CSE522, Winter 2011, Learning Theory	Lecture 14 - 02/17/2011

Derandomizing PAC-Bayes bounds and distribution dependent priors

Lecturer: Ofer Dekel *Scribe: Karthik Mohan*

1 Review of PAC Bayes Theorem

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

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

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

2 Derandomizing PAC Bayes bounds

Notation

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

Theorem 3. *Let $\mathcal{A}$ be any algorithm that takes in a sample $S \sim \mathcal{D}^m$ and outputs a hypothesis $h_{\tilde{w}}$ with $\tilde{w} \in [-1,1]^n$. Let $\mathcal{P}$ be uniformly distributed on $[-1,+1]^n$ and let $\mathcal{Q}$ be uniformly distributed on $(\tilde{w} \pm [-\frac{\gamma}{2n}, \frac{\gamma}{2n}]^n) \cap \mathcal{P}$. Then, $l_0(\tilde{w};\mathcal{D}) \leq l_\gamma(\tilde{w};S) + \sqrt{\frac{n\log(\frac{4n}{\gamma})+\log(\frac{m+1}{\delta})}{2m}}$.*

Note that the algorithm $\mathcal{A}$ 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:*

$$l_0(\tilde{w};\mathcal{D}) \leq l_{\frac{\gamma}{2}}(\mathcal{Q};\mathcal{D})$$
$$l_{\frac{\gamma}{2}}(\mathcal{Q};S) \leq l_\gamma(\tilde{w};S)$$

Proof. $\forall \hat{w} \in \mathcal{Q}, \forall x \in \mathcal{X}$ we have,

$$\begin{aligned} |\langle\tilde{w},x\rangle - \langle\hat{w},x\rangle| &= |\sum_{j=1}^n x_j(\tilde{w}_j - \hat{w}_j)| \\ &\leq \sum_{j=1}^n |x_j(\tilde{w}_j - \hat{w}_j)| \\ &\leq \sum_{j=1}^n |\tilde{w}_j - \hat{w}_j| \\ &\leq \sum_{j=1}^n \frac{\gamma}{2n} \\ &= \frac{\gamma}{2} \end{aligned} \tag{2}$$

Note that $l_\gamma(\tilde{w};(x,y)) = 0 \Rightarrow l_{\frac{\gamma}{2}}(\hat{w};(x,y)) = 0$. Indeed, let $y = 1$ then $\langle \tilde{w}, x\rangle \geq \gamma$. Hence from (2), $\langle \hat{w}, x\rangle \geq \langle \tilde{w}, x\rangle - \frac{\gamma}{2} \geq \frac{\gamma}{2}$. Similarly, $l_{\frac{\gamma}{2}}(\hat{w};(x,y)) = 0 \Rightarrow l_0(\tilde{w};(x,y)) = 0$. The previous two implications immediately imply that,

$$\begin{array}{ccccc} l_0(\tilde{w};\mathcal{D}) & \leq & l_{\frac{\gamma}{2}}(\hat{w};\mathcal{D}) & \leq & l_\gamma(\tilde{w};\mathcal{D}) \\ l_0(\tilde{w};S) & \leq & l_{\frac{\gamma}{2}}(\hat{w};S) & \leq & l_\gamma(\tilde{w};S) \end{array} \tag{3}$$

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

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

Proof. Note from definition that vol($\mathcal{P}$)$= 2^n$, vol($\mathcal{Q}$) $\geq (\frac{\gamma}{2n})^n$. Let $q(h), p(h)$ be the p.d.f of $\mathcal{Q}, \mathcal{P}$ respectively. $KL(\mathcal{Q}||\mathcal{P}) = \int_{h\in\mathcal{X}} q(h) \log \frac{q(h)}{p(h)} = \log \frac{\text{vol}(\mathcal{P})}{\text{vol}(\mathcal{Q})} \leq n \log \frac{4n}{\gamma}$. $\square$

Proof of Theorem 3. Note that (1) holds for $l = l_{\frac{\gamma}{2}}$. Along with Lemma 2, this implies that

$$l_{\frac{\gamma}{2}}(\mathcal{Q};\mathcal{D}) \quad \leq \quad l_{\frac{\gamma}{2}}(\mathcal{Q};S) + \sqrt{\frac{KL(\mathcal{Q}||\mathcal{P}) + \log \frac{m+1}{\delta}}{2m}} \tag{4}$$

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

$$\begin{aligned} l_0(\tilde{w};\mathcal{D}) &\leq l_{\frac{\gamma}{2}}(\mathcal{Q};\mathcal{D}) \\ &\leq l_{\frac{\gamma}{2}}(\mathcal{Q};S) + \sqrt{\frac{KL(\mathcal{Q}||\mathcal{P}) + \log \frac{m+1}{\delta}}{2m}} \\ &\leq l_\gamma(\tilde{w};S) + \sqrt{\frac{n \log \frac{4n}{\gamma} + \log \frac{m+1}{\delta}}{2m}} \end{aligned} \tag{5}$$

$\square$

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 Genereic prior

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

$$\mathcal{P}^* = \operatorname*{argmin}_{\mathcal{P}\in\mathcal{P}_\mathcal{H}} \mathbb{E}_{S\in\mathcal{D}^m}[KL(\mathcal{Q}||\mathcal{P})] \tag{6}$$

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

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

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

$$\begin{aligned} \int_{S\sim\mathcal{D}^m} \int_\mathcal{H} q(h) \log \frac{1}{p(h)} dh dS &= \int_\mathcal{H} \bar{q}(h) \log \frac{1}{p(h)} dh \\ &= \int_\mathcal{H} \bar{q}(h) \log \frac{1}{\bar{q}(h)} dh - \int_\mathcal{H} \bar{q}(h) \log \frac{p(h)}{\bar{q}(h)} dh \\ &\geq \int_\mathcal{H} \bar{q}(h) \log \frac{1}{\bar{q}(h)} dh \end{aligned} \tag{7}$$

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

Hence we have the following bound,

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

□

3.2 Distribution dependent prior for soft ERM

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

$$q(h) = \frac{1}{Z_{\mathcal{Q}}} e^{-\gamma l(h;S)}, \tag{9}$$

where $\gamma > 0$ and $Z_{\mathcal{Q}}$ is a normalization constant so that q is a p.d.f. Define the distribution dependent prior,

$$p(h) = \frac{1}{Z_{\mathcal{P}}} e^{-\gamma l(h;\mathcal{D})} \tag{10}$$

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.

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

Proof.

$$\begin{aligned} KL(\mathcal{Q}||\mathcal{P}) &= \mathbb{E}_{\mathcal{Q}} \log \tfrac{q(h)}{p(h)} \\ &= \mathbb{E}_{\mathcal{Q}}[\log \tfrac{e^{-\gamma l(h;S)}}{e^{-\gamma l(h;\mathcal{D})}}] - \log \tfrac{Z_{\mathcal{Q}}}{Z_{\mathcal{P}}} \\ &= \gamma(l(\mathcal{Q};\mathcal{D}) - l(\mathcal{Q};S)) - \log \tfrac{Z_{\mathcal{Q}}}{Z_{\mathcal{P}}} \end{aligned} \tag{12}$$

Note by definition that,

$$\begin{aligned} \log \tfrac{Z_{\mathcal{Q}}}{Z_{\mathcal{P}}} &= \log \int_{\mathcal{H}} \frac{1}{Z_{\mathcal{P}}} e^{-\gamma l(h;S)} dh \\ &= \log \int_{\mathcal{H}} p(h) e^{\gamma l(h;\mathcal{D})} e^{-\gamma l(h;S)} dh \\ &= \log \mathbb{E}_{\mathcal{P}}[e^{\gamma l(h;\mathcal{D})} e^{-\gamma l(h;S)}] \\ &\geq \mathbb{E}_{\mathcal{P}}[\gamma(l(h;\mathcal{D}) - l(h;S))] \\ &= \gamma(l(\mathcal{P};\mathcal{D}) - l(\mathcal{P};S)) \end{aligned} \tag{13}$$

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

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

$$KL(l(\mathcal{Q};S)||l(\mathcal{Q};\mathcal{D})) \leq \frac{\sqrt{2}\gamma}{m^{3/2}} \sqrt{\log\left(\frac{m+1}{\delta}\right)} + \frac{\gamma^2}{2m^2} + \frac{\log(\frac{m+1}{\delta})}{m} \tag{14}$$

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

$$l(\mathcal{Q};\mathcal{D}) - l(\mathcal{Q};S) \leq \sqrt{\frac{KL(\mathcal{Q}||\mathcal{P}) + \log\frac{m+1}{\delta}}{2m}} \tag{15}$$

$$|l(\mathcal{P};\mathcal{D}) - l(\mathcal{P};S)| \leq \sqrt{\frac{KL(\mathcal{P}||\mathcal{P}) + \log\frac{m+1}{\delta}}{2m}} \tag{16}$$

Combining Lemma 7 and (16) we have,

$$\begin{aligned} KL(\mathcal{Q}||\mathcal{P}) &\leq \gamma(l(\mathcal{Q};\mathcal{D}) - l(\mathcal{Q};S)) - \gamma(l(\mathcal{P};\mathcal{D}) - l(\mathcal{P};S)) \\ &\leq \gamma\sqrt{\tfrac{KL(\mathcal{Q}||\mathcal{P})+\log\frac{m+1}{\delta}}{2m}} + \gamma\sqrt{\tfrac{\log\frac{m+1}{\delta}}{2m}} \end{aligned} \tag{17}$$

Let $x = KL(\mathcal{Q}||\mathcal{P})$ and $L = \log\frac{m+1}{\delta}$. 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. □