1 Probably Approximately Correct (PAC) Bayes

Recall that in Theorem 2.1, we analyzed empirical risk minimization with a finite hypothesis class $\mathcal{F}$, i.e., $|\mathcal{F}| < +\infty$. Here, we will prove results for possibly infinite hypothesis classes. Although the PAC-Bayes framework is far more general, we will concentrate on the prediction problem as before, i.e., $(\forall f \in \mathcal{F}) f : \mathcal{X} \rightarrow \mathcal{Y}$.

Also, note that Theorem 2.1 could have been stated in a more general fashion and not only for the 0/1 risk $1[f(x) \neq y]$. Here, we will use a more general distortion function $d : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$. Under this setting, the 0/1 risk is given by $d(y, y') = 1[y \neq y']$.

Compared to Theorem 2.1, in the PAC-Bayes setting, instead of choosing a single predictor $f$, the learner picks a distribution $Q$. (It should be clear that the latter generalizes the previous setting.) Fix a prior distribution $P$ of support $\mathcal{F}$. After observing a training set of $n$ samples, the task of the learner is to choose a posterior distribution $Q$ of support $\mathcal{F}$. PAC-Bayes guarantees are given with respect to a prior distribution $P$ and simultaneously for all posterior distributions $Q$.

By looking at the following theorem statement, the reader would also notice the relationship between PAC-Bayes and KL-regularization.

**Theorem 4.1.** Assume that $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are arbitrary domains. Assume that the pair $(x, y)$ follows an arbitrary distribution $\mathcal{D}$. Assume that $(x_1, y_1) \ldots (x_n, y_n)$ are $n$ i.i.d. samples drawn from the distribution $\mathcal{D}$. Let $\mathcal{F}$ be a possibly infinite set of predictor functions $(\forall f \in \mathcal{F}) f : \mathcal{X} \rightarrow \mathcal{Y}$. Let $d : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ be a distortion function. Fix a prior distribution $P$ of support $\mathcal{F}$. The expected risk, the Gibbs expected risk and its minimizer are defined as:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[d(f(x), y)]$$
$$\mathcal{R}(Q) = \mathbb{E}_{f \sim Q}[\mathcal{R}(f)]$$
$$\mathcal{Q} = \arg \min_{Q} \mathcal{R}(Q)$$
The empirical risk, the Gibbs empirical risk and its minimizer are defined as:

\[
\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} d(f(x_i), y_i)
\]

\[
\hat{R}(Q) = \mathbb{E}_{f \sim Q}[\hat{R}(f)]
\]

\[
\hat{Q} = \arg \min_{Q} \hat{R}(Q) + \frac{1}{\sqrt{n}} \mathbb{KL}(Q \| P)
\]

Fix \( \delta \in (0, 1) \). With probability at least \( 1 - \delta \) over the choice of \( n \) samples, simultaneously for all posterior distributions \( Q \) of support \( F \), we have:

\[
\hat{R}(Q) - \overline{R}(Q) \leq \frac{1}{\sqrt{n}} \left( \mathbb{KL}(Q \| P) + \log \frac{e^{1/8}}{\delta} \right)
\]

Furthermore, with probability at least \( 1 - \delta \) over the choice of \( n \) samples, we have:

\[
\overline{R}(\hat{Q}) - \overline{R}(\overline{Q}) \leq \frac{2}{\sqrt{n}} \left( \mathbb{KL}(\overline{Q} \| P) + \log \frac{2e^{1/8}}{\delta} \right)
\]

Proof. Let \( t > 0 \) be a constant. Let \( S = (x_1, y_1) \ldots (x_n, y_n) \). Since each \( (x_i, y_i) \) for \( i = 1 \ldots n \) is sampled independently from \( \mathcal{D} \), we write \( S \sim \mathcal{D}^n \).

Note that the random variable \( \mathbb{E}_{f \sim P}[e^{t(\hat{R}(f) - \overline{R}(f))}] \) is non-negative. By Markov’s inequality,\(^1\) with probability at least \( 1 - \delta \) over the choice of \( n \) samples:

\[
\mathbb{E}_{f \sim P}[e^{t(\hat{R}(f) - \overline{R}(f))}] \leq \frac{1}{\delta} \mathbb{E}_{S \sim \mathcal{D}^n} \left[ \mathbb{E}_{f \sim P}[e^{t(\hat{R}(f) - \overline{R}(f))}] \right]
\]

\[
= \frac{1}{\delta} \mathbb{E}_{f \sim P} \left[ \mathbb{E}_{S \sim \mathcal{D}^n}[e^{t(\hat{R}(f) - \overline{R}(f))}] \right]
\]

By taking the logarithm on each side and since \( \mathbb{E}_{f \sim P}[\phi(f)] = \mathbb{E}_{f \sim Q} \left[ \frac{p(f)}{q(f)} \phi(f) \right] \) for every distribution \( Q \) and function \( \phi : F \rightarrow \mathbb{R} \), we have:

\[
(\forall Q) \quad \log \mathbb{E}_{f \sim Q} \left[ \frac{p(f)}{q(f)} e^{t(\hat{R}(f) - \overline{R}(f))} \right] \leq \log \left( \frac{1}{\delta} \mathbb{E}_{f \sim P} \left[ \mathbb{E}_{S \sim \mathcal{D}^n}[e^{t(\hat{R}(f) - \overline{R}(f))}] \right] \right)
\]

(1)

By using Jensen’s inequality on the (concave) log function,\(^2\) we can lower-bound the left-hand side of the above expression:

\[
(\forall Q) \quad \log \mathbb{E}_{f \sim Q} \left[ \frac{p(f)}{q(f)} e^{t(\hat{R}(f) - \overline{R}(f))} \right] \geq \mathbb{E}_{f \sim Q} \left[ \log \left( \frac{p(f)}{q(f)} e^{t(\hat{R}(f) - \overline{R}(f))} \right) \right]
\]

\[
= \mathbb{E}_{f \sim Q} \left[ \log \frac{p(f)}{q(f)} \right] + \mathbb{E}_{f \sim Q} \left[ t(\hat{R}(f) - \overline{R}(f)) \right]
\]

\[
= -\mathbb{KL}(Q \| P) + t \mathbb{E}_{f \sim Q}[\hat{R}(f) - \overline{R}(f)]
\]

\[
= -\mathbb{KL}(Q \| P) + t(\hat{R}(Q) - \overline{R}(Q))
\]

(2)

\(^1\)Note that by Markov’s inequality (Theorem 1.1), we have \( P[x \geq a] \leq \mathbb{E}[x]/a \). Make \( \mathbb{E}[x]/a = \delta \), we have \( P[x \geq \mathbb{E}[x]/\delta] \leq \delta \). Thus \( P[x < \mathbb{E}[x]/\delta] \geq 1 - \delta \).

\(^2\)I encourage you to look for Jensen’s inequality, the proof follows from the definition of convexity/concavity.
It remains to bound the term \( E_{S \sim D^n}[e^{t(\hat{R}(f) - \overline{R}(f))}] \) in eq.(1) by independence and Lemma 2.1 (Hoeffding’s lemma), in the following fashion:

\[
E_{S \sim D^n}[e^{t(\hat{R}(f) - \overline{R}(f))}] = E_{S \sim D^n}\left[e^{t\left(\frac{1}{n}\sum_{i=1}^{n} d(f(x_i),y_i) - \overline{R}(f)\right)}\right]
\]

\[
= \prod_{i=1}^{n} E_{(x_i,y_i) \sim D}\left[e^{\frac{t}{n}d(f(x_i),y_i) - \overline{R}(f)}\right]
\]

\[
\leq \prod_{i=1}^{n} e^{\frac{t^2}{n}}
\]

\[
= e^{\frac{t^2}{n}}
\]

By the above, eq.(1) and eq.(2), we have:

\[
(\forall Q) \ \hat{R}(Q) - \overline{R}(Q) \leq \frac{1}{t} \left(\text{KL}(Q\|P) + \log \frac{e^{t^2}}{\delta}\right)
\]

By setting \( t = \sqrt{n} \), we prove our first claim. For \( \varepsilon = \frac{1}{\sqrt{n}} \log 2e^{1/8}/\delta \), by the union bound we also have:

\[
(\forall Q) \ \left|\hat{R}(Q) - \overline{R}(Q)\right| \leq \frac{1}{\sqrt{n}} \text{KL}(Q\|P) + \varepsilon
\]

Finally since \( \hat{Q} \) minimizes \( \hat{R}(\cdot) + \frac{1}{\sqrt{n}} \text{KL}(\cdot\|P) \), we know that the following holds:

\[
\hat{R}(\hat{Q}) + \frac{1}{\sqrt{n}} \text{KL}(\hat{Q}\|P) \leq \hat{R}(\overline{Q}) + \frac{1}{\sqrt{n}} \text{KL}(\overline{Q}\|P)
\]

and therefore:

\[
\hat{R}(\hat{Q}) - \overline{R}(\overline{Q}) \leq \hat{R}(\hat{Q}) - \hat{R}(\overline{Q}) + \frac{1}{\sqrt{n}} \text{KL}(\hat{Q}\|P) + \varepsilon - \hat{R}(\overline{Q}) + \frac{1}{\sqrt{n}} \text{KL}(\overline{Q}\|P) + \varepsilon
\]

\[
\leq \hat{R}(\hat{Q}) - \hat{R}(\overline{Q}) + \frac{1}{\sqrt{n}} \text{KL}(\hat{Q}\|P) + \varepsilon - \hat{R}(\overline{Q}) + \frac{1}{\sqrt{n}} \text{KL}(\overline{Q}\|P) + \varepsilon
\]

\[
\leq \frac{2}{\sqrt{n}} \text{KL}(\overline{Q}\|P) + 2\varepsilon
\]

which proves our second claim.

2 Application: Structured Prediction

Assume we want to predict a parsing tree \( y \) from a sentence \( x \). A decoder is a machine for predicting the structured output \( y \) given the observed input \( x \).

We assume a distribution \( D \) on pairs \((x,y)\) where \( x \in X \) is the observed input and \( y \in Y \) is the latent structured output, i.e., \((x,y) \sim D\). We also assume that
we have a training set $S$ of $n$ i.i.d. samples drawn from the distribution $D$, i.e., $S \sim D^n$, and thus $|S| = n$.

We let $\mathcal{Y}(x) \neq \emptyset$ denote the countable set of feasible decodings of $x$. In other words, a decoder is a function that maps $x$ to an element in $\mathcal{Y}(x)$. We assume a fixed mapping $\phi$ from pairs to feature vectors, i.e., for any pair $(x,y)$ we have the feature vector $\phi(x,y) \in \mathbb{R}^k$. For a parameter $w \in \mathbb{R}^k$, we consider linear decoders of the form:

$$f_w(x) \equiv \arg \max_{y \in \mathcal{Y}(x)} \langle \phi(x,y), w \rangle$$  \hspace{1cm} (3)

We will continue using the distortion function $d : \mathcal{Y} \times \mathcal{Y} \to [0,1]$. We define the margin $m(x,y,y',w)$ as the amount by which $y$ is preferable to $y'$ under the parameter $w$. More formally:

$$m(x,y,y',w) \equiv \langle \phi(x,y), w \rangle - \langle \phi(x,y'), w \rangle$$

Let $c(p,x,y)$ be a nonnegative integer that gives the number of times that the part $p \in \mathcal{P}$ appears in the pair $(x,y)$. For a part $p \in \mathcal{P}$, we define the feature $p$ as follows:

$$\phi_p(x,y) \equiv c(p,x,y)$$

We let $\mathcal{P}(x) \neq \emptyset$ denote the set of $p \in \mathcal{P}$ such that there exists $y \in \mathcal{Y}(x)$ with $c(p,x,y) > 0$. We define the Hamming distance $H$ as follows:

$$H(x,y,y') \equiv \sum_{p \in \mathcal{P}(x)} |c(p,x,y) - c(p,x,y')|$$

Let $\mathcal{P}$ be a zero-mean and unit-variance Gaussian distribution of parameters $w' \in \mathbb{R}^k$. Let $\alpha > 0$ and let $\mathcal{Q}(w)$ be a unit-variance Gaussian distribution centered at $\alpha w$ of parameters $w' \in \mathbb{R}^k$. Define the Gibbs expected and empirical risks as before:

$$\mathbb{R}(\mathcal{Q}(w)) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y,f_{w'}(x))] \right]$$

$$\hat{\mathbb{R}}(\mathcal{Q}(w)) = \frac{1}{n} \sum_{(x,y) \in S} \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y,f_{w'}(x))]$$

First, we prove an auxiliary lemma.

**Lemma 4.1** (Adapted from Lemma 6 in [2]). Assume that there exists a finite integer value $\ell$ such that $|\bigcup_{(x,y) \in S} \mathcal{P}(x)| \leq \ell$. Let $\mathcal{Q}(w)$ be a unit-variance Gaussian distribution centered at $\alpha w$ for $\alpha = \sqrt{2 \log(2n\ell/||w||_2^2)}$. Simultaneously for all $(x,y) \in S$ and $w \in \mathbb{R}^k$, we have:

$$\mathbb{P}_{w' \sim \mathcal{Q}(w)}[H(x,y,f_{w'}(x)) - m(x,y,f_{w'}(x),w) < 0] \leq \frac{||w||_2^2}{n}$$
or equivalently:

\[ P_{w' \sim Q(w)}[H(x, y, f_{w'}(x)) - m(x, y, f_{w'}(x), w) \geq 0] \geq 1 - \frac{\|w\|^2}{n} \tag{4} \]

**Proof.** First, note that \( w' - \alpha w \) is a zero-mean and unit-variance Gaussian random vector. By Corollary 1.2, for any \( p \in P(x) \) we have:

\[ P_{w' \sim Q(w)}[|w'_p - \alpha w_p| \geq \varepsilon] \leq 2e^{-\varepsilon^2/2} \]

By the union bound and setting \( \varepsilon = \alpha = \sqrt{2 \log (2n\ell/\|w\|^2)} \), we have:

\[ P_{w' \sim Q(w)}[\exists p \in \cup_{(x,y) \in S} P(x)] [w'_p - \alpha w_p | \geq \alpha] \leq 2|\cup_{(x,y) \in S} P(x)|e^{-\alpha^2/2} \]

or equivalently:

\[ P_{w' \sim Q(w)}[\forall p \in \cup_{(x,y) \in S} P(x)] [w'_p - \alpha w_p | < \alpha] \geq 1 - \frac{\|w\|^2}{n} \]

The high-probability statement in eq.(4) can be written as:

\[ y' = f_{w'}(x) \Rightarrow H(x, y, y') - m(x, y, y', w) \geq 0 \]

Next, we use proof by contradiction, i.e., we will assume:

\[ y' = f_{w'}(x) \text{ and } H(x, y, y') - m(x, y, y', w) < 0 \]

and arrive to a contradiction \( y' \neq f_{w'}(x) \). From the above, we have:

\[
m(x, y, y', w') = m(x, y, y', \alpha w + (w' - \alpha w))
\]

\[
= \alpha m(x, y, y', w) - \langle \phi(x, y) - \phi(x, y'), \alpha w - w' \rangle
\]

\[
> \alpha H(x, y, y') - \langle \phi(x, y) - \phi(x, y'), \alpha w - w' \rangle
\]

\[
= \alpha H(x, y, y') - \sum_{p \in P(x)} (c(p, x, y) - c(p, x, y'))(\alpha w_p - w'_p)
\]

\[
\geq \alpha H(x, y, y') - \sum_{p \in P(x)} |c(p, x, y) - c(p, x, y')|\alpha w_p - w'_p
\]

\[
\geq \alpha H(x, y, y') - \sum_{p \in P(x)} |c(p, x, y) - c(p, x, y')|\alpha
\]

\[
= 0
\]

Note that \( m(x, y, y', w') > 0 \) if and only if \( \langle \phi(x, y), w' \rangle > \langle \phi(x, y'), w' \rangle \). Therefore \( y' \neq f_{w'}(x) \) since it does not maximize \( \langle \phi(x, \cdot), w \rangle \) as defined in eq.(3). Thus, we prove our claim. \( \square \)
Next, we prove the main result.

**Theorem 4.2** (Adapted from Theorem 2 in [2]). Assume that there exists a finite integer value $\ell$ such that $|\bigcup_{(x,y) \in S} P(x)| \leq \ell$. Let the prior distribution $P$ be a zero-mean and unit-variance Gaussian distribution of parameters $w' \in \mathbb{R}^k$. Fix $\delta \in (0, 1)$. With probability at least $1 - \delta$ over the choice of $n$ samples, simultaneously for all parameters $w \in \mathbb{R}^k$ and unit-variance Gaussian posterior distributions $Q(w)$ centered at $\alpha w$ for $\alpha = \sqrt{2 \log \left( \frac{2n\ell/\|w\|^2_2}{\log e} \right)}$, we have:

$$
\mathcal{R}(Q(w)) \leq \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in \mathcal{Y}(x)} d(y, \hat{y}) \left( H(x, y, \hat{y}) - m(x, y, \hat{y}, w) \geq 0 \right)
$$

$$
+ \frac{\|w\|^2_2}{n} + \frac{1}{\sqrt{n}} \left( \frac{\|w\|^2_2 \log \left( \frac{2n\ell/\|w\|^2_2}{\log e} \right)}{2} + \log \frac{e^{1/8}}{\delta} \right)
$$

**Proof.** Fix $\alpha > 0$. Since $P$ is a zero-mean and unit-variance Gaussian distribution and since $Q(w)$ be a unit-variance Gaussian distribution centered at $\alpha w$, we have by Lecture 3, eq.(3):

$$
p(w) = \frac{1}{\sqrt{2\pi}} e^{-\|w\|^2/2}
$$

$$
g(w'|w) = \frac{1}{\sqrt{2\pi}} e^{-\|w' - \alpha w\|^2/2}
$$

$$
\text{KL}(Q(w)||P) = \frac{\|w\|^2_2 \alpha^2}{2}
$$

Fix $\delta \in (0, 1)$. By Theorem 4.1 with probability at least $1 - \delta$ over the choice of $n$ samples, simultaneously for all parameters $w \in \mathbb{R}^k$, and unit-variance Gaussian posterior distributions $Q(w)$ centered at $\alpha w$, we have:

$$
\mathcal{R}(Q(w)) \leq \hat{R}(Q(w)) + \frac{1}{\sqrt{n}} \left( \text{KL}(Q(w)||P) + \log \frac{e^{1/8}}{\delta} \right)
$$

$$
= \hat{R}(Q(w)) + \frac{1}{\sqrt{n}} \left( \frac{\|w\|^2_2 \alpha^2}{2} + \log \frac{e^{1/8}}{\delta} \right)
$$

Thus, an upper bound of $\hat{R}(Q(w))$ would lead to an upper bound of $\mathcal{R}(Q(w))$. In order to upper-bound $\hat{R}(Q(w))$, we can upper-bound each of its summands, i.e., we can upper-bound $E_{w' \sim Q(w)}[d(y, f_w(x))]$ for each $(x, y) \in S$. For clarity of presentation, define:

$$
u(x, y, y', w) \equiv H(x, y, y') - m(x, y, y', w)
$$
Let $u \equiv u(x, y, f_{w'}(x), w)$. Simultaneously for all $(x, y) \in S$, we have:

\[
\mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))] = \mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))] 1[u \geq 0] + d(y, f_{w'}(x)) 1[u < 0] \\
\leq \mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))] 1[u \geq 0] + 1[u < 0] \\
= \mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))] 1[u \geq 0] + \mathbb{P}_{w' \sim Q(w)}[u < 0] \\
\leq \mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))] 1[u \geq 0] + \|w\|_2^2/n \\
= \mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))] 1[u \geq 0] + \|w\|_2^2/n \\
\leq \max_{\tilde{y} \in \mathcal{Y}(x)} d(y, \tilde{y}) 1[u(x, y, \tilde{y}, w) \geq 0] + \|w\|_2^2/n
\]

(6.a)

where the step in eq.(6.a) holds since $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. The step in eq.(6.b) follows from Lemma 4.1. Let $g : \mathcal{Y} \to [0, 1]$ be some arbitrary function, the step in eq.(6.c) uses the fact that $\mathbb{E}_y[g(y)] \leq \max_y g(y)$.

By eq.(5) and eq.(6.c), we prove our claim. \qed

References
