1 Review of PAC Bayes Theorem

**Theorem 1.** ∀ distributions \( \mathcal{D} \), ∀ hypothesis \( \mathcal{H} \), ∀ priors \( \mathcal{P} \) on \( \mathcal{H} \), ∀\( \delta > 0 \) w.p. \( \geq 1 - \delta \), it holds for all posteriors \( \mathcal{Q} \) on \( \mathcal{H} \) that

\[
KL(l(\mathcal{Q}; S) || l(\mathcal{Q}; \mathcal{D})) \leq KL(\mathcal{Q}||\mathcal{P}) + \log \frac{m+1}{\delta}
\]

**Lemma 2.** For any scalars, \( \alpha, \beta \) let it hold that \( KL(\alpha || \beta) \leq x \). Then, \( |\alpha - \beta| \leq \sqrt{\frac{x}{2}} \). Also if \( \beta > \alpha \), \( \beta - \alpha \leq \sqrt{2} x \alpha + 2 x \).

2 Derandomizing PAC Bayes bounds

**Notation**

\( X = [-1, 1]^n \subset \mathbb{R}^n = \{ w \in \mathbb{R}^n : \|w\|_\infty \leq 1 \} \). \( \mathcal{H} \) is a linear hypothesis class, so that any element, \( h_w \) applied to \( x \) has the form, \( h_w(x) = \langle w, x \rangle \) with \( w \in X \). For any feature vector \( x \), \( \text{sgn}(h_w(x)) \) is the binary prediction and \( |h_w(x)| \) is the confidence. Denote by \( l_\gamma \) the \( \gamma \)-margin 0-1 loss. That is, \( l_\gamma(h_w; (x, y)) = 1_{y h_w(x) \leq \gamma} \). Note that \( l_0 \) is the standard error-indicator loss. For a uniform distribution, \( \mathcal{P} \) let \( \text{vol}(\mathcal{P}) \) denote the volume of the sample space having non-zero probability mass.

**Theorem 3.** Let \( A \) be any algorithm that takes in a sample \( S \sim \mathcal{D}^m \) and outputs a hypothesis \( ˜w \) with \( ˜w \in [-\frac{\gamma}{2n}, \frac{\gamma}{2n}]^n \). Let \( \mathcal{P} \) be uniformly distributed on \( [-\frac{\gamma}{2n}, \frac{\gamma}{2n}]^n \) \( \cap \mathcal{P} \). Then, \( l_0(˜w; \mathcal{D}) \leq l_\gamma(˜w; S) + \sqrt{\frac{n \log(\frac{2n}{\delta}) + \log(\frac{m+1}{\delta})}{2m}} \).

Note that the algorithm \( A \) needn’t know anything about the prior \( \mathcal{P} \) or posterior, \( \mathcal{Q} \). These two quantities are chosen in the theorem to give good de-randomized PAC-Bayes bounds. The proof of the theorem follows from two lemmas given below.

**Lemma 4.** The following inequalities hold true:

\[
\begin{align*}
    l_0(˜w; \mathcal{D}) &\leq l_\gamma(\mathcal{Q}; \mathcal{D}) \\
    l_\gamma(\mathcal{Q}; S) &\leq l_\gamma(˜w; S)
\end{align*}
\]

**Proof.** \( \forall ˜w \in \mathcal{Q}, \forall x \in X \) we have,

\[
\begin{align*}
    |⟨ ˜w, x⟩ - ⟨ ˜w, x⟩| &= \left| \sum_{j=1}^{n} x_j (˜w_j - ˜w_j) \right| \\
    &\leq \sum_{j=1}^{n} |x_j (˜w_j - ˜w_j)| \\
    &\leq \sum_{j=1}^{n} | ˜w_j - ˜w_j| \\
    &\leq \sum_{j=1}^{n} \frac{\gamma}{2n} \\
    &= \frac{\gamma}{2}
\end{align*}
\]
Note that $l_1(\hat{w}; (x, y)) = 0 \Rightarrow l_2(\hat{w}; (x, y)) = 0$. Indeed, let $y = 1$ then $\langle \hat{w}, x \rangle \geq \gamma$. Hence from (2), $\langle \hat{w}, x \rangle \geq \langle \hat{w}, x \rangle - \hat{\gamma} \geq \frac{\gamma}{2}$. Similarly, $l_2(\hat{w}; (x, y)) = 0 \Rightarrow l_0(\hat{w}; (x, y)) = 0$. The previous two implications immediately imply that,

$$
\begin{align*}
    l_0(\hat{w}; D) & \leq l_2(\hat{w}; D) \\
    l_0(\hat{w}; S) & \leq l_2(\hat{w}; S)
\end{align*}
$$

Taking expectation of above inequalities over $\hat{w} \sim Q$, the lemma follows. \hfill \Box

**Lemma 5.** $KL(Q||P) \leq n \log \left( \frac{4n}{\gamma} \right)$.

**Proof.** Note from definition that $\text{vol}(P) = 2^n$, $\text{vol}(Q) \geq \left( \frac{\gamma}{2n} \right)^n$. Let $q(h), p(h)$ be the p.d.f of $Q, P$ respectively.

$$
KL(Q||P) = \int_{h \in \mathcal{X}} q(h) \log \frac{q(h)}{p(h)} = \log \frac{\text{vol}(P)}{\text{vol}(Q)} \leq n \log \frac{4n}{\gamma}.
$$

\hfill \Box

**Proof of Theorem 3.** Note that (1) holds for $l = l_2$. Along with Lemma 2, this implies that

$$
l_2(Q; D) \leq l_2(Q; S) + \sqrt{\frac{KL(Q||P) + \log \frac{n+1}{2m}}{2m}}
$$

Using (4) along with Lemma 4 and Lemma 5 we have that,

$$
\begin{align*}
l_0(\hat{w}; D) & \leq l_2(Q; D) \\
& \leq l_2(Q; S) + \sqrt{\frac{KL(Q||P) + \log \frac{n+1}{2m}}{2m}} \\
& \leq l_2(\hat{w}; S) + \sqrt{\frac{n \log \frac{4n}{\gamma} + \log \frac{n+1}{2m}}{2m}}
\end{align*}
$$

\hfill \Box

### 3 Distribution dependent priors

In this section, we give two examples of distribution dependent priors on the hypothesis space that give good PAC-Bayes bounds.

#### 3.1 Generically prior

Given a sample $S \sim D^n$ and an algorithm $A(S)$, the posterior $Q$ is a function of $A(S)$. The bound on the right hand side of (1) can be minimized by choosing $P$ appropriately. Set,

$$
P^* = \arg\min_{P \in \mathcal{P}} \mathbb{E}_{S \sim D^n} [KL(Q||P)]
$$

The following lemma shows that $P^*$ would be dependent on the distribution $D$ but not on the sample $S$.

**Lemma 6.** $P^* = \mathbb{E}_{S \sim D^n} [Q]$.

**Proof.** Let $q(h)$ and $p(h)$ be the p.d.f of $Q$ and $P$ respectively. Note that minimizing $\mathbb{E}_{S \sim D^n} [KL(Q||P)] = \int_{S \sim D^n} \int_{A} q(h) \log \frac{q(h)}{p(h)} dh dS$ with respect to $P$ is equivalent to minimizing $\int_{S \sim D^n} \int_{A} q(h) \log \frac{1}{p(h)} dh dS$ with respect to $P$. Note that $\mathbb{E}_{S}[q(h)] = \bar{q}(h) = \int_{S \sim D^n} q(h) dS$. Hence,

$$
\begin{align*}
    \int_{S \sim D^n} \int_{A} q(h) \log \frac{1}{p(h)} dh dS &= \int_{A} \bar{q}(h) \log \frac{1}{p(h)} dh \\
    &= \int_{A} \bar{q}(h) \log \frac{1}{\bar{q}(h)} dh - \int_{A} \bar{q}(h) \log \frac{p(h)}{\bar{q}(h)} dh \\
    &\geq \int_{A} \bar{q}(h) \log \frac{1}{\bar{q}(h)} dh
\end{align*}
$$

2
where the last inequality follows from Jensen’s inequality. Since the equality is achieved for \( p(h) = \bar{q}(h) \) it follows that \( \mathcal{P}^* = \mathbb{E}_{S \sim \mathcal{D}^m}[Q] \).

Hence we have the following bound,

\[
KL(l(Q; S)\| l(Q; \mathcal{D})) \leq \frac{KL(Q\|\mathbb{E}_S[Q]) + \log(m+1)}{m} \tag{8}
\]

### 3.2 Distribution dependent prior for soft ERM

Consider the posterior coming out of the soft Empirical Risk Minimization:

\[
q(h) = \frac{1}{Z_Q} e^{-\gamma l(h; S)} \tag{9}
\]

where \( \gamma > 0 \) and \( Z_Q \) is a normalization constant so that \( q \) is a p.d.f. Define the distribution dependent prior,

\[
p(h) = \frac{1}{Z_P} e^{-\gamma l(h; \mathcal{D})} \tag{10}
\]

Note that although \( p(h) \) is not the expectation of \( q(h) \) over \( S \sim \mathcal{D}^m \), the exponent \( l(h; \mathcal{D}) = \mathbb{E}_{S \sim \mathcal{D}^m} l(h; S) \).

**Lemma 7.**

\[
KL(Q\|P) \leq \gamma (l(Q; \mathcal{D}) - l(Q; S)) - \gamma (l(P; \mathcal{D}) - l(P; S)) \tag{11}
\]

**Proof.**

\[
KL(Q\|P) = \mathbb{E}_Q \log \frac{q(h)}{p(h)} = \mathbb{E}_Q [\log \frac{e^{-\gamma l(h; S)}}{e^{-\gamma l(h; \mathcal{D})}}] - \log \frac{Z_Q}{Z_P} \tag{12}
\]

Note by definition that,

\[
\log \frac{Z_Q}{Z_P} = \log \int_{\mathcal{H}} \frac{1}{Z_P} e^{-\gamma l(h; S)} dh = \log \int_{\mathcal{H}} p(h) e^{\gamma (l(h; \mathcal{D}) - l(h; S))} dh = \log \mathbb{E}_P [e^{\gamma (l(h; \mathcal{D}) - l(h; S))}] \geq \mathbb{E}_P [\gamma (l(h; \mathcal{D}) - l(h; S))] = \gamma (l(P; \mathcal{D}) - l(P; S)) \tag{13}
\]

where the above inequality follows from Jensen’s inequality. Combining (12) and (13), the lemma follows.

**Theorem 8.** For the posterior \( Q \) with p.d.f as defined in (9), it holds that,

\[
KL(l(Q; S)\| l(Q; \mathcal{D})) \leq \sqrt{2} \gamma \sqrt{\log \left( \frac{m+1}{\delta} \right)} + \frac{\gamma^2}{2m^2} + \frac{\log(m+1)}{m} \tag{14}
\]

3
Proof. The PAC-Bayes bounds in (1) along with Lemma 2 gives,

\[ l(Q; D) - l(Q; S) \leq \sqrt{\frac{KL(Q||P) + \log \frac{m+1}{2}}{2m}} \]  

(15)

\[ |l(P; D) - l(P; S)| \leq \sqrt{\frac{KL(P||P) + \log \frac{m+1}{2}}{2m}} \]  

(16)

Combining Lemma 7 and (16) we have,

\[ KL(Q||P) \leq \gamma (l(Q; D) - l(Q; S)) - \gamma (l(P; D) - l(P; S)) \]

\[ \leq \gamma \sqrt{\frac{KL(Q||P) + \log \frac{m+1}{2}}{2m}} + \gamma \sqrt{\frac{\log \frac{m+1}{2}}{2m}} \]  

(17)

Let \( x = KL(Q||P) \) and \( L = \log \frac{m+1}{2} \). Then, \( x - \gamma \sqrt{\frac{L}{2m}} \leq \gamma \sqrt{\frac{x+L}{2m}} \). Assume \( x \geq \gamma \sqrt{\frac{L}{2m}} \). Squaring the previous inequality on both sides, we get that \( x \leq 2\gamma \sqrt{\frac{L}{2m}} + \frac{\gamma^2}{2m} \). Plugging this back into (1) the theorem follows. \( \square \)