# Machine learning theory

## PAC-Bayesian Theory

Hamid Beigy

Sharif university of technology

May 29, 2023

# Table of contents

# Introduction

1. PAC (Probably Approximately Correct) learning provides guarantees on the expected error (approximately) of prediction rules that hold with high probability (probably) with respect to representativeness of the observed sample.
2. In PAC approach, we choose hypothesis class $H$ as the prior knowledge.
3. The PAC approach has the advantage that one can prove guarantees for generalization error without assuming the truth of the prior.
4. How to incorporate more complicated prior knowledge.

1. The Bayesian approach has the advantage of using arbitrary domain knowledge in the form of a Bayesian prior.
2. A PAC-Bayesian approach to machine learning attempts to combine the advantages of both PAC and Bayesian approaches.
3. A PAC-Bayesian approach bases the bias of the learning algorithm on an arbitrary prior distribution, thus allowing the incorporation of domain knowledge, and yet provides a guarantee on generalization error that is independent of any truth of the prior.

# Bayesian methods

1. Let the data is drawn from a distribution that comes from some parametric family.

**Example (Gaussian distribution)**

Let $\sigma$ be a known fixed parameter. Then, $\mathbb{P}[y \mid \mathbf{x}; \mathbf{w}] = \mathcal{N}\left(\langle \mathbf{w}, \mathbf{x} \rangle, \sigma^2\right) = \langle \mathbf{w}, \mathbf{x} \rangle + \mathcal{N}\left(0, \sigma^2\right)$ is a parametric family.

2. Given a sample $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, we define the likelihood of $\mathbf{w}$ as

$$\mathcal{L}(\mathbf{w}, S) = \log\left(\mathbb{P}[y_1, \ldots, y_m \mid \mathbf{x}_1, \ldots, \mathbf{x}_m; \mathbf{w}]\right) = \sum_{i=1}^{m} \log\left(\mathbb{P}[y_i \mid \mathbf{x}_i; \mathbf{w}]\right)$$

3. The maximum likelihood maximizes $\mathcal{L}(\mathbf{w}, S)$ given value of $S$

$$\mathbf{w} = \operatorname*{argmax}_{\mathbf{w}'} \mathcal{L}(\mathbf{w}', S)$$

**Example (Gaussian distribution)**

1. Let $\sigma$ be a known fixed parameter. Then, $\mathbb{P}[y \mid \mathbf{x}; \mathbf{w}] = \mathcal{N}\left(\langle \mathbf{w}, \mathbf{x}\rangle, \sigma^2\right) = \langle \mathbf{w}, \mathbf{x}\rangle + \mathcal{N}\left(0, \sigma^2\right)$ is a parametric family.
2. This means that $\mathbb{P}[y_i \mid \mathbf{x}_i; \mathbf{w}] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle \mathbf{w}, \mathbf{x}\rangle)^2}{\sigma^2}\right)$ and the likelihood is $\mathcal{L}(\mathbf{w}, S) = -\sum_{i=1}^{m} \frac{1}{\sigma^2} \frac{(y_i - \langle \mathbf{w}, \mathbf{x}\rangle)^2}{\sigma^2} + C$, where $C$ is a normalization factor that does not depend on $\mathbf{w}$.
3. This means that maximum likelihood is equivalent to minimizing square loss.
4. We want to maximize $\mathbb{P}[\mathbf{w} \mid \mathbf{x}, y]$.

1. To find $\mathbb{P}\left[\mathbf{w} \mid \mathbf{x}, y\right]$, we need to a prior distribution $\mathbb{P}\left[\mathbf{w}\right]$.
2. We have $\mathbb{P}\left[y \mid \mathbf{x}, \mathbf{w}\right]$ and $\mathbb{P}\left[\mathbf{w}\right]$ from Bayes Theorem, hence, we have

$$\begin{aligned} \mathbb{P}\left[\mathbf{w} \mid \mathbf{x}, y\right] &= \frac{\mathbb{P}\left[y \mid \mathbf{x}, \mathbf{w}\right] \mathbb{P}\left[\mathbf{w}\right]}{\mathbb{P}\left[y \mid \mathbf{x}\right]} \\ &\propto \mathbb{P}\left[y \mid \mathbf{x}, \mathbf{w}\right] \mathbb{P}\left[\mathbf{w}\right]. \end{aligned}$$

3. The maximum a posteriori (MAP) model is

$$\begin{aligned} \mathbf{w} &= \underset{\mathbf{w}'}{\operatorname{argmax}}\ \mathbb{P}\left[\mathbf{y} \;\middle|\; \mathbf{X}, \mathbf{w}'\right] \mathbb{P}\left[\mathbf{w}'\right] \\ &= \underset{\mathbf{w}'}{\operatorname{argmax}}\ \mathcal{L}(\mathbf{w}', S) + \log \mathbb{P}\left[\mathbf{w}'\right] \end{aligned}$$

**Example (Gaussian distribution (cont.))**

1. Let $\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(0, \sigma^2_{\mathbf{w}}\mathbf{I})$ be prior distribution on $\mathbf{w}$.
2. Now, we have

$$\begin{aligned}\mathbf{w} &= \underset{\mathbf{w}'}{\operatorname{argmax}} \; -\sum_{i=1}^{m} \frac{1}{\sigma^2} \frac{\left(y_i - \langle \mathbf{w}', \mathbf{x} \rangle\right)^2}{\sigma^2} - \frac{1}{\sigma^2} \left\|\mathbf{w}'\right\|_2^2 \\ &= \arg\min \mathbf{w}' \sum_{i=1}^{m} \frac{1}{\sigma^2} \frac{\left(y_i - \langle \mathbf{w}', \mathbf{x} \rangle\right)^2}{\sigma^2} + \frac{1}{\sigma^2} \left\|\mathbf{w}'\right\|_2^2\end{aligned}$$

3. This is equivalent to doing regularized ERM with $L_2$ regularization.
4. If we use Laplacian distribution instead of Gaussian, we will get $L_1$ regularization.

1. MAP picks the best model, given our model and data.
2. Why do we have to pick one model?
3. We have seen that the optimal classifier can be calculated given $\mathbb{P}[y \mid \mathbf{x}]$.
4. The Bayesian approach does exactly that, so we get

$$\mathbb{P}[y \mid \mathbf{x}, S] = \int_{\mathbf{w}} \mathbb{P}[y \mid \mathbf{x}, \mathbf{w}]\, \mathbb{P}[\mathbf{w} \mid S]\, d\, \mathbb{P}[\mathbf{w}]$$

5. In some cases (such as Guassian), this as an analytic solution, but most of the time there isn't any.

# PAC-Bayes theory

1. In agnostic PAC learning, this prior is defined as selecting the hypothesis class $H$.
2. In SRM learning, this prior is defined as the weights assigned to different hypothesis class $H_n$.
3. In MDL, this prior is defined as the description length of hypothesis $h$.
4. In the above models, the output of the learning algorithm is a single hypothesis $h$, i.e $h = A(S)$.
5. In PAC-Bayes, algorithms return a distribution $Q$ on $H$.
6. The learning algorithm is
   - Define prior distribution $P$ on $H$.
   - Get sample $S \sim \mathcal{D}^m$.
   - Define/find posterior distribution $Q$ on $H$.
7. Note that distributions play two different semantic roles:
   - $\mathcal{D}$ is a model of the world;
   - $P$ and $Q$ express our beliefs about the correct answer.

**Example (Loss of posterior)**

1. Let $Q$ be a distribution on $H$, $\mathcal{D}$ a distribution on $\mathcal{X} \times \mathcal{Y}$ and $S$ a finite sample.
2. Define

$$\mathbf{R}(Q) = \mathop{\mathbb{E}}_{h \sim Q} [\mathbf{R}(h)] = \mathop{\mathbb{E}}_{h \sim Q} \left[ \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell(h, z)] \right]$$

$$\hat{\mathbf{R}}(Q) = \mathop{\mathbb{E}}_{h \sim Q} \left[ \hat{\mathbf{R}}(h) \right] = \mathop{\mathbb{E}}_{h \sim Q} \left[ \frac{1}{m} \sum_{i=1}^{m} \ell(h, z) \right]$$

We can turn a posterior into a learning algorithm.

**Definition (Gibbs classifier)**

Let $Q$ be a distribution on $H$. The Gibbs classifier is the following randomized hypothesis

- Pick $h \in H$ according to $Q(h)$.
- Observe $\mathbf{x}$.
- Return $h(\mathbf{x})$.

It is straightforward to show that the expected loss Gibbs classifier equals to $\mathbf{R}(Q)$.

**Example**

1. Let $H = \{h_1, \ldots, h_k\}$.
2. Let $P$ be a uniform distribution over $H$.
3. Let $Q$ be defined as

$$Q(h) = \begin{cases} 1 & \text{if } h = h_{erm} \\ 0 & \text{if } h \neq h_{erm} \end{cases}$$

**Example**

1. For $\mathbf{w} \in \mathbb{R}^n$, define

$$h_{\mathbf{w}}(\mathbf{x}) = \begin{cases} +1 & \text{with probability } \frac{1}{Z} e^{\langle \mathbf{w}, \mathbf{x} \rangle} \\ -1 & \text{with probability } \frac{1}{Z} e^{-\langle \mathbf{w}, \mathbf{x} \rangle} \end{cases}$$

2. The prior $P$ is $\mathcal{N}\left(0, \sigma^2 \mathbf{I}\right)$, i.e. $P(h_{\mathbf{w}}) \propto \exp(-\left\|\mathbf{w}\right\|^2 / \sigma^2)$.
3. Given sample $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \sim \mathcal{D}^m$, obtain $Q$, sample $h \sim Q$, and output $h(x)$. Then likelihood equals to

$$\mathbb{P}\left[y_1, \ldots, y_m \mid h_{\mathbf{w}}, \mathbf{x}_1, \ldots, \mathbf{x}_m\right] = \prod_i \frac{1}{Z} e^{\langle \mathbf{w}, \mathbf{x}_i \rangle} \propto \exp\left(\sum_i y_i \left\langle \mathbf{w}, \mathbf{x}_i \right\rangle\right).$$

4. Using Bayes' rule, we can form the posterior

$$\begin{aligned}\mathbb{P}\left[h_{\mathbf{w}} \mid y_1, \ldots, y_m, \mathbf{x}_1, \ldots, \mathbf{x}_m\right] &\propto \left(\exp\left(\sum_i y_i \left\langle \mathbf{w}, \mathbf{x}_i \right\rangle\right)\right)\left(\exp\left(-\frac{\left\|\mathbf{w}\right\|^2}{\sigma^2}\right)\right) \\ &\propto \left(\exp\left(\sum_i y_i \left\langle \mathbf{w}, \mathbf{x}_i \right\rangle\right) - \frac{\left\|\mathbf{w}\right\|^2}{\sigma^2}\right)\end{aligned}$$

1. We want to show that if $Q$ is similar to $P$, the classifier generalizes well.
2. We will see that the critical factor determining the complexity of the learning algorithm will become $KL(Q||P)$, the Kullback-Liebler divergence from $Q$ to $P$ instead of the Rademacher complexity.
3. Kullback-Leibler (KL) divergence is how to measure the similarity of two distributions.

**Definition (KL divergence)**

Let $P$ and $Q$ be continuous or discrete distributions. Then, KL divergence of distributions $P$ and $Q$ defined as

$$KL(Q||P) = \mathop{\mathbb{E}}_{x\sim Q}\left[\ln\left(\frac{Q(x)}{P(x)}\right)\right].$$

4. Note that KL divergence is not symmetric, i.e. $KL(Q||P) \neq KL(P||Q)$.
5. The intuition behind this definition comes from information theory.

1. Assume we have a finite alphabet and message $x$ is sent with probability $P(x)$.
2. Shannon's coding theorem states that code of $x$ with $\log_2(1/P(x))$ bits is an optimal coding and the expected bits per letter is $\mathbb{E}_{x\sim P}\left[\log_2\left(\frac{1}{P(x)}\right)\right] = H(P)$.
3. Consider now that we use the optimal code for $P$, but the letters where sent according to $Q$.
4. The expected bits per letter is now

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim Q}\left[\log_2\left(\frac{1}{P(x)}\right)\right] &= \mathop{\mathbb{E}}_{x\sim Q}\left[\log_2\left(\frac{Q(x)}{P(x)}\right) + \log_2\left(\frac{1}{Q(x)}\right)\right] \\ &= H(Q) + KL(Q||P).\end{aligned}$$

5. $KL(Q||P)$ is the extra number of bits expected per letter from using $P$ instead of $Q$ to create the codebook.
6. This shows that $KL(Q||P) \geq 0$.

**Example**

Let $P$ be a distribution on $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and $Q(\mathbf{x}_i) = 1$. Then, $KL(Q||P) = \ln\left(\frac{1}{P(x_i)}\right)$.

**Example**

Let $P(\mathbf{x}_i) = 0$ and $Q(\mathbf{x}_i) > 0$, then $KL(Q||P) = \infty$.

**Example**

Let $\alpha, \beta \in [0, 1]$, then $KL(\alpha||\beta) = KL(Ber(\alpha)||Ber(\beta)) = \alpha \ln\left(\frac{\alpha}{\beta}\right) + (1-\alpha)\ln\left(\frac{1-\alpha}{1-\beta}\right)$.
**Show the above equation.**

**Example**

Let $Q = \mathcal{N}(\mu_0, \Sigma_0)$ and $P = \mathcal{N}(\mu_1, \Sigma_1)$ be two $n$-dimensional Gaussian distributions. Then, (**Show the following equation.**)

$$KL(Q||P) = \frac{1}{2}\left(\text{Tr}\left[\Sigma_1^{-1}\Sigma_0\right] + (\mu_1 - \mu_0)\Sigma_1^{-1}(\mu_1 - \mu_0) - n - \frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right)$$

**Lemma**

*If $X$ is a real valued random number satisfying $\mathbb{P}[X \leq x] \leq e^{-mf(x)}$, then $\mathbb{E}\left[e^{(m-1)f(x)}\right] \leq m$.*

**Lemma**

*With probability greater than $(1-\delta)$ over $S$,*

$$\mathop{\mathbb{E}}_{h\sim P}\left[e^{(m-1)KL(\hat{\mathbf{R}}(h)||\mathbf{R}(h))}\right] \leq \frac{m}{\delta}.$$

**Lemma (Shift of measure)**

$$\mathop{\mathbb{E}}_{x\sim Q}[f(x)] \leq KL(Q||P) + \ln \mathop{\mathbb{E}}_{x\sim P}\left[e^{f(x)}\right].$$

**Theorem (PAC Bayes bound)**

*Let $Q$ and $P$ be distributions on $H$ and $\mathcal{D}$ be a distribution on$\mathcal{X} \times \mathcal{Y}$. Also let $\ell(h, z) \in [0, 1]$ and $S \sim \mathcal{D}^m$ be a sample of size $m$, then with probability greater or equal to $(1 - \delta)$ over $S$ we have*

$$KL(\hat{\mathbf{R}}(Q) || \mathbf{R}(Q)) \leq \frac{KL(P||Q) + \ln\left(\frac{m+1}{\delta}\right)}{m}.$$

1. The left-hand side is the KL divergence between two numbers; while the right-hand side is the KL divergence between distributions.
2. We assume no connection between $\mathcal{D}$ and $P$ (an agnostic analysis).

**Proof (PAC Bayes bound).**

1. Define $f(h) = KL(\hat{\mathbf{R}}(h)||\mathbf{R}(h))$. Using the Lemma Shift of measure and its preceding lemma, we get
$$\mathop{\mathbb{E}}_{h\sim Q}[mf(h)] \leq KL(Q||P) + \ln \mathop{\mathbb{E}}_{h\sim P}\left[e^{mf(h)}\right] \leq KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)$$
2. Since KL divergence is convex, so from the Jensen inequality
$$\begin{aligned} KL(\hat{\mathbf{R}}(Q)||\mathbf{R}(Q)) &= KL(\mathop{\mathbb{E}}_{h\sim Q}\left[\hat{\mathbf{R}}(h)\right] || \mathop{\mathbb{E}}_{h\sim Q}[\mathbf{R}(h)]) \\ &\leq \mathop{\mathbb{E}}_{h\sim Q}\left[KL(\hat{\mathbf{R}}(h)||\mathbf{R}(h))\right] = \mathop{\mathbb{E}}_{h\sim Q}[f(h)] \end{aligned}$$

□

We bounded $KL(\hat{\mathbf{R}}(Q)||\mathbf{R}(Q))$ and then bound $\mathbf{R}(Q) - \hat{\mathbf{R}}(Q)$.

**Lemma**

*Let $a, b \in [0,1]$ and $KL(a||b) \leq x$, then $b \leq a + \sqrt{\frac{x}{2}}$ and $b \leq a + 2x + \sqrt{2ax}$.*

The second is much stronger if $a$ is very small.

**Theorem (Generalization bounds)**

*Let $Q$ and $P$ be distributions on $H$ and $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Let also $\ell(h,z) \in [0,1]$ and $S \sim \mathcal{D}^m$ be a sample, then with probability greater or equal to $(1-\delta)$ over $S$ we have*

$$\mathbf{R}(Q) \leq \hat{\mathbf{R}}(Q) + \sqrt{\frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{2m}}$$

$$\mathbf{R}(Q) \leq \hat{\mathbf{R}}(Q) + 2\frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{m} + \sqrt{2\hat{\mathbf{R}}(Q)\frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{m}}$$

**Example (Soft-ERM)**

1. In Soft-ERM, we have $Q(h) = \frac{1}{Z_Q} e^{-\beta \hat{\mathbf{R}}(h)}$, where $Z_Q$ is the normalization constant.
2. When $\beta \to 0$, $Q$ is uniform.
3. When $\beta \to \infty$, $Q$ is concentrated on the ERM.
4. Its natural counterpart is the prior $P(h) = \frac{1}{Z_P} e^{-\beta \mathbf{R}(h)}$.
5. We do not know $P$, but we only use it for theoretical analysis.

**Theorem**

*Let $Q$ be the Soft-ERM posterior, then with probability greater or equal to $(1-\delta)$ over $S$ we have*

$$KL(\hat{\mathbf{R}}(Q)||\mathbf{R}(Q)) \leq \frac{\sqrt{2}\beta}{m^{3/2}} \sqrt{\ln\left(\frac{2m+2}{\delta}\right)} + \frac{\beta^2}{2m^2} + \frac{\ln\left(\frac{2m+2}{\delta}\right)}{m}$$

**Homework:** It seems like Soft-ERM is a universal learner! What doesn't it contradict the fundamental theorem in statistical learning?

# Summary

1. Shawe-Taylor et al. gave PAC analysis of Bayesian estimators.
2. McAllester gave PAC-Bayesian bound.
3. PAC-Bayes bounds hold even if prior incorrect; while Bayesian inference must assume prior is correct.
4. PAC-Bayes bounds hold for all posteriors; while in Bayesian learning, posterior computed by Bayesian inference, depends on statistical modeling
5. PAC-Bayes bounds can be used to define prior, hence no need to be known explicitly; while in Bayesian learning, input effectively excluded from the analysis, randomness lies in the noise model generating the output.
6. We analyzed Gibbs classifier. Another solution is to sample many $h_i \sim Q$ i.i.d. and output the majority vote.
7. PAC-Bayes theory gives the tightest known generalization bounds for SVMs, with fairly simple proofs.
8. PAC-Bayesian analysis applies directly to algorithms that output distributions on the hypothesis class, rather than a single best hypothesis.
9. However, it is possible to de-randomize the PAC-Bayes bound to get bounds for algorithms that output deterministic hypothesis.

# Readings

1. Chapter 31 of Shai Shalev-Shwartz and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
2. The papers given in References McAllester 1999, 2003a,b, 2013.

McAllester, David A. (1999). "Some PAC-Bayesian Theorems". In: *Machine Learning* 37.3, pp. 355–363.

— (2003a). "PAC-Bayesian Stochastic Model Selection". In: *Machine Learning* 51.1, pp. 5–21.

— (2003b). "Simplified PAC-Bayesian Margin Bounds". In: *Lecture Notes in Computer Science*. Vol. 2777. Springer, pp. 203–215.

— (2013). "A PAC-Bayesian Tutorial with A Dropout Bound". In: *CoRR* abs/1307.2118.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

**Questions?**