Hands-On Learning Theory
Fall 2016, Lecture 4

Jean Honorio jhonorio@purdue.edu

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 of the prediction problem as before, i.e., $(\forall f \in \mathcal{F})\ f : \mathcal{X} \to \mathcal{Y}$.

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 $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. 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 $\mathcal{Q}$. (It should be clear that the latter generalizes the previous setting.) Fix a prior distribution $\mathcal{P}$ of support $\mathcal{F}$. After observing a training set of n samples, the task of the learner is to choose a posterior distribution $\mathcal{Q}$ of support $\mathcal{F}$. PAC-Bayes guarantees are given with respect to a prior distribution $\mathcal{P}$ and simultaneously for all posterior distributions $\mathcal{Q}$.

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 $(\forall f \in \mathcal{F})\ f : \mathcal{X} \to \mathcal{Y}$. Let $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a distortion function. Fix a prior distribution $\mathcal{P}$ of support $\mathcal{F}$. The expected risk, the Gibbs expected risk and its minimizer are defined as:*

$$
\begin{aligned}
\overline{R}(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[d(f(x), y)] \\
\overline{R}(\mathcal{Q}) &= \mathbb{E}_{f\sim\mathcal{Q}}[\overline{R}(f)] \\
\overline{\mathcal{Q}} &= \arg\min_{\mathcal{Q}} \overline{R}(\mathcal{Q})
\end{aligned}
$$

The empirical risk, the Gibbs empirical risk and its minimizer are defined as:

$$\begin{aligned}
\widehat{R}(f) &= \frac{1}{n}\sum_{i=1}^{n} d(f(x_i), y_i) \\
\widehat{R}(\mathcal{Q}) &= \mathbb{E}_{f\sim\mathcal{Q}}[\widehat{R}(f)] \\
\widehat{\mathcal{Q}} &= \arg\min_{\mathcal{Q}} \widehat{R}(\mathcal{Q}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\mathcal{Q}\|\mathcal{P})
\end{aligned}$$

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

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}_{f\sim\mathcal{P}}[e^{t(\widehat{R}(f)-\overline{R}(f))}] &\leq \frac{1}{\delta}\mathbb{E}_{S\sim\mathcal{D}^n}\left[\mathbb{E}_{f\sim\mathcal{P}}[e^{t(\widehat{R}(f)-\overline{R}(f))}]\right] \\
&= \frac{1}{\delta}\mathbb{E}_{f\sim\mathcal{P}}\left[\mathbb{E}_{S\sim\mathcal{D}^n}[e^{t(\widehat{R}(f)-\overline{R}(f))}]\right]
\end{aligned}$$

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

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

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

$$\begin{aligned}
(\forall\mathcal{Q})\ \log\mathbb{E}_{f\sim\mathcal{Q}}\left[\frac{p(f)}{q(f)}e^{t(\widehat{R}(f)-\overline{R}(f))}\right] &\geq \mathbb{E}_{f\sim\mathcal{Q}}\left[\log\left(\frac{p(f)}{q(f)}e^{t(\widehat{R}(f)-\overline{R}(f))}\right)\right] \\
&= \mathbb{E}_{f\sim\mathcal{Q}}\left[\log\frac{p(f)}{q(f)}\right] + \mathbb{E}_{f\sim\mathcal{Q}}\left[t(\widehat{R}(f)-\overline{R}(f))\right] \\
&= -\mathbb{KL}(\mathcal{Q}\|\mathcal{P}) + t\ \mathbb{E}_{f\sim\mathcal{Q}}[\widehat{R}(f)-\overline{R}(f)] \\
&= -\mathbb{KL}(\mathcal{Q}\|\mathcal{P}) + t(\widehat{R}(\mathcal{Q})-\overline{R}(\mathcal{Q})) \qquad (2)
\end{aligned}$$

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

[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 $\mathbb{E}_{S\sim\mathcal{D}^n}[e^{t(\widehat{R}(f)-\overline{R}(f))}]$ in eq.(1) by independence and Lemma 2.1 (Hoeffding's lemma), in the following fashion:

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^n}[e^{t(\widehat{R}(f)-\overline{R}(f))}] &= \mathbb{E}_{S\sim\mathcal{D}^n}\left[e^{t(\frac{1}{n}\sum_{i=1}^n d(f(x_i),y_i)-\overline{R}(f))}\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}^n}\left[\prod_{i=1}^n e^{\frac{t}{n}(d(f(x_i),y_i)-\overline{R}(f))}\right] \\
&= \prod_{i=1}^n \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}\left[e^{\frac{t}{n}(d(f(x_i),y_i)-\overline{R}(f))}\right] \\
&\leq \prod_{i=1}^n e^{\frac{1}{8}(\frac{t}{n})^2} \\
&= e^{\frac{t^2}{8n}}
\end{aligned}$$

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

$$(\forall \mathcal{Q})\ \widehat{R}(\mathcal{Q}) - \overline{R}(\mathcal{Q}) \leq \frac{1}{t}\left(\mathbb{KL}(\mathcal{Q}\|\mathcal{P}) + \log\frac{e^{\frac{t^2}{8n}}}{\delta}\right)$$

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

$$(\forall \mathcal{Q})\ \left|\widehat{R}(\mathcal{Q}) - \overline{R}(\mathcal{Q})\right| \leq \frac{1}{\sqrt{n}}\ \mathbb{KL}(\mathcal{Q}\|\mathcal{P}) + \varepsilon$$

Finally since $\widehat{\mathcal{Q}}$ minimizes $\widehat{R}(\cdot) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\cdot\|\mathcal{P})$, we know that the following holds $\widehat{R}(\widehat{\mathcal{Q}}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\widehat{\mathcal{Q}}\|\mathcal{P}) \leq \widehat{R}(\overline{\mathcal{Q}}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\overline{\mathcal{Q}}\|\mathcal{P})$ and therefore:

$$\begin{aligned}
\overline{R}(\widehat{\mathcal{Q}}) - \overline{R}(\overline{\mathcal{Q}}) &\leq \widehat{R}(\widehat{\mathcal{Q}}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\widehat{\mathcal{Q}}\|\mathcal{P}) + \varepsilon - \widehat{R}(\overline{\mathcal{Q}}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\overline{\mathcal{Q}}\|\mathcal{P}) + \varepsilon \\
&\leq \widehat{R}(\overline{\mathcal{Q}}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\overline{\mathcal{Q}}\|\mathcal{P}) + \varepsilon - \widehat{R}(\overline{\mathcal{Q}}) + \frac{1}{\sqrt{n}}\ \mathbb{KL}(\overline{\mathcal{Q}}\|\mathcal{P}) + \varepsilon \\
&\leq \frac{2}{\sqrt{n}}\ \mathbb{KL}(\overline{\mathcal{Q}}\|\mathcal{P}) + 2\varepsilon
\end{aligned}$$

which proves our second claim. $\square$

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 $\mathcal{D}$ on pairs (x, y) where $x \in \mathcal{X}$ is the observed input and $y \in \mathcal{Y}$ 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 $\mathcal{D}$, i.e., $S \sim \mathcal{D}^n$, 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 ϕ 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:

$$f_w(x) \equiv \operatorname*{arg\,max}_{y \in \mathcal{Y}(x)} \langle \phi(x, y), w \rangle \tag{3}$$

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 $p \in P$ appears in the pair (x, y). For a part $p \in P$, we define the feature p as follows:

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

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

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

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 αw of parameters $w' \in \mathbb{R}^k$. Define the Gibbs expected and empirical risks as before:

$$\begin{aligned} \overline{R}(\mathcal{Q}(w)) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{E}_{w'\sim\mathcal{Q}(w)}[d(y, f_{w'}(x))]\right] \\ \widehat{R}(\mathcal{Q}(w)) &= \frac{1}{n} \sum_{(x,y)\in S} \mathbb{E}_{w'\sim\mathcal{Q}(w)}[d(y, f_{w'}(x))] \end{aligned}$$

First, we prove an auxiliary lemma.

Lemma 4.1 (Adapted from Lemma 6 in [2])**.** *Assume that there exists a finite integer value ℓ such that $|\cup_{(x,y)\in S} P(x)| \leq \ell$. Let $\mathcal{Q}(w)$ be a unit-variance Gaussian distribution centered at αw 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:*

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

or equivalently:

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

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:

$$\mathbb{P}_{w' \sim \mathcal{Q}(w)}[|w'_p - \alpha w_p| \geq \varepsilon] \leq 2e^{-\varepsilon^2/2}$$

By the union bound and setting $\varepsilon = \alpha = \sqrt{2\log\left(2n\ell/\|w\|_2^2\right)}$, we have:

$$\begin{aligned}
\mathbb{P}_{w' \sim \mathcal{Q}(w)}[(\exists p \in \cup_{(x,y)\in S} P(x)) \; |w'_p - \alpha w_p| \geq \alpha] &\leq 2|\cup_{(x,y)\in S} P(x)|e^{-\alpha^2/2} \\
&= |\cup_{(x,y)\in S} P(x)|\frac{\|w\|^2}{\ell n} \\
&\leq \frac{\|w\|^2}{n}
\end{aligned}$$

or equivalently:

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

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:

$$\begin{aligned}
m(x,y,y',w') &= m(x,y,y',\alpha w + (w' - \alpha w)) \\
&= \alpha m(x,y,y',w) - \langle \phi(x,y) - \phi(x,y'), \alpha w - w' \rangle \\
&> \alpha H(x,y,y') - \langle \phi(x,y) - \phi(x,y'), \alpha w - w' \rangle \\
&= \alpha H(x,y,y') - \sum_{p \in P(x)} (c(p,x,y) - c(p,x,y'))(\alpha w_p - w'_p) \\
&\geq \alpha H(x,y,y') - \sum_{p \in P(x)} |c(p,x,y) - c(p,x,y')||\alpha w_p - w'_p| \\
&\geq \alpha H(x,y,y') - \sum_{p \in P(x)} |c(p,x,y) - c(p,x,y')|\alpha \\
&= 0
\end{aligned}$$

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. □

Next, we prove the main result.

Theorem 4.2 (Adapted from Theorem 2 in [2])**.** *Assume that there exists a finite integer value ℓ such that $|\cup_{(x,y)\in S} P(x)| \leq \ell$. Let the prior distribution $\mathcal{P}$ 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 $\mathcal{Q}(w)$ centered at αw for $\alpha = \sqrt{2\log(2n\ell/\|w\|_2^2)}$, we have:*

$$\begin{aligned}\overline{R}(\mathcal{Q}(w)) \leq \frac{1}{n} \sum_{(x,y)\in S} \max_{\hat{y}\in\mathcal{Y}(x)} d(y,\hat{y})\ 1[H(x,y,\hat{y}) - m(x,y,\hat{y},w) \geq 0] \\ + \frac{\|w\|_2^2}{n} + \frac{1}{\sqrt{n}}\left(\frac{\|w\|_2^2 \log(2n\ell/\|w\|_2^2)}{2} + \log\frac{e^{1/8}}{\delta}\right)\end{aligned}$$

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

$$\begin{aligned}p(w) &= \frac{1}{\sqrt{2\pi}} e^{-\|w\|^2/2} \\ q(w'|w) &= \frac{1}{\sqrt{2\pi}} e^{-\|w'-\alpha w\|^2/2} \\ \mathbb{KL}(\mathcal{Q}(w)\|\mathcal{P}) &= \frac{\|w\|_2^2\alpha^2}{2}\end{aligned}$$

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 $\mathcal{Q}(w)$ centered at αw, we have:

$$\begin{aligned}\overline{R}(\mathcal{Q}(w)) &\leq \widehat{R}(\mathcal{Q}(w)) + \frac{1}{\sqrt{n}}\left(\mathbb{KL}(\mathcal{Q}(w)\|\mathcal{P}) + \log\frac{e^{1/8}}{\delta}\right) \\ &= \widehat{R}(\mathcal{Q}(w)) + \frac{1}{\sqrt{n}}\left(\frac{\|w\|_2^2\alpha^2}{2} + \log\frac{e^{1/8}}{\delta}\right) \qquad (5)\end{aligned}$$

Thus, an upper bound of $\widehat{R}(\mathcal{Q}(w))$ would lead to an upper bound of $\overline{R}(\mathcal{Q}(w))$. In order to upper-bound $\widehat{R}(\mathcal{Q}(w))$, we can upper-bound each of its summands, i.e., we can upper-bound $\mathbb{E}_{w'\sim\mathcal{Q}(w)}[d(y, f_{w'}(x))]$ 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:

$$
\begin{aligned}
\mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y, f_{w'}(x)] &= \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y, f_{w'}(x)) \ 1[u \geq 0] + d(y, f_{w'}(x)) \ 1[u < 0]] \\
&\leq \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y, f_{w'}(x)) \ 1[u \geq 0] + 1[u < 0]] \qquad (6.a) \\
&= \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y, f_{w'}(x)) \ 1[u \geq 0]] + \mathbb{P}_{w' \sim \mathcal{Q}(w)}[u < 0] \\
&\leq \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y, f_{w'}(x)) \ 1[u \geq 0]] + \|w\|_2^2/n \qquad (6.b) \\
&= \mathbb{E}_{w' \sim \mathcal{Q}(w)}[d(y, f_{w'}(x)) \ 1[u(x, y, f_{w'}(x), w) \geq 0]] + \|w\|_2^2/n \\
&\leq \max_{\hat{y} \in \mathcal{Y}(x)} d(y, \hat{y}) \ 1[u(x, y, \hat{y}, w) \geq 0] + \|w\|_2^2/n \qquad (6.c)
\end{aligned}
$$

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. □

References

[1] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J. Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(Apr):787–860, 2015.

[2] D. McAllester. Generalization bounds and consistency. In *Predicting Structured Data*, pages 247–261. MIT Press, 2007.