

Lecture 8: PAC-Bayes Bounds

Han Shao

Chp 31, UML

1 From a Hypothesis Class to a Distribution over Hypotheses

Our earlier generalization bounds hold simultaneously for every $h \in \mathcal{H}$. For example, when $\mathcal{H}$ is finite, Hoeffding's inequality and a union bound give

$$\text{err}_{\mathcal{D}}(h) \leq \text{err}_S(h) + \sqrt{\frac{\ln |\mathcal{H}| + \ln(1/\delta)}{2m}} \quad \text{for all } h \in \mathcal{H}$$

with probability at least $1 - \delta$. Rademacher complexity can replace the cardinality term by a quantity that reflects the observed instances. However, within a fixed class and sample, that complexity term is still the same for every hypothesis.

PAC-Bayes gives a different way to measure complexity. We first fix a distribution P over hypotheses, expressing which hypotheses we favor before observing the sample. The learner then selects a distribution Q after observing the sample. The bound depends on how much Q differs from P .

We retain the notation from the previous lectures: $S = (z_1, \dots, z_m) \sim \mathcal{D}^m$, with $z_i = (x_i, y_i)$ in the classification setting. We write $\ell_h(z)$ for the loss, $\text{err}_{\mathcal{D}}(h)$ for its population expectation, and $\text{err}_S(h)$ for its empirical average. For binary classification, $\ell_h(x, y) = \mathbf{1}\{h(x) \neq y\}$ with labels in $\{-1, +1\}$. The argument below works for any measurable loss $\ell_h(z) \in [0, 1]$; we use the same err notation for its expected and empirical values. All logarithms are natural.

A *prior* P is a distribution over $\mathcal{H}$ fixed independently of S . A *posterior* Q is a distribution over $\mathcal{H}$ that may depend arbitrarily on S . Here “posterior” does not require a Bayesian update rule. Any data-dependent choice of Q is allowed. The data distribution $\mathcal{D}$ is a distribution over examples; P and Q are distributions over hypotheses.

A distribution Q defines a randomized predictor: on an input x , draw $h \sim Q$ and predict $h(x)$. Its loss averages over this internal randomness:

$$\ell_Q(z) = \mathbb{E}_{h \sim Q}[\ell_h(z)].$$

Consequently,

$$\text{err}_{\mathcal{D}}(Q) = \mathbb{E}_{h \sim Q}[\text{err}_{\mathcal{D}}(h)], \quad \text{err}_S(Q) = \mathbb{E}_{h \sim Q}[\text{err}_S(h)].$$

These quantities concern the randomized predictor. They are not, in general, the error of a majority vote or of a single hypothesis drawn from Q .

Definition 1 (Kullback–Leibler divergence). For a finite or countable hypothesis class, define

$$\text{KL}(Q \| P) = \sum_{h: Q(h) > 0} Q(h) \ln \frac{Q(h)}{P(h)} = \mathbb{E}_{h \sim Q} \left[\ln \frac{Q(h)}{P(h)} \right].$$

If $Q(h) > 0$ but $P(h) = 0$ for some h , the divergence is $+\infty$. For distributions with densities, replace the sum by an integral and use the density ratio. More generally, $\text{KL}(Q\|P) = \mathbb{E}_Q[\ln(dQ/dP)]$ when $Q \ll P$, and it is $+\infty$ otherwise.

KL divergence is nonnegative and vanishes when $Q = P$. It is generally not symmetric, so it is not a metric. In this lecture it measures the complexity of choosing Q relative to the fixed reference P .

2 The PAC-Bayes Generalization Bound

Theorem 1 (PAC-Bayes bound). *Let $\mathcal{H}$ be a hypothesis class, let $\ell_h(z) \in [0, 1]$, and fix a prior P independently of the training sample. For every $m \geq 2$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, simultaneously for every distribution Q over $\mathcal{H}$,*

$$\text{err}_{\mathcal{D}}(Q) \leq \text{err}_S(Q) + \sqrt{\frac{\text{KL}(Q\|P) + \ln((2m-1)/\delta)}{2(m-1)}}. \quad (1)$$

In particular, Q may be chosen after observing S .

The bound balances two quantities:

- $\text{err}_S(Q)$ rewards distributions that place weight on hypotheses with small training error.
- $\text{KL}(Q\|P)$ penalizes departing from the prior, especially by concentrating on hypotheses that had little prior probability.

The prior cannot be chosen using this same training sample under the theorem's assumptions. It may, for example, be constructed from independent data and then treated as fixed when applying the theorem.

3 Proof of Theorem 1

The proof has three ingredients. First, an exponential moment under P controls expectations under every Q , at a cost of $\text{KL}(Q\|P)$. Second, Hoeffding's inequality bounds this exponential moment for each fixed hypothesis. Finally, Markov's inequality gives one event on which all posteriors are controlled.

Changing the distribution over hypotheses

Lemma 1 (Change of measure). *For any bounded measurable function $g : \mathcal{H} \rightarrow \mathbb{R}$ and any distributions P, Q with $\text{KL}(Q\|P) < \infty$,*

$$\mathbb{E}_{h \sim Q}[g(h)] \leq \text{KL}(Q\|P) + \ln \mathbb{E}_{h \sim P}[e^{g(h)}]. \quad (2)$$

Proof. For a finite or countable class, expand the KL divergence and apply Jensen's inequality to the concave logarithm:

$$\begin{aligned}
\mathbb{E}_Q[g(h)] - \text{KL}(Q\|P) &= \mathbb{E}_Q\left[\ln\left(e^{g(h)}\frac{P(h)}{Q(h)}\right)\right] \\
&\leq \ln \mathbb{E}_Q\left[e^{g(h)}\frac{P(h)}{Q(h)}\right] \\
&= \ln \sum_{h:Q(h)>0} P(h)e^{g(h)} \\
&\leq \ln \sum_{h\in\mathcal{H}} P(h)e^{g(h)} = \ln \mathbb{E}_P[e^{g(h)}].
\end{aligned}$$

The last step is an inequality because Q may assign zero mass to hypotheses with positive prior mass. For a general hypothesis space, the same proof uses $r = dQ/dP$ and

$$\mathbb{E}_Q[e^g/r] = \int_{\{r>0\}} e^g dP \leq \mathbb{E}_P[e^g]. \quad \square$$

This lemma is deterministic. We can therefore fix a sample S , choose a function g depending on S , and apply it to every Q , including posteriors selected using that sample.

From a tail bound to an exponential moment For each fixed h , define the generalization gap

$$\Delta_S(h) = \text{err}_{\mathcal{D}}(h) - \text{err}_S(h).$$

This quantity may be positive or negative. We use the two-sided Hoeffding inequality to control its square.

Lemma 2 (Exponential moment of the gap). *For every fixed $h \in \mathcal{H}$ and every $m \geq 2$,*

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left[e^{2(m-1)(\Delta_S(h))^2} \right] \leq 2m - 1. \quad (3)$$

Proof. Since the m losses are independent and lie in $[0, 1]$, Hoeffding's inequality gives, for every $t > 0$,

$$\Pr_S[|\Delta_S(h)| > t] \leq 2e^{-2mt^2}.$$

Let $W = e^{2(m-1)(\Delta_S(h))^2} \geq 1$. For every $u > 1$,

$$\begin{aligned}
\Pr_S[W > u] &= \Pr_S\left[|\Delta_S(h)| > \sqrt{\frac{\ln u}{2(m-1)}}\right] \\
&\leq 2 \exp\left(-\frac{m}{m-1} \ln u\right) = 2u^{-m/(m-1)}.
\end{aligned}$$

For any nonnegative random variable, $\mathbb{E}[W] = \int_0^\infty \Pr[W > u] du$. Thus

$$\begin{aligned}
\mathbb{E}_S[W] &= 1 + \int_1^\infty \Pr_S[W > u] du \\
&\leq 1 + 2 \int_1^\infty u^{-m/(m-1)} du \\
&= 1 + 2(m-1) = 2m - 1. \quad \square
\end{aligned}$$

The coefficient $2(m-1)$ is slightly smaller than the $2m$ in Hoeffding's exponent. This makes the last integral converge and explains the $m-1$ in the theorem's denominator.

One event that controls all posteriors

Proof of Theorem 1. Set

$$g_S(h) = 2(m-1)(\Delta_S(h))^2, \quad Z(S) = \mathbb{E}_{h \sim P}[e^{g_S(h)}].$$

Since P is fixed independently of S , we may average the fixed-hypothesis moment bound over P :

$$\begin{aligned} \mathbb{E}_S[Z(S)] &= \mathbb{E}_{h \sim P} \mathbb{E}_S[e^{g_S(h)}] \\ &\leq \mathbb{E}_{h \sim P}[2m-1] = 2m-1. \end{aligned}$$

Here exchanging the expectations is justified by nonnegativity; independence ensures that the inner sample distribution remains $\mathcal{D}^m$ for every fixed h .

By Markov's inequality,

$$\Pr_S[Z(S) > (2m-1)/\delta] \leq \delta.$$

Fix any sample in the event $Z(S) \leq (2m-1)/\delta$, which has probability at least $1 - \delta$. For *every* Q with finite KL divergence, Lemma 1 gives

$$\begin{aligned} 2(m-1)\mathbb{E}_{h \sim Q}[(\Delta_S(h))^2] &\leq \text{KL}(Q\|P) + \ln Z(S) \\ &\leq \text{KL}(Q\|P) + \ln((2m-1)/\delta). \end{aligned}$$

Finally, Jensen's inequality gives

$$\begin{aligned} |\text{err}_{\mathcal{D}}(Q) - \text{err}_S(Q)| &= |\mathbb{E}_{h \sim Q}[\Delta_S(h)]| \\ &\leq \sqrt{\mathbb{E}_{h \sim Q}[(\Delta_S(h))^2]} \\ &\leq \sqrt{\frac{\text{KL}(Q\|P) + \ln((2m-1)/\delta)}{2(m-1)}}. \end{aligned}$$

If $\text{KL}(Q\|P) = +\infty$, the claimed bound is vacuous. Therefore the same event establishes the theorem for every Q . $\square$

Remark 1 (A two-sided guarantee). The proof already bounds the absolute generalization gap. Thus, on the same event, simultaneously for every Q ,

$$|\text{err}_{\mathcal{D}}(Q) - \text{err}_S(Q)| \leq \sqrt{\frac{\text{KL}(Q\|P) + \ln((2m-1)/\delta)}{2(m-1)}}$$

with probability at least $1 - \delta$. No additional union bound is needed.

4 Examples and Interpretation

4.1 Recovering a finite-class bound

Suppose $|\mathcal{H}| = N$ and P is uniform over $\mathcal{H}$. Let Q put all its mass on a single hypothesis h . Then

$$\text{KL}(Q\|P) = \ln \frac{1}{P(h)} = \ln N.$$

Theorem 1 gives, simultaneously for all $h \in \mathcal{H}$,

$$\text{err}_{\mathcal{D}}(h) \leq \text{err}_S(h) + \sqrt{\frac{\ln N + \ln((2m-1)/\delta)}{2(m-1)}}.$$

This recovers the dependence on $\ln N$ from the finite-class bound, with an extra $\ln(2m-1)$ term and $m-1$ in place of m . This slack comes from the particular version of the PAC-Bayes bound proved here.

More generally, for any fixed discrete prior and any h with $P(h) > 0$, the same point-mass choice gives

$$\text{err}_{\mathcal{D}}(h) \leq \text{err}_S(h) + \sqrt{\frac{\ln(1/P(h)) + \ln((2m-1)/\delta)}{2(m-1)}}.$$

A hypothesis with larger prior probability receives a smaller complexity penalty. This is the connection with Occam bounds.

4.2 Why a distribution can have smaller complexity

Continue with a uniform prior on N hypotheses. Suppose the learner finds a subset $A \subseteq \mathcal{H}$ of size k and chooses Q uniformly on A . The subset may depend on S . Then

$$\text{KL}(Q\|P) = \sum_{h \in A} \frac{1}{k} \ln \frac{1/k}{1/N} = \ln \frac{N}{k}.$$

Thus

$$\frac{1}{k} \sum_{h \in A} \text{err}_{\mathcal{D}}(h) \leq \frac{1}{k} \sum_{h \in A} \text{err}_S(h) + \sqrt{\frac{\ln(N/k) + \ln((2m-1)/\delta)}{2(m-1)}}.$$

If all k hypotheses have small training error, spreading the posterior over them keeps the empirical error small while reducing the KL term. For instance, if $N = 1024$ and $k = 256$, this term is $\ln 4$, compared with $\ln 1024$ for a point mass. The conclusion controls their average population error; it does not certify each member of A individually.

4.3 Gaussian perturbations and a norm penalty

For a continuous example, index predictors by parameters $w \in \mathbb{R}^d$, treating the parameter space as the hypothesis space. Fix $\tau > 0$ before observing S , and consider

$$P = \mathcal{N}(0, \tau^2 I_d), \quad Q = \mathcal{N}(w, \tau^2 I_d),$$

where the center w may depend on S . For a parameter v ,

$$\ln \frac{q(v)}{p(v)} = \frac{\|v\|_2^2 - \|v - w\|_2^2}{2\tau^2}.$$

Writing $v = w + \xi$ with $\xi \sim \mathcal{N}(0, \tau^2 I_d)$ and averaging eliminates the cross term, so

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

Hence, with probability at least $1 - \delta$, simultaneously for all centers w ,

$$\mathbb{E}_\xi[\text{err}_\mathcal{D}(h_{w+\xi})] \leq \mathbb{E}_\xi[\text{err}_S(h_{w+\xi})] + \sqrt{\frac{\|w\|_2^2/(2\tau^2) + \ln((2m-1)/\delta)}{2(m-1)}}.$$

This is a guarantee for the predictor with Gaussian parameter noise. Its empirical term measures how well the predictor performs under those perturbations. The formula contains no explicit factor of d , but the perturbed empirical error and the norm may depend strongly on dimension.

The theorem does not immediately give the same guarantee for the unperturbed predictor h_w : a point mass at w has infinite KL relative to this continuous prior. Also, τ is fixed here because changing it changes the prior.

4.4 A learning objective suggested by the bound

For a fixed prior, the bound suggests selecting Q to minimize

$$\text{err}_S(Q) + \sqrt{\frac{\text{KL}(Q\|P) + \ln((2m-1)/\delta)}{2(m-1)}}.$$

One can optimize over a convenient family of posteriors; the theorem still applies to the resulting choice, whether or not a global minimum is found. The examples show two ways to keep the KL term small: retain substantial prior mass among good hypotheses, or use a Gaussian posterior whose center has small norm. In both cases, the goal is to achieve small empirical error while limiting how much the posterior departs from the prior.