10 Stability and PAC-Bayes Bounds

In this lecture, we will cover two different types for generalization bounds. The first one is based on uniform stability which is based on a small modification of the proof we did for Rademacher complexity.

10.1 Stability based generalization bounds

We define the algorithmic stability as follows.

Definition 41 (Uniform stability)**.** *We say that an empirical risk minimization algorithm given as*

$$\hat{f}_{\mathcal{D}} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) \quad \textit{for } \mathcal{D} = \{z_1, z_2, ..., z_n\} \in \mathcal{Z}^n,$$

is uniformly β-stable if for all training sets $\mathcal{D} \in \mathcal{Z}^n$, and their j-th sample perturbations denoted by $\mathcal{D}_j = \{z_1, .., z_j', .., z_n\}$, we have

$$\sup_{z \in \mathcal{Z}} \Big| \ell(z, \hat{f}_{\mathcal{D}}) - \ell(z, \hat{f}_{\mathcal{D}_j}) \Big| \leq \beta. \tag{10.1}$$

Remark. It should be understood that smaller β corresponds to a more stable algorithm.

- We emphasize that the above notion is not for a specific empirical risk minimizer, rather for the **minimization algorithm** which is why we refer to it as algorithmic stability. The difference is that $\hat{f}$ is data specific whereas an algorithm outputs different minimizers for different data inputs. We make this dependence explicit by using the same notation $\hat{f}_{\mathcal{D}}$.
- Moreover, the above condition (10.1) is uniform over data $z \in \mathcal{Z}$, and all possible datasets $\mathcal{D}$ and their perturbations $\mathcal{D}_j'$, for all j. Needless to say, it is a very strong assumption, but can be easily verified for several algorithms of interest.

Example. [Revisiting Gaussian mean estimation]

- Consider the Gaussian mean estimation problem where we observe n data points $\mathcal{D} = \{z_1, z_2, ..., z_n\}$. Standard assumption in this problem is $z_i \sim \mathcal{N}(\mu, \sigma^2 I)$, when coupled with an ℓ_2-regularization, the MLE yields the following algorithm

$$\hat{\mu}_{\mathcal{D}} = \operatorname*{argmin}_{\mu \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \|z_i - \mu\|^2 = \frac{1}{n} \sum_{i=1}^{n} z_i \triangleq \bar{z},$$

  where we denote the sample mean estimator with $\bar{z}$.
- In this problem, we notice that the loss function is given as $\ell(z, \mu) = \|z - \mu\|^2$. For simplicity, lets assume that data points are uniformly bounded, i.e.

$$\|z_i\| \leq \kappa \text{ almost surely.}$$

This assumptions is clearly violated for Gaussian data; however, similar bounds can be obtained under high probability. Denoting the sample mean estimator over the perturbed data $\mathcal{D}'_j$ with $\bar{z}'_j$, we verify the uniform stability condition as follows. For $z \in \mathcal{Z}$, we write

$$\begin{aligned}\big|\ell(z,\hat{\mu}_{\mathcal{D}}) - \ell(z,\hat{\mu}_{\mathcal{D}_j})\big| =& \big|\|z-\hat{\mu}_{\mathcal{D}}\|^2 - \|z-\hat{\mu}_{\mathcal{D}_j}\|^2\big|,\\ =&\big|\|z-\bar{z}\|^2 - \|z-\bar{z}'_j\|^2\big|,\\ =&\big|\langle 2z-\bar{z}-\bar{z}'_j,\ \underbrace{\bar{z}-\bar{z}'_j}_{=(z_j-z'_j)/n}\rangle\big|, \quad \text{by Cauchy-Schwartz} \downarrow\\ \le& \tfrac{1}{n}\underbrace{\|2z-\bar{z}-\bar{z}'_j\|}_{\le 4\kappa}\underbrace{\|z_j-z'_j\|}_{\le 2\kappa} \le \frac{8\kappa^2}{n} := \beta.\end{aligned}$$

- We observe that larger the sample size n, smaller the parameter β; thus, more stable the algorithm. Another observation we can make is that the radius of the support κ has a negative effect on the stability of an algorithm.

Example. [Stability of Lipschitz loss & linear functions]

- We assume that the loss is Lipschitz in its second argument, i.e.

$$\big|\ell(z,f)-\ell(z,f')\big| \le L\|f-f'\|_\infty \triangleq L\sup_{x\in\mathbb{R}^d}\big|f(x)-f'(x)\big|.$$

  If we consider an SVM classifier where $y \in \{\pm 1\}$ and the loss is Hinge loss $\ell(z=(y,x),f) = \max\{0, 1-yf(x)\}$, we have

$$\begin{aligned}\big|\ell(z,f)-\ell(z,f')\big| =& \big|\max\{0,1-yf(x)\} - \max\{0,1-yf'(x)\}\big|\\ \le& \big|yf(x)-yf'(x)\big| \le \sup_{x\in\mathbb{R}^d}\big|f(x)-f'(x)\big|.\end{aligned}$$

- Now let's focus our attention to the class of linear functions $\mathcal{F} = \{x \to \langle x, \theta\rangle : \theta \in \mathbb{R}^d\}$. Any function $f \in \mathcal{F}$ can be characterized by the parameter θ; so let's switch notation $f \to \theta$.

- SVMs are generally coupled with ℓ_2-regularization; thus the resulting empirical risk minimization algorithm reduces to

$$\hat{\theta}_{\mathcal{D}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \max\{0, 1-y_i\langle\theta, x_i\rangle\} + \frac{\lambda}{2}\|\theta\|^2$$

- Therefore the resulting loss function becomes

$$\ell(\underbrace{z}_{=(y,x)}, f) = \max\{0, 1-y\underbrace{\langle\theta,x\rangle}_{=f(x)}\} + \frac{\lambda}{2n}\|\theta\|^2.$$

- If we assume $\|x_i\| \le \kappa$, Bousquet and Elisseeff showed that this algorithm has uniform stability with parameter

$$\beta = \frac{\kappa^2}{\lambda n}.$$

  This is nontrivial, and skipped in class. Similar to the Gaussian mean estimation example, stability gets better with the number of samples. But another important observation we can make is that stability gets better with more regularization.

The following result provides a generalization bound based on uniform β-stability.

Theorem 42 (Generalization based on Uniform Stability)**.** *Assume that an empirical risk minimization algorithm is uniformly β-stable, and the loss is bounded, i.e., $0 \leq \ell(z, f) \leq B$. Then with probability at least $1 - \delta$, we have*

$$R(\hat{f}) - R(f_*) \leq \beta + (\beta n + 3B)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Remark. We make the following remarks.

- Notice that for above bound to be useful, one needs $\beta = o(1/\sqrt{n})$. This is because of the term $\beta\sqrt{n}$ in the coefficient of the second term on the right hand side.
- In general, we have $\beta = \mathcal{O}(1/n)$ which gives the familiar rate of generalization error, $\mathcal{O}(1/\sqrt{n})$.
- In the case of linear SVMs (previous example), we have $\beta = \frac{\kappa^2}{\lambda n}$. This yields a bound of order

$$\mathcal{O}\left(\frac{\kappa^2}{\lambda n} + (\kappa^2 + B)\sqrt{\frac{2\log(1/\delta)}{n}}\right) = \mathcal{O}\left(\frac{(\kappa^2 + B)\sqrt{\log(1/\delta)}}{\lambda\sqrt{n}}\right).$$

  This bound is the same order as previous generalization bounds we obtained, but it is worse in terms of dependence on κ.

Proof. The proof of this theorem is very similar to that of Theorem 23, the generalization results based on Rademacher complexity. Recall the notation $\hat{R}(f; \mathcal{D})$ which means the empirical risk of f over the dataset $\mathcal{D}$. For example, $\hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}_j)$ is the empirical risk of $\hat{f}_{\mathcal{D}}$ over the single-data perturbed dataset $\mathcal{D}_j$.

The main observation is again to write the following decomposition of the excess risk

$$R(\hat{f}_{\mathcal{D}}) - R(f_*) = \underbrace{[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D})]}_{\text{not iid sum}} + \underbrace{[\hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \hat{R}(f_*; \mathcal{D})]}_{\leq 0} + \underbrace{[\hat{R}(f_*; \mathcal{D}) - R(f_*)]}_{\text{iid sum}/n}, \quad (10.2)$$

$$\leq 2\sup_{f \in \mathcal{F}} |\hat{R}(f; \mathcal{D}) - R(f)| \quad \text{which is what we did previously.}$$

Before, we proceeded by bounding both of the above nontrivial terms with the supremum of the empirical process, $\sup_{f \in \mathcal{F}} |\hat{R}(f; \mathcal{D}) - R(f)|$. This time though, we will handle them separately. Bounding the second term above is quite easy since f_* is deterministic, and therefore it becomes an iid average, i.e.,

$$\hat{R}(f_*; \mathcal{D}) - R(f_*) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, f_*) - \mathbb{E}[\ell(z_i, f_*)],$$

which we know how to deal with.

For the first term $R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D})$, we will invoke the uniform stability together with McDiarmid's inequality.

The proof relies on three key steps as before: 1-Concentration, 2-Control over expectation, and 3- Uniform conv. (10.2) $\implies$ generalization.

1. **Concentration**: Let's recall the main concentration tool that we will relied on in our efforts to derive a generalization bound based on Rademacher complexity.

Lemma 43 (Recall: McDiarmid's inequality (Lemma 24))**.** *Let $g : \mathcal{Z} \times ... \times \mathcal{Z} \to \mathbb{R}$ be a function satisfying the bounded difference property*

$$|g(z_1, \dots, z_j, \dots, z_n)| - g(z_1, \dots, z_j', \dots, z_n)| \leq c_j$$

Then for independent random variables $z_1, z_2 \dots, z_n$, we have

$$\mathbb{P}\Big(g(z_1, \dots, z_n) - \mathbb{E}[g(z_1, \dots, z_n)] \geq \epsilon\Big) \leq \exp\left\{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right\}.$$

Recall that Hoeffding's inequality is an application of the above lemma. We can invoke either and immediately obtain a bound on the second term. Let's get that out of the way.

Warm-up: Getting the third term in (10.2) out of way. By McDiarmid's (or by Hoeffding's) inequality, we have

$$\mathbb{P}\Big(\hat{R}(f_*; \mathcal{D}) - R(f_*) \geq \frac{\epsilon}{2}\Big) \leq \exp\Big\{-\frac{n\epsilon^2}{2B^2}\Big\} \triangleq \frac{\delta}{2}.$$

This translates to, with probability at least $1 - \delta/2$, we have

$$\hat{R}(f_*; \mathcal{D}) - R(f_*) \leq B\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Bounding the first term in (10.2). Recall that previously, we needed to bound the empirical process in (10.2). For this, we'd let the g function from McDiarmid's inequality be the function of interest. That is,

$$\text{Previously: } g(z_1, \dots, z_n) = \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f).$$

This time though, we are dealing with another function, so we let

$$\text{This time: } g(z_1, \dots, z_n) = R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}).$$

Notice that, by the uniform β-stability assumption, we have

$$\Big|\hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D})\Big| \leq \beta \text{ and } \Big|R(\hat{f}_{\mathcal{D}}) - R(\hat{f}_{\mathcal{D}_j})\Big| \leq \beta.$$

Let's verify the second one as the first one follows from the same argument.

$$\begin{aligned}\Big|R(\hat{f}_{\mathcal{D}}) - R(\hat{f}_{\mathcal{D}_j})\Big| =& \Big|\mathbb{E}[\ell(z, \hat{f}_{\mathcal{D}}) - \ell(z, \hat{f}_{\mathcal{D}_j})]\Big| \\ \leq& \mathbb{E}[|\ell(z, \hat{f}_{\mathcal{D}}) - \ell(z, \hat{f}_{\mathcal{D}_j})|] \text{ by triangle ineq.} \\ \leq& \beta \text{ by uniform } \beta\text{-stability.}\end{aligned}$$

We proceed by first verifying the bounded difference property which is needed by McDiarmid's inequality.

$$
\begin{aligned}
&|g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z_j', \ldots, z_n)| \\
&= \Big| R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \big[R(\hat{f}_{\mathcal{D}_j}) - \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D}_j) \big] \Big| \\
&\leq \underbrace{\Big| R(\hat{f}_{\mathcal{D}}) - R(\hat{f}_{\mathcal{D}_j}) \Big|}_{\leq \beta \text{ by stability}} + \Big| \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D}_j) \pm \hat{R}(\hat{f}_{\mathcal{D}_j}, \mathcal{D}) \Big| \quad \text{by triangle ineq.} \\
&\leq \beta + \underbrace{\Big| \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D}) \Big|}_{\leq \beta \text{ by stability}} + \underbrace{\Big| \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D}_j) \Big|}_{= \frac{1}{n} |\ell(z_j, \hat{f}) - \ell(z_j', \hat{f})| \leq \frac{2B}{n}} \quad \text{by triangle ineq.} \\
&\leq 2\beta + \frac{2B}{n} \triangleq c_j \quad \text{in McDiarmid's inequality.}
\end{aligned}
$$

Hence, by the McDiarmid's inequality, we obtain

$$
\begin{aligned}
\mathbb{P}\Bigg(R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) \geq \epsilon + \overbrace{\mathbb{E}\Big[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) \Big]}^{\text{Need to control}} \Bigg) &\leq \exp\left\{ \frac{-2\epsilon^2}{n(2\beta + 2B/n)^2} \right\} \qquad (10.3) \\
&\leq \exp\left\{ \frac{-n\epsilon^2}{2(\beta n + B)^2} \right\} \triangleq \frac{\delta}{2}.
\end{aligned}
$$

The above bound is obtained under uniform stability; yet, it is not surprising at all given the McDiarmid's inequality. We still need to control the additional expectation above. This was previously done by the symmetrization argument. In the following we use stability property of the algorithm.

2. **Controlling the expectation via stability**: We denote our dataset with $\mathcal{D} = \{z_1, \ldots, z_n\}$, and let $\mathcal{D}' = \{z_1', \ldots, z_n'\}$ be the iid copy of the $\mathcal{D}$, and the perturbation is given as $\mathcal{D}_j = \{z_1, .., z_j', .., z_n\}$. We have

$$
\hat{R}(f; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) \quad \text{and} \quad \hat{R}(f; \mathcal{D}_j') = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i', f).
$$

For a fixed f, the population risk will be identical for these datasets,

$$
R(f) = \mathbb{E}[\ell(z, f)] = \mathbb{E}[\hat{R}(f, \mathcal{D})] = \mathbb{E}[\hat{R}(f, \mathcal{D}_j')].
$$

Let's investigate the quantity we would like to bound.

$$
\begin{aligned}
\mathbb{E}\Big[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) \Big] &= \mathbb{E}_{\text{all}}\Big[\mathbb{E}_z[\ell(z, \hat{f}_{\mathcal{D}})] - \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, \hat{f}_{\mathcal{D}}) \Big] \\
&= \mathbb{E}\Big[\mathbb{E}_{z_i'}\Big[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i', \hat{f}_{\mathcal{D}}) \Big] - \frac{1}{n} \sum_{i=1}^{n} \ell(z_i', \hat{f}_{\mathcal{D}_i}) \Big] \\
&= \mathbb{E}\Big[\mathbb{E}_{z_i'}\Big[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i', \hat{f}_{\mathcal{D}}) - \ell(z_i', \hat{f}_{\mathcal{D}_i}) \Big] \Big] \\
&\leq \beta \quad \text{by stability.}
\end{aligned}
$$

The second inequality is because z_i' is independent from $\mathcal{D}$, and $\mathcal{D}$ and $\mathcal{D}_j'$ are exchangable. Therefore we get, with probability at least $1-\delta/2$

$$R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}) \le \mathbb{E}\Big[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D})\Big] + \frac{\epsilon}{2} \le \beta + (\beta n + B)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

3. **Uniform convergence $\implies$ generalization** (but almost): Combining this with (10.3), we write out generalization bound with probability at least $1-\delta$,

$$\begin{aligned} R(\hat{f}_{\mathcal{D}}) - R(f_*) &\le [R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D})] + 0 + [\hat{R}(f_*;\mathcal{D}) - R(f_*)] \\ &\le \beta + (\beta n + B)\sqrt{\frac{\log(2/\delta)}{2n}} + B\sqrt{\frac{2\log(2/\delta)}{n}} \\ &\le \beta + (\beta n + 3B)\sqrt{\frac{\log(2/\delta)}{2n}} \end{aligned}$$

which concludes the proof.

□

10.2 PAC-Bayes bounds

In this section, we scratch the surface of PAC-Bayesian bounds. The PAC-Bayes theory is originally developed as an attempt to explain Bayesian learning from a learning theory perspective. But these tools have to be proved very useful in various context. The main idea is to place a prior distribution π_0 over the function class $\mathcal{F}$, which encodes our prior knowledge over the set of hypotheses. After observing data $\mathcal{D}$, we update our view of the function class, which is referred to as the posterior distribution $\pi_{\mathcal{D}}$.

The bounds that rely on the concept "uniform convergence $\implies$ generalization" hold for all functions in the function class. Consider for example a finite function class. By a simple application of the union bound, we were able to derive a generalization error bound of (ignoring constants)

$$R(\hat{f}) - R(f_*) < \sqrt{\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{n}},$$

which we proved as Theorem 18. However, the main building block of this theorem was to show the uniform convergence, which reads (again ignoring constants), with probability at least $1-\delta$

$$\forall f \in \mathcal{F}, \ : \ R(f) \le \hat{R}(f) + \sqrt{\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{n}}. \tag{10.4}$$

This is equivalent to saying $\sup_{f\in\mathcal{F}} R(f) - \hat{R}(f) \le$ the last term above. However, we notice that the above bound gives a worst case bound, in other words it gives a bound for all functions by treating them all the same. But we know some are more likely than the others!

If we had a prior distribution $\pi_0(f)$ over the class of functions $\mathcal{F}$ that are available to us, we can incorporate this to our bound. Intuitively, if there is a function $f \in \mathcal{F}$ that we are certain it is not going to be returned by our algorithm, it shouldn't count towards the size of the function class which appears in the numerator of (10.4).

Let's start with the simplest of PAC-Bayes style bounds, Occam's bound.

Theorem 44 (Occam's bound)**.** *For a countable function class $\mathcal{F}$, and a bounded loss function $0 \leq \ell \leq B$, if we have the prior distribution π_0 over the function class $\mathcal{F}$, then with probability at least $1-\delta$, we have*

$$\forall f \in \mathcal{F} \;:\; R(f) \leq \hat{R}(f) + B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta)}{2n}}. \tag{10.5}$$

Remark. We make the following immediate remarks.

1. The bound is not for the excess risk. The difference between training and the test error is small for a function f, if its prior is large.

2. If the prior distribution $\pi_0(f)$ is uniform over $\mathcal{F}$, i.e. each function is equally likely and $\pi_0(f) = \mathbb{P}(f = f_i) = 1/|\mathcal{F}|$, the above bound reduces to the bound in (10.4).

3. If the prior distribution is uniform over a subset $\mathcal{G}$ of $\mathcal{F}$, bound reduces to $\sqrt{\frac{\log(|\mathcal{G}|)+\log(1/\delta)}{2n}}$. This was exactly our intuition; the functions that are unlikely to come up shouldn't count towards the complexity of the function class.

4. If the prior puts all its mass on a single function f_0, i.e. $\pi_0(f_0) = 1$, then the bound reduces to just a concentration result, since we only have a single function that is available to us.

5. This bound allows $\mathcal{F}$ to have large size as long as the prior behaves nicely for a specific function $f \in \mathcal{F}$. For that particular function, above result will yield a good bound. However, if the prior is somewhat close to uniform distribution, then $\pi_0(f) \approx 1/|\mathcal{F}|$ will get worse with an increase in the size of the function class.

Proof. The main idea in this proof is to simply allocate the confidence parameter δ over different functions based on their prior.

For a fixed (non-random) function $f \in \mathcal{F}$, by the Hoeffding's inequality, we have

$$\mathbb{P}\Big(R(f) \geq \hat{R}(f) + \epsilon\Big) \leq \exp\Big\{-\frac{2n\epsilon^2}{B^2}\Big\} := \delta_f = \pi_0(f)\delta.$$

Notice that $\sum_f \delta_f = \delta$ since π_0 is a probability distribution. The above bound reads,

$$\mathbb{P}\left(R(f) \geq \hat{R}(f) + B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta_f)}{2n}}\right) \leq \delta_f.$$

Note that the above bound holds for a fixed f. By applying the union bound over $f \in \mathcal{F}$, we obtain

$$\mathbb{P}\left(\forall f \in \mathcal{F} : R(f) \geq \hat{R}(f) + B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta_f)}{2n}}\right) \leq \sum_{f\in\mathcal{F}} \delta_f = \delta,$$

which completes the proof.

□

Let's recall our objective: We want to minimize the population risk (aka test error). The bound (10.6) upper bounds the quantity we would like to minimize. Therefore, we can minimize this upper bound, and hope that we get close to minimizing itself! That is, the above theorem suggest to minimize the following objective

$$\hat{R}(f) + \underbrace{B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta)}{2n}}}_{\text{regularizer}}. \tag{10.6}$$

The bound $B\sqrt{\frac{\log(1/\pi_0(f))+\log(1/\delta_f)}{2n}}$ will serve as a regularizer by penalizing functions that are less likely according to the prior π_0. We also observe that its impact decreases with increased sample size n.

There are two shortcomings of the Occam's bound.

- First, it relies on the union bound which requires the function class $\mathcal{F}$ to be countable.
- Second, it only allows an algorithm to return a single function rather than a posterior distribution. These are addressed in the following theorem.

Theorem 45 (McAllester's PAC-Bayes theorem)**.** *For any prior π_0 and any posterior $\pi_{\mathcal{D}}$, and a bounded loss function $0 \leq \ell(z, f) \leq 1$, with probability at least $1 - \delta$, we have*

$$\mathbb{E}_{f\sim\pi_{\mathcal{D}}}[R(f)] \leq \mathbb{E}_{f\sim\pi_{\mathcal{D}}}[\hat{R}(f)] + \sqrt{\frac{KL(\pi_{\mathcal{D}}||\pi_0) + \log(4n/\delta)}{2n-1}}.$$

Remark.

- Compared to Occam's bound, instead of for all f, this one is for expectation under the posterior.
- If the posterior puts all its mass on one function f_0 in $\mathcal{F}$, the above bound recovers Occam's bound. Say for example, π_0 is uniform over a finite set $\mathcal{F}$. Then,

$$\begin{aligned}\mathrm{KL}(\pi_{\mathcal{D}}||\pi_0) &= \sum_f \log\left(\frac{\pi_{\mathcal{D}}(f)}{\pi_0(f)}\right)\pi_{\mathcal{D}}(f)\\ &= \log\left(\frac{\pi_{\mathcal{D}}(f_0)}{\pi_0(f_0)}\right)\pi_{\mathcal{D}}(f_0) = \log(|\mathcal{F}|).\end{aligned}$$

- Converting the above bound to a (kind of) bound on the excess risk requires characterizing the expected suboptimality,

$$\mathbb{E}_{f\sim\pi_{\mathcal{D}}}[R(f)] - R(f_*).$$

- In literature, the expectations are generally denoted with $\mathbb{E}_{f\sim\pi}[R(f)] = R(\pi)$.

Proof. Skipped in class. To be added. □