# 泛化理论

## 第四章 PAC-Bayesian

## §4.1.2 PAC-Bayesian Bound (proof)

@ 滕佳烨

[ref] McAllester, D. A. (1999, July). PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory* (pp. 164-170).

**Recall:**

**Theorem (PAC-Bayesian).** Given prior distribution $P$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training), for all posterior distribution $Q$,

$$L_D(Q) - L_S(Q) \le \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

**Today's topic:**
Its proof.

Key idea: change the distribution from Q to P.

**Theorem (PAC-Bayesian).** Given prior distribution $P$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training), for all posterior distribution $Q$,

$$L_D(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

**The core of the proof is still "decouple".**

Denote generalization gap for h by $\Delta(h) = L_D(\mathrm{h}) - L_S(h)$. We want to bound $\mathbb{E}_Q \Delta(h)$.

However, we can only bound $\mathbb{E}_P \Delta(h)$ since $P$ is independent of the training.

How to transfer from the distribution P to Q? KL→ similar idea in optimal transport.

$$\mathbb{E}_P f(x) = \int f(x) p(x) dx = \int f(x) \frac{p(x)}{q(x)} q(x) dx = \mathbb{E}_Q f(x) \frac{p(x)}{q(x)}.$$

Not good? Recall $KL(Q||P) = \mathbb{E}_Q \log \frac{q(x)}{p(x)}$. Where is the log? Jensen's inequality!

$$\log \mathbb{E}_P f(x) = \log \mathbb{E}_Q f(x) \frac{p(x)}{q(x)} \geq \mathbb{E}_Q \log f(x) \frac{p(x)}{q(x)} = \mathbb{E}_Q \log f(x) - KL(Q||P).$$

**Theorem (PAC-Bayesian).** Given prior distribution $P$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training), for all posterior distribution $Q$,

$$L_D(Q) - L_S(Q) \le \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

**Goal: bound the term $\mathbb{E}_Q \Delta(h)$, where $\Delta(h) = L_D(\mathrm{h}) - L_S(h)$.**

$$\log \mathbb{E}_P f(x) = \log \mathbb{E}_Q f(x) \frac{p(x)}{q(x)} \ge \mathbb{E}_Q \log f(x) \frac{p(x)}{q(x)} = \mathbb{E}_Q \log f(x) - KL(Q||P).$$

Setting $\log f(x)$ as $\mathrm{c}\Delta(h)^2$, where $c$ is a constant to be determined,

$$\log \mathbb{E}_P \exp\left(c\Delta(h)^2\right) \ge c\mathbb{E}_Q \Delta(h)^2 - KL(Q||P).$$

We next consider the bound for $\mathbb{E}_P \exp\left(c\Delta(h)^2\right)$, which is easier since $P$ is independent.

However, note that we need the bound hold for *any* distribution Q, therefore, we need sup:

$$\sup_Q \mathrm{c}\Delta(h)^2 - KL(Q||P) \le ?$$

**Theorem (PAC-Bayesian).** Given prior distribution $P$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training), for all posterior distribution $Q$,

$$L_D(Q) - L_S(Q) \le \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

**Goal: bound the term $\mathbb{E}_Q \Delta(h)$, where $\Delta(h) = L_D(\mathrm{h}) - L_S(h)$.**

$$\log \mathbb{E}_P \exp\left(c\Delta(h)^2\right) \ge c\mathbb{E}_Q \Delta(h)^2 - KL(Q||P).$$

- Let $\mathrm{f(S)} = \sup_Q \mathrm{c}\mathbb{E}_Q \Delta(h)^2 - KL(Q||P)$, then $\mathbb{E}_S \exp\left(f(S)\right) \le \mathbb{E}_S \mathbb{E}_P \exp\left(c\Delta(h)^2\right)$

where the RHS is independent of $Q$ (so we can take sup).

- Due to Hoeffding inequality, with prob (over $P$) at most $\exp(-2\mathrm{n}t^2)$, $\Delta(h) \ge t$.
- Plug it into the above equation, we can derive that $\mathbb{E}_S \exp\left(f(S)\right) \le \frac{c}{2n-c}$ (if $c < 2n$).

where we use $\mathbb{E}X \le \int P(X \ge t)dt$ (note that $t > 1$ when $X = \exp\left(c\Delta(h)^2\right)$).

- Therefore, by choosing $c = 2(n-1)$, we have that $\mathbb{E}_S \exp\left(f(S)\right) \le n$.

**Theorem (PAC-Bayesian).** Given prior distribution $P$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training), for all posterior distribution $Q$,

$$L_D(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

**Goal: bound the term $\mathbb{E}_Q \Delta(h)$, where $\Delta(h) = L_D(\mathrm{h}) - L_S(h)$.**

$$\log \mathbb{E}_P \exp\left(c\Delta(h)^2\right) \geq c\mathbb{E}_Q \Delta(h)^2 - KL(Q||P).$$

$$\mathrm{f(S)} = \sup_Q 2(\mathrm{n}-1)\mathbb{E}_Q \Delta(h)^2 - KL(Q||P), \mathbb{E}_S \exp\left(f(S)\right) \leq n.$$

Therefore, by Markov inequality, $\mathbb{P}(f(S) \geq u) \leq \frac{n}{\exp u}$. By setting $u = \log\frac{n}{\delta}$, we have for any $Q$, with probability at least $1-\delta$,

$$2(\mathrm{n}-1)\mathbb{E}_Q \Delta(h)^2 - KL(Q||P) \leq f(S) \leq \log\frac{n}{\delta}.$$

We finish the proof by equation $\left(\mathbb{E}_Q \Delta(h)\right)^2 \leq \mathbb{E}_Q \Delta(h)^2$.

## Take-away messages

**Theorem (PAC-Bayesian).** Given prior distribution $P(h)$ and posterior $Q(h)$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training),

$$L_D(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

**Proof sketch:**

(1) Go from distribution $P$ to distribution $Q$, which causes loss $KL(Q||P)$.

(2) Sup over Q; expectation over P (concentration inequality)

(3) Does $\log n$ comes from the sup term?

Note: from the derivation we can see, PAC-Bayesian still need a sup operator on the distribution Q, and therefore **similar to uniform convergence**.

@ 滕佳烨

All the slides will be available at [www.tengjiaye.com/generalization](www.tengjiaye.com/generalization) soon.

Thanks!