) \leq e^{-2m\epsilon^2}$ where $m \geq 1$ and $\epsilon > 0$, we have

$$\mathbb{E}[e^{2(m-1)X^2}] \leq 2m.$$

Proof.

$$\begin{aligned}
\mathbb{E}[e^{2(m-1)X^2}] &= \int_0^\infty \mathbb{P}\left(e^{2(m-1)X^2} \geq \nu\right) \mathrm{d}\nu \qquad \text{(Lemma 1.5)} \\
&= \int_0^\infty \mathbb{P}\left(X^2 \geq \frac{\ln \nu}{2(m-1)}\right) \mathrm{d}\nu \\
&= \int_0^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu + \int_0^\infty \mathbb{P}\left(X \leq -\sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu \qquad (3)
\end{aligned}$$

$$\begin{aligned}
\int_0^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu &= \int_0^1 \mathbb{P}\left(X \geq \sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu + \int_1^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu \\
&\leq 1 + \int_1^\infty \mathbb{P}\left(X \geq \sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu \\
&\leq 1 + \int_1^\infty e^{-2m\frac{\ln \nu}{2(m-1)}} \mathrm{d}\nu \\
&= 1 + \left(-(m-1)\nu^{-\frac{1}{m-1}}\Big|_1^\infty\right) \\
&= m \qquad (4)
\end{aligned}$$

Similarly, we can show that

$$\int_0^\infty \mathbb{P}\left(X \leq -\sqrt{\frac{\ln \nu}{2(m-1)}}\right) \mathrm{d}\nu \leq m \qquad (5)$$

Combining Eq. (3) and Eq. (4), we finish the proof. □

Remark. One can also obtain the 1-side version $\mathbb{E}[e^{2(m-1)X^2}] \leq m$ by, *e.g.*, adding the assumption that $X \geq 0$ and following the same proof.

Lemma 1.7. [2-side] Let X be a random variable defined on $(0, 1)$ with mean μ satisfying $\mathbb{P}(X \geq \mu + \epsilon) \leq e^{-m\,\mathrm{KL}^+(\mu+\epsilon\|\mu)}$ and $\mathbb{P}(X \leq \mu - \epsilon) \leq e^{-m\,\mathrm{KL}^+(\mu-\epsilon\|\mu)}$ where $m \geq 1$, $\epsilon > 0$, $\mu + \epsilon < 1$ and $\mu - \epsilon > 0$, we have

$$\mathbb{E}[e^{(m-1)\,\mathrm{KL}^+(X\|\mu)}] \leq 2m.$$

Proof. Denoting $f(p) = \mathrm{KL}^+(p\|\mu)$, we have

$$\begin{aligned}
f'(p) &= \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{\mu}{1-\mu}\right) \\
f''(p) &= \frac{1}{p(1-p)}.
\end{aligned}$$

Since $f''(p) \geq 0$ for all $p \in (0, 1)$, $f(p)$ decreases when $0 < p \leq \mu$, increases when $\mu \leq p < 1$ and attains its minimum 0 at $p = \mu$. Therefore, based on the inverse function theorem, there exists the inverse function ϕ of $\mathrm{KL}^+(p\|\mu)$ when $p \in (0, \mu)$, *i.e.*, $p = \phi(\mathrm{KL}^+(p\|\mu))$, $\forall p \in (0, \mu)$. Similarly, there exists another inverse function ψ of $\mathrm{KL}^+(p\|\mu)$ when $p \in (\mu, 1)$, *i.e.*, $p = \psi(\mathrm{KL}^+(p\|\mu))$, $\forall p \in (\mu, 1)$. Note that the functional forms of ϕ and ψ may or may not be the same (*i.e.*, whether

$\mathrm{KL}^+(p\|\mu)$ is symmetric w.r.t. $p=\mu$) depending on the value of μ. As shown below, the exact functional form of the inverse function does not matter as long as it exists.

$$\begin{aligned}
\mathbb{E}[e^{(m-1)\,\mathrm{KL}^+(X\|q)}] &= \int_0^\infty \mathbb{P}\left(e^{(m-1)\,\mathrm{KL}^+(X\|q)} \geq \nu\right) \mathrm{d}\nu \qquad \text{(Lemma 1.5)} \\
&= \int_0^\infty \mathbb{P}\left(\mathrm{KL}^+(X\|q) \geq \frac{\ln\nu}{m-1}\right) \mathrm{d}\nu \\
&= \int_0^\infty \mathbb{P}\left(X \geq \phi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu + \int_0^\infty \mathbb{P}\left(X \leq \psi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu
\end{aligned} \tag{6}$$

We use the same trick again as in Lemma 1.6,

$$\begin{aligned}
\int_0^\infty \mathbb{P}\left(X \geq \phi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu &= \int_0^1 \mathbb{P}\left(X \geq \phi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu + \int_1^\infty \mathbb{P}\left(X \geq \phi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu \\
&\leq 1 + \int_1^\infty \mathbb{P}\left(X \geq \phi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu \\
&= 1 + \int_1^\infty e^{-\frac{m\ln\nu}{m-1}} \mathrm{d}\nu \\
&= 1 + \left(-(m-1)\nu^{-\frac{1}{m-1}}\Big|_1^\infty\right) \\
&= m
\end{aligned} \tag{7}$$

Similarly, we have $\int_0^\infty \mathbb{P}\left(X \leq \psi\left(\frac{\ln\nu}{m-1}\right)\right) \mathrm{d}\nu \leq m$. Therefore, combining it with Eq. (6) and Eq. (7), we prove the claim. $\square$

Remark. Similarly, one can obtain the 1-side version $\mathbb{E}[e^{(m-1)\,\mathrm{KL}^+(X\|\mu)}] \leq m$ by, *e.g.*, adding the assumption that $X \geq \mu$ and following the same proof.

2 Main Result

We only consider the binary classification problem. Let us first introduce some basic setup.

- Data z, $z=(x,y)$ input $x \in \mathbb{R}^d$ and the output $y \in \{0,1\}$
- Data Space $\mathcal{Z}$, $z \in \mathcal{Z}$
- Data Distribution D, $z \overset{iid}{\sim} D$
- Hypothesis h, model
- Hypothesis Class $\mathcal{H}$, $h \in \mathcal{H}$
- Training Set S with size m, $S=\{z_1,\ldots,z_m\}$
- Loss ℓ, $\ell : \mathcal{H} \times \mathcal{Z} \to \{0,1\}$

As usual, we care about the *generalization error*, *i.e.*, *error rate* or *misclassification rate* in this case,

$$L_D(h) = \mathbb{P}_{z\sim D}(h(x) \neq y), \tag{8}$$

where we use the subscript to emphasize the dependency on the data distribution. However, since we can not observe it directly, we approximate it using the empirical distribution, a.k.a., *empirical error*,

$$L_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h,z_i) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left[h(x_i) \neq y_i\right]. \tag{9}$$

where ℓ is the 0-1 loss.

Bayesian View We are interested in bounding the generalization error using the empirical error. PAC-Bayes [4, 3] takes a Bayesian view of PAC theory [7]. In particular, it assumes that we have a prior distribution P over the hypothesis class $\mathcal{H}$ and we gonna obtain the posterior Q after the learning process on the training set. We can define the generalization error and empirical error in terms of this Bayesian view as,

$$L_S(Q) = \mathbb{E}_{h\sim Q}[L_S(h)]$$
$$L_D(Q) = \mathbb{E}_{h\sim Q}[L_D(h)].$$

Before introducing the main results, let us prove some useful lemmas. First, note that $L_S(h)$ and $L_D(h)$ are distributions over Bernoulli random variables themselves.

Lemma 2.1. For any h and any $\epsilon > 0$ with $L_D(h) + \epsilon < 1$ and $L_D(h) - \epsilon > 0$, we have

$$\mathbb{P}(L_S(h) \geq L_D(h) + \epsilon) \leq e^{-m\,\mathrm{KL}^+(L_D(h)+\epsilon\|L_D(h))}$$
$$\mathbb{P}(L_S(h) \leq L_D(h) - \epsilon) \leq e^{-m\,\mathrm{KL}^+(L_D(h)-\epsilon\|L_D(h))}$$

Proof. First, recall the definition of $L_S(h)$ in Eq. (9) and observe that for any h we have $\mathbb{E}_S[L_S(h)] = L_D(h)$. Then we replace μ and $\bar{X}_m$in Lemma 1.4 with $L_D(h)$ and $L_S(h)$, we finish the proof. □

Lemma 2.2. For any distribution Q over $\mathcal{H}$, $\forall h \in \mathcal{H}$, $p(h) \in (0,1)$ and $q(h) \in (0,1)$, we have

$$\mathrm{KL}^+\left(\mathbb{E}_{h\sim Q}[p(h)]\|\mathbb{E}_{h\sim Q}[q(h)]\right) \leq \mathbb{E}_{h\sim Q}\left[\mathrm{KL}^+(p(h)\|q(h))\right]$$

Proof. Denoting $f(p) = \mathrm{KL}^+(p\|q)$, from Lemma 1.3, we know $f(p)$ is strongly convex for any q. Therefore, based on the Jensen's inequality, we have

$$\mathrm{KL}^+\left(\mathbb{E}_{h\sim Q}[p(h)]\|\mathbb{E}_{h\sim Q}[q(h)]\right) \leq \mathbb{E}_{h\sim Q}\left[\mathrm{KL}^+\left(p(h)\|\mathbb{E}_{h\sim Q}[q(h)]\right)\right]. \tag{10}$$

Denoting $g(q) = \mathrm{KL}^+(p\|q)$, from Lemma 1.3, we know $g(q)$ is convex for any p,

$$\mathrm{KL}^+\left(p(h)\|\mathbb{E}_{h\sim Q}[q(h)]\right) \leq \mathbb{E}_{h\sim Q}\left[\mathrm{KL}^+\left(p(h)\|q(h)\right)\right]. \tag{11}$$

Combining Eq. (10) and Eq. (11), we have

$$\begin{aligned}\mathrm{KL}^+\left(\mathbb{E}_{h\sim Q}[p(h)]\|\mathbb{E}_{h\sim Q}[q(h)]\right) &\leq \mathbb{E}_{h\sim Q}\left[\mathbb{E}_{h\sim Q}\left[\mathrm{KL}^+\left(p(h)\|q(h)\right)\right]\right] \\ &= \mathbb{E}_{h\sim Q}\left[\mathrm{KL}^+\left(p(h)\|q(h)\right)\right],\end{aligned}$$

which proves the claim. □

2.1 Generalization Bound with KL divergence

Now we introduce the generalization bound which uses KL^+ to measure the distance between prior P and posterior Q over models.

Theorem 2.3. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Then, with probability $1-\delta$ over the choice of an i.i.d. training set S according to D, for all distributions Q over $\mathcal{H}$, we have

$$\mathrm{KL}^+(L_S(Q)\|L_D(Q)) \leq \frac{\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}}{m-1}$$

Proof. Fix a hypothesis h, based on Lemma 2.1, we have

$$\mathbb{P}\left(L_S(h) \geq L_D(h) + \epsilon\right) \leq e^{-m\,\mathrm{KL}^+(L_D(h)+\epsilon\|L_D(h))}$$
$$\mathbb{P}\left(L_S(h) \geq L_D(h) - \epsilon\right) \leq e^{-m\,\mathrm{KL}^+(L_D(h)-\epsilon\|L_D(h))}.$$

Then, based on Lemma 1.7, we have

$$\mathbb{E}_S[e^{(m-1)\,\mathrm{KL}^+(L_S(h)\|L_D(h))}] \leq 2m. \tag{12}$$

Therefore, for all S and any $\delta > 0$, we have,

$$\mathbb{P}\left(e^{(m-1)\,\mathrm{KL}^+(L_S(h)\|L_D(h))} \geq \frac{2m}{\delta}\right) \leq \frac{\mathbb{E}_S[e^{(m-1)\,\mathrm{KL}^+(L_S(h)\|L_D(h))}]\delta}{2m} \leq \delta \tag{13}$$

For any function $f(h)$, we have

$$\begin{aligned}
\mathbb{E}_{h\sim Q}[f(h)] &= \mathbb{E}_{h\sim Q}[\ln e^{f(h)}] \\
&= \mathbb{E}_{h\sim Q}[\ln e^{f(h)} + \ln \frac{Q}{P} + \ln \frac{P}{Q}] \\
&= \mathrm{KL}(Q\|P) + \mathbb{E}_{h\sim Q}\left[\ln\left(\frac{P}{Q}e^{f(h)}\right)\right] \\
&\leq \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim Q}\left[\frac{P}{Q}e^{f(h)}\right] \\
&= \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim P}\left[e^{f(h)}\right].
\end{aligned} \tag{14}$$

Let $f(h) = (m-1)\,\mathrm{KL}^+(L_S(h)\|L_D(h))$. From Eq. (13) and Eq. (14), we have, with at least probability $1-\delta$,

$$\begin{aligned}
(m-1)\mathbb{E}_{h\sim Q}[\mathrm{KL}^+(L_S(h)\|L_D(h))] &\leq \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim P}\left[e^{(m-1)\,\mathrm{KL}^+(L_S(h)\|L_D(h))}\right] \\
&\leq \mathrm{KL}(Q\|P) + \ln\left(\frac{2m}{\delta}\right).
\end{aligned} \tag{15}$$

From Lemma 2.2, we have

$$\begin{aligned}
(m-1)\,\mathrm{KL}^+(\mathbb{E}_{h\sim Q}[L_S(h)]\|\mathbb{E}_{h\sim Q}[L_D(h)]) &\leq (m-1)\mathbb{E}_{h\sim Q}[\mathrm{KL}^+(L_S(h)\|L_D(h))] \\
&\leq \mathrm{KL}(Q\|P) + \ln\left(\frac{2m}{\delta}\right),
\end{aligned} \tag{16}$$

which proves the theorem. □

Remark. The proof follows the original proof of [3]. Again, we can have a one-side version,

$$\mathrm{KL}^+(L_S(Q)\|L_D(Q)) \leq \frac{\mathrm{KL}(Q\|P) + \ln\frac{m}{\delta}}{m-1},$$

by adding the assumption $L_S(Q) > L_D(Q)$ which is reasonable in practice.

One can also prove a slight different generalization bound by using a different technique as below.

Theorem 2.4. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Then, with probability $1-\delta$ over the choice of an i.i.d. training set S according to D, for all distributions Q over $\mathcal{H}$, we have

$$\mathrm{KL}^+(L_S(Q)\|L_D(Q)) \leq \frac{\mathrm{KL}(Q\|P) + \ln\frac{m+1}{\delta}}{m}$$

Proof. From Eq. (14) in the proof of Theorem 2.3, for any function $f(h)$, we have

$$\mathbb{E}_{h\sim Q}[f(h)] \leq \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim P}\left[e^{f(h)}\right]. \tag{17}$$

Based on Lemma 2.2 and let $f(h) = m\,\mathrm{KL}^+(L_S(h)\|L_D(h))$, we have

$$\begin{aligned}
\mathrm{KL}^+(L_S(Q)\|L_D(Q)) &\leq \mathbb{E}_{h\sim Q}[\mathrm{KL}^+(L_S(h)\|L_D(h))] \\
&= \mathbb{E}_{h\sim Q}[\frac{1}{m}f(h)] \\
&\leq \frac{\mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim P}\left[e^{f(h)}\right]}{m}.
\end{aligned} \tag{18}$$

Note that since we consider 0-1 loss and samples are i.i.d., $mL_S(h)$ can only take values from $\{0, 1, 2, \dots, m\}$ and follows the binomial distribution $B(m, L_D(h))$. Hence, we have,

$$\begin{aligned}
\mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{f(h)}\right]\right] &= \mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{m\,\mathrm{KL}^+(L_S(h)\|L_D(h))}\right]\right] \\
&= \mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{mL_S(h)\ln\frac{L_S(h)}{L_D(h)}+m(1-L_S(h))\ln\frac{1-L_S(h)}{1-L_D(h)}}\right]\right] \\
&= \mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[\left(\frac{L_S(h)}{L_D(h)}\right)^{mL_S(h)}\left(\frac{1-L_S(h)}{1-L_D(h)}\right)^{m(1-L_S(h))}\right]\right] \\
&= \mathbb{E}_{h\sim P}\left[\mathbb{E}_S\left[\left(\frac{L_S(h)}{L_D(h)}\right)^{mL_S(h)}\left(\frac{1-L_S(h)}{1-L_D(h)}\right)^{m(1-L_S(h))}\right]\right] \\
&= \mathbb{E}_{h\sim P}\left[\sum_{k=0}^{m}\binom{m}{k}L_D(h)^k(1-L_D(h))^{m-k}\left(\frac{k/m}{L_D(h)}\right)^k\left(\frac{1-k/m}{1-L_D(h)}\right)^{m-k}\right] \\
&= \mathbb{E}_{h\sim P}\left[\sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(\frac{m-k}{m}\right)^{m-k}\right] \\
&\le m+1, \qquad (19)
\end{aligned}$$

where the last inequality uses the fact that $\binom{m}{k}\left(\frac{k}{m}\right)^k\left(\frac{m-k}{m}\right)^{m-k}$ is the probability of a binomial random variable (following $B(m, \frac{k}{m})$) taking the value as k, thus being no larger than 1. The second to the last line makes use of the law of the unconscious statistician (LOTUS).

Based on Markov's inequality, for any $\delta > 0$, we have

$$\mathbb{P}\left(\mathbb{E}_{h\sim P}\left[e^{f(h)}\right] \ge \frac{m+1}{\delta}\right) \le \frac{\delta\mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{f(h)}\right]\right]}{m+1} \le \delta, \qquad (20)$$

which along with Eq. (18) proves the theorem. ☐

Remark. This proof largely follows the one in [1]. The observation that, for any h, $mL_S(h)$ is distributed according to the binomial distribution $B(m, L_D(h))$ is really insightful. One can further improves Eq. (19) to show $\sqrt{m} \le \mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{f(h)}\right]\right] \le \sqrt{2m}$. How?

2.2 Canonical Generalization Bound

Theorem 2.5. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0, 1)$. Then, with probability $1-\delta$ over the choice of an i.i.d. training set S according to D, for all distributions Q over $\mathcal{H}$, we have

$$L_D(Q) \le L_S(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}}{2(m-1)}}$$

Proof. (Proof-I) From Lemma 1.1 and Theorem 2.3, we have

$$2(L_S(Q) - L_D(Q))^2 \le \mathrm{KL}^+(L_S(Q)\|L_D(Q)) \le \frac{\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}}{m-1}.$$

Therefore, we have,

$$L_D(Q)) \le L_S(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}}{2(m-1)}},$$

which proves the theorem. ☐

Proof. (Proof-II) Let $\Delta(h) = L_D(h) - L_S(h)$. From Eq. (14), for any function $f(h)$, we have

$$\mathbb{E}_{h\sim Q}[f(h)] \le \mathrm{KL}(Q\|P) + \ln\mathbb{E}_{h\sim P}\left[e^{f(h)}\right]. \qquad (21)$$

Let $f(h) = 2(m-1)\Delta(h)^2$. We have

$$2(m-1)\mathbb{E}_{h\sim Q}[\Delta(h)]^2 \le 2(m-1)\mathbb{E}_{h\sim Q}[\Delta(h)^2] \qquad \text{(Jensen's inequality)}$$
$$\le \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]. \tag{22}$$

Since $L_D(h) \in [0,1]$, based on Hoeffding's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}(\Delta(h) \ge \epsilon) \le e^{-2m\epsilon^2}$$
$$\mathbb{P}(\Delta(h) \le -\epsilon) \le e^{-2m\epsilon^2}$$

Hence, based on Lemma 1.6, we have

$$\mathbb{E}_S\left[e^{2(m-1)\Delta(h)^2}\right] \le 2m \;\Rightarrow\; \mathbb{E}_{h\sim P}\left[\mathbb{E}_S\left[e^{2(m-1)\Delta(h)^2}\right]\right] \le 2m$$
$$\Leftrightarrow\; \mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]\right] \le 2m$$

Based on Markov's inequality, we have

$$\mathbb{P}\left(\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right] \ge \frac{2m}{\delta}\right) \le \frac{\delta\mathbb{E}_S\left[\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]\right]}{2m} \le \delta. \tag{23}$$

Combining Eq. (22) and Eq. (23), with probability $1-\delta$, we have

$$\mathbb{E}_{h\sim Q}[\Delta(h)]^2 \le \frac{\mathrm{KL}(Q\|P) + \ln\left(\frac{2m}{\delta}\right)}{2(m-1)} \tag{24}$$

which proves the theorem. □

Remark. This theorem has a simple form and is more similar to the majority of generalization bounds. Therefore, it is frequently used in the literature. Proof-I is based on the one in [3, 1]. Proof-II is based on the one in the chapter 31 of [6]. The proof technique in Proof-II is more general in a sense that one can generalize the loss beyond 0-1 loss under this framework and derive similar results.

Theorem 2.6. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Then, with probability $1-\delta$ over the choice of an i.i.d. training set S according to D, for all distributions Q over $\mathcal{H}$, we have

$$L_D(Q) \le L_S(Q) + \sqrt{\frac{2L_S(Q)\left(\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}\right)}{m-1}} + \frac{2\left(\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}\right)}{m-1}$$

Proof. From Lemma 1.2, we have for any $p, q \in [0,1]$ with $p \le q$,

$$\mathrm{KL}^+(p\|q) \ge \frac{(p-q)^2}{2q}.$$

If $\mathrm{KL}^+(p\|q) \le x$, then we have

$$x \ge \frac{(p-q)^2}{2q} \;\Leftrightarrow\; 2qx \ge (p-q)^2 \;\Leftrightarrow\; q \le p + \sqrt{2qx}. \tag{25}$$

Note that

$$q \le p + \sqrt{2qx} \;\Leftrightarrow\; \left(\sqrt{q} - \frac{\sqrt{2x}}{2}\right)^2 \le p + \frac{x}{2}$$
$$\Leftrightarrow\; \sqrt{2q} \le \sqrt{2p+x} + \sqrt{x} \tag{26}$$

Based on Eq. (25) and Eq. (26), we have

$$\begin{aligned} q &\le p + \sqrt{2qx} \\ &\le p + (\sqrt{2p+x} + \sqrt{x})\sqrt{x} \\ &= p + \sqrt{2px + x^2} + x \\ &\le p + \sqrt{2px} + 2x \qquad \text{(Subadditivity: } \sqrt{x+y} < \sqrt{x} + \sqrt{y}\text{)} \end{aligned} \tag{27}$$

Let $p = L_S(Q)$, $q = L_D(Q)$, and $x = \frac{\mathrm{KL}(Q\|P) + \ln\frac{2m}{\delta}}{m-1}$. Theorem 2.3 shows $\mathrm{KL}^+(p\|q) \le x$. Then Eq. (27) proves the theorem. □

Remark. Note that whether Theorem 2.5 or Theorem 2.6 provides a sharper bound depends on the actual value of $L_S(Q)$. But Theorem 2.5 has a simpler form.

2.3 Generalization Bound of Deterministic Models

Let us first review a result from [5] which generalizes the PAC-Bayes bound to a general class of deterministic models. We define the model to be $f_w \in \mathcal{H} : \mathcal{X} \to \mathbb{R}^k$ where w are the parameters of the model, $\mathcal{X}$ is the input space, and $k \geq 1$. We also define the γ-margin loss for the k-category classification as,

$$\tilde{L}_D(f_w, \gamma) = \mathbb{P}_{z\sim D}\left(f_w(x)[y] \leq \gamma + \max_{j\neq y} f_w(x)[j]\right),$$

where $\gamma > 0$ and $f_w(x)[j]$ means the j-th output of the model. Accordingly, we can define the empirical version,

$$\tilde{L}_S(f_w, \gamma) = \frac{1}{m}\sum_{z_i\in S} \mathbf{1}\left(f_w(x)[y] \leq \gamma + \max_{j\neq y} f_w(x)[j]\right),$$

We use $\mathbb{N}_m^+$ to denote the first m positive integers, *i.e.*, $\mathbb{N}_m^+ = \{1, 2, \ldots, m\}$.

Theorem 2.7. Let $f_w(x) : \mathcal{X} \to \mathbb{R}^k$ be any model with parameters w, and P be any distribution on the parameters that is independent of the training data. For any w, we construct a posterior $Q(w + u)$ by adding any random perturbation u to w, *s.t.*, $\mathbb{P}(\max_{x\in\mathcal{X}} \|f_{w+u}(x) - f_w(x)\|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$. Then, for any $\gamma, \delta > 0$, with probability at least $1 - \delta$ over the size-m training set S, for any w, we have:

$$\tilde{L}_D(f_w, 0) \leq \tilde{L}_S(f_w, \gamma) + \sqrt{\frac{2\,\mathrm{KL}(Q\|P) + \ln\frac{8m}{\delta}}{2(m-1)}} \tag{28}$$

Proof. Let $\tilde{w} = w + u$. Let $\mathcal{C}$ be the set of perturbation with the following property,

$$\mathcal{C} = \left\{w' \middle| \max_{x\in\mathcal{X}} \|f_{w'}(x) - f_w(x)\|_\infty < \frac{\gamma}{4}\right\}. \tag{29}$$

$\tilde{w} = w + u$ (w is deterministic and u is stochastic) is distributed according to $Q(\tilde{w})$. We now construct a new posterior $\tilde{Q}$ as follows,

$$\tilde{Q}(\tilde{w}) = \begin{cases} \frac{1}{Z}Q(\tilde{w}) & \tilde{w} \in \mathcal{C} \\ 0 & \tilde{w} \in \bar{\mathcal{C}}. \end{cases} \tag{30}$$

Here $Z = \int_{\tilde{w}\in\mathcal{C}} \mathrm{d}\tilde{Q}(\tilde{w}) = \mathbb{P}(\tilde{w} \in \mathcal{C})$. We know from the assumption that $Z > \frac{1}{2}$. $\bar{\mathcal{C}}$ is the complement set of $\mathcal{C}$. Therefore, for any $\tilde{w} \sim \tilde{Q}$, we have

$$\begin{aligned}
&\max_{i\in\mathbb{N}_k^+, j\in\mathbb{N}_k^+, x\in\mathcal{X}} \Big| |f_{\tilde{w}}(x)[i] - f_{\tilde{w}}(x)[j]| - |f_w(x)[i] - f_w(x)[j]| \Big| \\
\leq &\max_{i\in\mathbb{N}_k^+, j\in\mathbb{N}_k^+, x\in\mathcal{X}} \Big| f_{\tilde{w}}(x)[i] - f_{\tilde{w}}(x)[j] - f_w(x)[i] + f_w(x)[j] \Big| \\
\leq &\max_{i\in\mathbb{N}_k^+, j\in\mathbb{N}_k^+, x\in\mathcal{X}} \Big| f_{\tilde{w}}(x)[i] - f_w(x)[i] \Big| + \Big| f_{\tilde{w}}(x)[j] - f_w(x)[j] \Big| \\
\leq &\max_{i\in\mathbb{N}_k^+, x\in\mathcal{X}} \Big| f_{\tilde{w}}(x)[i] - f_w(x)[i] \Big| + \max_{j\in\mathbb{N}_k^+, x\in\mathcal{X}} \Big| f_{\tilde{w}}(x)[j] - f_w(x)[j] \Big| \\
&< \frac{\gamma}{4} + \frac{\gamma}{4} = \frac{\gamma}{2}
\end{aligned} \tag{31}$$

Recall that

$$\begin{aligned}
\tilde{L}_D(f_w, 0) &= \mathbb{P}_{z\sim D}\left(f_w(x)[y] \leq \max_{j\neq y} f_w(x)[j]\right) \\
\tilde{L}_D(f_{\tilde{w}}, \frac{\gamma}{2}) &= \mathbb{P}_{z\sim D}\left(f_{\tilde{w}}(x)[y] \leq \frac{\gamma}{2} + \max_{j\neq y} f_{\tilde{w}}(x)[j]\right),
\end{aligned}$$

Denoting $j_1^* = \arg\max_{j \neq y} f_{\tilde{w}}(x)[j]$ and $j_2^* = \arg\max_{j \neq y} f_w(x)[j]$, from Eq. (31), we have

$$\begin{aligned} &\left| f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_2^*] - f_w(x)[y] + f_w(x)[j_2^*] \right| < \frac{\gamma}{2} \\ \Rightarrow\ & f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_2^*] < f_w(x)[y] - f_w(x)[j_2^*] + \frac{\gamma}{2} \end{aligned} \tag{32}$$

Note that since $f_{\tilde{w}}(x)[j_1^*] \geq f_{\tilde{w}}(x)[j_2^*]$, we have

$$\begin{aligned} f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] &\leq f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_2^*] \\ &\leq f_w(x)[y] - f_w(x)[j_2^*] + \frac{\gamma}{2} \qquad \text{(Eq. (32))} \end{aligned}$$

Therefore, we have

$$f_w(x)[y] - f_w(x)[j_2^*] \leq 0 \ \Rightarrow\ f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] \leq \frac{\gamma}{2},$$

which indicates $\underset{z\sim D}{\mathbb{P}}\left(f_w(x)[y] \leq f_w(x)[j_2^*]\right) \leq \underset{z\sim D}{\mathbb{P}}\left(f_{\tilde{w}}(x)[y] \leq f_{\tilde{w}}(x)[j_1^*] + \frac{\gamma}{2}\right)$, or equivalently

$$\tilde{L}_D(f_w, 0) \leq \tilde{L}_D(f_{\tilde{w}}, \frac{\gamma}{2}). \tag{33}$$

Note that this holds for any perturbation $\tilde{w} \sim \tilde{Q}$.

Again, recall that

$$\begin{aligned} \tilde{L}_D(f_{\tilde{w}}, \frac{\gamma}{2}) &= \underset{z\sim D}{\mathbb{P}}\left(f_{\tilde{w}}(x)[y] \leq \frac{\gamma}{2} + \max_{j \neq y} f_{\tilde{w}}(x)[j]\right) \\ \tilde{L}_D(f_w, \gamma) &= \underset{z\sim D}{\mathbb{P}}\left(f_w(x)[y] \leq \gamma + \max_{j \neq y} f_w(x)[j]\right) \end{aligned}$$

From Eq. (31), we have

$$\begin{aligned} &\left| f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] - f_w(x)[y] + f_w(x)[j_1^*] \right| < \frac{\gamma}{2} \\ \Rightarrow\ & f_w(x)[y] - f_w(x)[j_1^*] < f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] + \frac{\gamma}{2} \end{aligned} \tag{34}$$

Note that since $f_w(x)[j_2^*] \geq f_w(x)[j_1^*]$, we have

$$\begin{aligned} f_w(x)[y] - f_w(x)[j_2^*] &\leq f_w(x)[y] - f_w(x)[j_1^*] \\ &\leq f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] + \frac{\gamma}{2} \qquad \text{(Eq. (34))} \end{aligned}$$

Therefore, we have

$$f_{\tilde{w}}(x)[y] - f_{\tilde{w}}(x)[j_1^*] \leq \frac{\gamma}{2} \ \Rightarrow\ f_w(x)[y] - f_w(x)[j_2^*] \leq \gamma,$$

which indicates $\tilde{L}_D(f_{\tilde{w}}, \frac{\gamma}{2}) \leq \tilde{L}_D(f_w, \gamma)$. Therefore, from the perspective of the empirical estimation of the probability, for any $\tilde{w} \sim \tilde{Q}$, we almost surely have

$$\tilde{L}_S(f_{\tilde{w}}, \frac{\gamma}{2}) \leq \tilde{L}_S(f_w, \gamma). \tag{35}$$

Now with probability at least $1 - \delta$, we have

$$\begin{aligned} \tilde{L}_D(f_w, 0) &\leq \mathbb{E}_{\tilde{w}\sim\tilde{Q}}\left[\tilde{L}_D(f_{\tilde{w}}, \frac{\gamma}{2})\right] \qquad \text{(Eq. (33))} \\ &\leq \mathbb{E}_{\tilde{w}\sim\tilde{Q}}\left[\tilde{L}_S(f_{\tilde{w}}, \frac{\gamma}{2})\right] + \sqrt{\frac{\mathrm{KL}(\tilde{Q}\|P) + \ln\frac{2m}{\delta}}{2(m-1)}} \qquad \text{(Theorem 2.5)} \\ &\leq \tilde{L}_S(f_w, \gamma) + \sqrt{\frac{\mathrm{KL}(\tilde{Q}\|P) + \ln\frac{2m}{\delta}}{2(m-1)}} \qquad \text{(Eq. (35))} \end{aligned} \tag{36}$$

Note that

$$\begin{aligned}
\mathrm{KL}(Q\|P) &= \int\limits_{\tilde{w}\in\mathcal{C}} Q\ln\frac{Q}{P}\mathrm{d}\tilde{w} + \int\limits_{\tilde{w}\in\bar{\mathcal{C}}} Q\ln\frac{Q}{P}\mathrm{d}\tilde{w} \\
&= \int\limits_{\tilde{w}\in\mathcal{C}} \frac{QZ}{Z}\ln\frac{Q}{ZP}\mathrm{d}\tilde{w} + \int\limits_{\tilde{w}\in\mathcal{C}} Q\ln Z\mathrm{d}\tilde{w} + \int\limits_{\tilde{w}\in\bar{\mathcal{C}}} \frac{Q(1-Z)}{1-Z}\ln\frac{Q}{(1-Z)P}\mathrm{d}\tilde{w} + \int\limits_{\tilde{w}\in\bar{\mathcal{C}}} Q\ln(1-Z)\mathrm{d}\tilde{w} \\
&= Z\,\mathrm{KL}(\tilde{Q}\|P) + (1-Z)\,\mathrm{KL}(\bar{Q}\|P) - H(Z),
\end{aligned} \tag{37}$$

where $\bar{Q}$ denotes the normalized density of Q restricted to $\bar{\mathcal{C}}$. $H(Z)$ is the entropy of a Bernoulli random variable with parameter Z. Since we know $\frac{1}{2} \le Z \le 1$ from the beginning, $0 \le H(Z) \le \ln 2$, and KL is nonnegative, from Eq. (37), we have

$$\begin{aligned}
\mathrm{KL}(\tilde{Q}\|P) &= \frac{1}{Z}\left[\mathrm{KL}(Q\|P) + H(Z) - (1-Z)\,\mathrm{KL}(\bar{Q}\|P)\right] \\
&\le \frac{1}{Z}\left[\mathrm{KL}(Q\|P) + H(Z)\right] \\
&\le 2\,\mathrm{KL}(Q\|P) + 2\ln 2.
\end{aligned} \tag{38}$$

Combining Eq. (36) and Eq. (38), we have

$$\tilde{L}_D(f_w, 0) \le \tilde{L}_S(f_w, \gamma) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \frac{1}{2}\ln\frac{8m}{\delta}}{m-1}}, \tag{39}$$

which finishes the proof. □

Remark. Note that the constants are slightly different from the one in [5] due to the facts that we use two-side version of Theorem 2.5 and we use natural logarithm rather than the one with base 2.

References

[1] Ofer Dekel. Pac-bayes analysis. https://courses.cs.washington.edu/courses/cse522/11wi/scribes/lecture13.pdf, 2011.

[2] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[3] David McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003.

[4] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[5] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[6] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[7] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[8] Robert L Wolpert. Markov, chebychev and hoeffding inequalities. https://www2.stat.duke.edu/courses/Spring09/sta205/lec/hoef.pdf, 2009.