ANALYSIS OF PAC-BAYESIAN BOUNDS FOR GAUSSIAN PROCESSES

FINAL REPORT

Shubh Gupta
Department of Electrical and Computer Engineering
University of Illinois, Urbana-Champaign
Champaign, IL 61820
shubhg2@illinois.edu

May 10, 2019

ABSTRACT

This report explores the PAC-Bayesian theorem which blends Bayesian and frequentist approaches to the theory of machine learning, and analyzes its applications to a non-parametric data driven prior model which is Gaussian Processes. Furthermore, it adapts the problem of estimating the risk associated with the obtained posterior distribution over position space in presence of noisy controls to the above framework with applications to safe autonomy.

***K*eywords** PAC-Bayes · Gaussian Processes · Safe localisation

1 Introduction

Probably Approximately Correct (PAC) learnability is a notion introduced by Valiant in the mathematical theory of learning, which is at the crossroads of computer science, optimization and statistics. The PAC bound can be intuitively understood as the upper-bound of the performance of a learning algorithm; as obtained by a loss function; which decays to an optimal value as more samples are fed (or in other words, is approximately correct) with an arbitrarily high probability. These bounds dont assume any prior knowledge about the hypothesis apart from independent sampling in their derivation, as well as have very little constraints on the hypothesis class and data distribution. Hence, it is widely applicable and is a valuable tool for deriving theoretical guarantees in various learning problems.

Another way to approach a learning problem is the Bayesian perspective, which assumes a joint distribution over input and hypothesis with the objective of inferring the conditional on hypothesis given observations. It gives a principled way of managing randomness and uncertainty and hence has been very useful in a variety of learning problems. The notion of *generalised* Bayes extends the canonical Bayes theory to improve on the predictive capability while trading off interpretability. It does this by dealing with arbitrary measures of quality of performance instead of the likelihood,

such as the tempered likelihood, and allows a shift from the model-based procedure of canonical Bayes to a model-free procedure.

PAC-Bayesian inequalities, introduced by McAllester, combine the theoretical performance deriving capability of PAC bounds with the generalised Bayes framework, and hence provide a powerful tool to derive theoretical measures of the performance of a learning algorithm in the presence of some knowledge about the prior hypothesis distribution. These types of setting are more common in the real world than the former, since we can characterize some form of a bias towards certain concepts in a problem most of the time using previous or auxiliary information.

2 Problem Setting

Assume that the input data $\mathcal{D}_n$ is generated from a list of pairs $(X_i, Y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ each of which is iid sampled from an underlying distribution $\mathbb{P}$. The objective is to determine an optimal estimator $\hat{f} \in \mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}\}$ of Y' for any new sample (X', Y') or in other words, generalises. Now, similar to the bayesian setting, we also have access to some prior information π_0 operating on some $\mathcal{F}_0 \in \mathcal{F}$ about the distribution of $\hat{f}$. The algorithm tries to infer the posterior distribution over the estimator $\pi(\hat{f}|\mathcal{D}_n)$ or simply π

To assess the generalisation capability, we define a loss $l \in \mathbb{R} \times \mathbb{R} \to [0, 1]$ which allows us to define the risk, or expected loss, associated with a hypothesis f as

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(f(x), y)]$$

The empirical risk associated with f is further defined as

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$$

To account for stochastic nature of the hypothesis class, we define an expected risk associated with a posterior distribution π over the hypothesis class

$$\mathcal{L}(\pi) = \mathbb{E}_{f\sim\pi} R(f)$$

and corresponding expected empirical risk

$$\hat{\mathcal{L}}_n(\pi) = \mathbb{E}_{f\sim\pi} \hat{R}_n(f)$$

The PAC-Bayesian bounds deal with estimating (with arbitrary probability) the upper-bound on $\mathcal{L}(\pi)$; which cannot be computed due to lack of knowledge about $\mathcal{D}$; using $\hat{\mathcal{L}}_n(\pi)$ and other terms which can be computed.

3 PAC-Bayesian bounds

We discuss three kinds of PAC-Bayesian bounds depending upon different constraints on the learning problem. The bounds build incrementally, with each subsequent bound adding a level of complexity over the former.

3.1 Occam Bound[1]

The Occam PAC-Bayesian bounds are derived assuming a discrete(countable) hypothesis class $\mathcal{F} = \{f_i : i \in \mathbb{N}\}$.

Theorem 1. *With probability at least 1-δ over iid draws from $\mathcal{D}_n$ we have that the following holds for all f.*

$$R(f) \leq \inf_{\lambda>0.5} \frac{1}{1-\frac{1}{2\lambda}} \left(\hat{R}_n(f) + \frac{\lambda}{N} \left(ln(\frac{1}{\pi_0(f_i)}) + ln(\frac{1}{\delta}) \right) \right) \tag{1}$$

Remarks Above theorem is derived from a simple application of Chernoff bound and equating with the $\pi_0(f_i)\delta$ contribution of ith hypothesis to get

$$\epsilon(f_i) = \sqrt{\frac{2R(f_i)\left(ln(\frac{1}{\pi_0(f_i)}) + ln(\frac{1}{\delta}) \right)}{N}}$$

The expression can then be converted to the theorem by application of the following property which holds by Jensen's inequality

$$\sqrt{ab} = \inf_{\lambda>0} \frac{a}{2\lambda} + \frac{\lambda b}{2}$$

An interesting application of this theorem is in bounding the finite precision problem where the parameter characterizing the hypothesis class is in $\mathbb{R}^d$ and each of its components is represented in atmost b bits. Hence, we can assume that all possible 2^{bd} hypothesis are equiprobable, which gives us the bounds

$$R(f) \leq \inf_{\lambda>0.5} \frac{1}{1-\frac{1}{2\lambda}} \left(\hat{R}_n(f) + \frac{\lambda}{N} \left(ln(2)bd + ln(\frac{1}{\delta}) \right) \right) \tag{2}$$

Another interesting bound is obtained if we know that our test hypothesis has a sparsity level of s in the representation. The probability can then be broken into a probability of selecting s uniformly from d choices, selecting those s components from d choices one by one and drawing b-bit components for each choice. Hence the bound becomes

$$R(f) \leq \inf_{\lambda>0.5} \frac{1}{1-\frac{1}{2\lambda}} \left(\hat{R}_n(f) + \frac{\lambda}{N} \left(lnd + s(lnd + (ln(2)b) + ln(\frac{1}{\delta}) \right) \right) \tag{3}$$

4 Cantoni's bound[2]

McAllester followed by Cantoni derived the bounds on expected risk for the continuous hypothesis class, which are said to be the first PAC-Bayesian bounds. A variant is as follows

Theorem 2. *With probability at least 1-δ over iid draws from $\mathcal{D}_n$ we have that the following holds for all distributions π on $\mathcal{F}$.*

$$\mathcal{L}(\pi) \leq \inf_{\lambda>0.5} \frac{1}{1-\frac{1}{2\lambda}} \left(\hat{\mathcal{L}}_n(\pi) + \frac{\lambda}{N} \left(KL(\pi\|\pi_0) + ln(\frac{1}{\delta}) \right) \right) \tag{4}$$

Remarks The bound is quite similar to Occam bound for discrete hypothesis, with the only change being KL divergence term instead of negative log likelihood of the hypothesis, which are essentially similar in spirit. Intuitively, KL divergence assigns a negative log likelihood to continuous hypothesis distributions on the basis of their distance from the prior. Hence, similarity to prior lowers the bound and vice versa.

The theorem can be proved by bounding the bernoulli KL divergence between expected risk and expected empirical risk $KL_\gamma(\hat{\mathcal{L}}\|\mathcal{L})$ by $KL(\pi\|\pi_0)$ using the shift of measure lemma of KL divergence and convexity. Finally the bound can be derived using the property for any $\lambda > 0.5$

$$KL_{-1/\lambda}(p\|q) \leq c \Rightarrow p \leq \frac{1}{1-\frac{1}{2\lambda}}(q + \lambda c)$$

which can be derived by simple algebraic manipulations. An interesting application of this bound is in the derivation of generalisation bounds for Dropout in neural networks. Assuming that the weights of a layer are dropped to ϵ with probability α and are $\theta_i + \epsilon$ otherwise where θ is the trainable parameter and $\epsilon \sim \mathcal{N}(0,1)$. The prior here can be taken to $N(0,1)$ while posterior is distribution of r.v. $s \cdot \theta + \bar{\epsilon}$ where s_i is 0 w.p. α. KL divergence can be computed as

$$KL(\pi\|\pi_0) = \mathbb{E}_s[\frac{1}{2}\|s \cdot \theta\|^2] = \frac{1-\alpha}{2}\|\theta\|^2$$

Hence the PAC-Bayesian bound can be computed as

$$\mathcal{L}(\pi_\theta) \leq \inf_{\lambda > 0.5} \frac{1}{1-\frac{1}{2\lambda}} \left(\hat{\mathcal{L}}_n(\pi_\theta) + \frac{\lambda}{N} \left(\frac{1-\alpha}{2}\|\theta\|^2 + ln(\frac{1}{\delta}) \right) \right) \tag{5}$$

4.1 Training Variance Bound[1]

The third type of bound is derived from the Cantoni bound when the learning algorithm $\mathcal{A}$ is specified. In presence of the algo, the bound is obtained by noting that the KL divergence term $KL(\pi\|\pi_0)$ is minimized in expectation for $\pi_0 = \mathbb{E}_{\mathcal{A}}[\pi]$. The obtained $KL(\pi\|\mathbb{E}_{\mathcal{A}}[\pi])$ can be intuitively seen as a measure of variance in the induced hypothesis subclass by the algorithm A which gives us a tighter bound

Theorem 3. *With probability at least 1-δ over iid draws from $\mathcal{D}_n$ we have that the following holds for all distributions π on $\mathcal{F}_{\mathcal{A}}$ induced by learning algorithm.*

$$\mathcal{L}(\pi) \leq \inf_{\lambda > 0.5} \frac{1}{1-\frac{1}{2\lambda}} \left(\hat{\mathcal{L}}_n(\pi) + \frac{\lambda}{N} \left(KL(\pi\|\mathbb{E}_{\mathcal{A}}[\pi]) + ln(\frac{1}{\delta}) \right) \right) \tag{6}$$

However, the term $\mathbb{E}_{\mathcal{A}}[\pi]$ cannot be computed since the distribution of hypothesis under A in unknown. The term can however be approximated as a weighted average of samples from the distribution $\{f_i\}_{i=1}^N$, with weights $w_i = e^{\frac{-N\mathcal{L}(f_i)}{\lambda}}$ This bound can be further combined with inequalities on this KL divergence to lead to a KL divergence independent bound, albeit vacuous.

5 Gaussian Process PAC-Bayesian bounds

Gaussian processes serve as a discriminative model (modelling conditional of labels given input) to formalise prior knowledge about the task in order to develop data dependent complexity measures for generalisation error[3]. Gaussian processes impose a non-parametric model over the label distribution, such that distribution of labels given any finite subset of inputs is a gaussian[4].

The gaussian process is characterized by a kernel function K. Hence for given training data $(X,Y)^n$ prior over hypothesis f is specified by the GP prior as $\pi_0 = \mathcal{N}(0, K_{XX})$. The posterior $\pi =$

$\mathcal{N}(K_{XX}\alpha, \Sigma)$ is also paramatrized by a gaussian where α, Σ are arbitrary. The final classification is performed by a Gibbs classifier on GP output. Now,

$$KL(\pi\|\pi_0) = \int \left[-\frac{1}{2}\log\frac{|\Sigma|}{|K_{XX}|} - \frac{1}{2}(f)^T\Sigma^{-1}(f) + \frac{1}{2}(f - K_{XX}\alpha)^T K_{XX}^{-1}(f - K_{XX}\alpha)\right] \times \pi_0(f)df \tag{7}$$

$$= -\frac{1}{2}\log\frac{|\Sigma|}{|K_{XX}|} - \frac{1}{2}tr\left\{E[ff^T]\Sigma^{-1}\right\} + \frac{1}{2}E[(f - K_{XX}\alpha)^T K_{XX}^{-1}(f - K_{XX}\alpha)] \tag{8}$$

$$= -\frac{1}{2}\log\frac{|\Sigma|}{|K_{XX}|} - \frac{1}{2}tr\{I_n\} + \frac{1}{2}(K_{XX}\alpha)^T K_{XX}^{-1}(K_{XX}\alpha) + \frac{1}{2}tr\{\Sigma^{-1}K_{XX}\}^{-1} \tag{9}$$

$$= \frac{1}{2}\log|\Sigma^{-1}K_{XX}| + \frac{1}{2}tr\{\Sigma^{-1}K_{XX}\}^{-1} + \frac{1}{2}\alpha^T K_{XX}\alpha - \frac{n}{2} \tag{10}$$

Plugging in the PAC-Bayesian bound by Cantoni we get our GP bound[5]

$$\mathcal{L}(\pi) \leq \inf_{\lambda>0.5} \frac{1}{1-\frac{1}{2\lambda}}\left(\hat{\mathcal{L}}_n(\pi) + \frac{\lambda}{N}\left(\frac{1}{2}\log|\Sigma^{-1}K_{XX}| + \frac{1}{2}tr\{\Sigma^{-1}K_{XX}\}^{-1} + \frac{1}{2}\alpha^T K_{XX}\alpha - \frac{n}{2} + ln(\frac{1}{\delta})\right)\right) \tag{11}$$

6 Experiment

As an attempt at adapting the generalisation bounds obtained for GP classification to different domains, we apply the bounds on the problem of risk evaluation in localisation of an agent under noisy controls. The hypothesis space $\mathcal{F}$ consists of distributions over $\mathbb{R}^2$. The input is a control U in $\mathbb{R}^2$ which along with measurement generates a resultant distribution of the agent's position over the position space. The label Y associated with a position X is determined as a function of this resultant distribution. Dataset $(U, Y)^n$ is generated for X by randomly sampling U and generating labels. A Gaussian Process for generating these labels is now learnt over controls U with Radial Basis Function (RBF) based Kernel. The risk associated with labelling X as the true position is computed as risk in the GP learning process as well as negative classifications for highly probable inputs.

The results exhibit a good detection of True positives, however, false positives are also highly detected indicating that the bound is vacuous.

Case	**frequency (%)**
High risk, False position(TP)	32
High risk, True position(FP)	27
Low risk, False position(FN)	18
Low risk, True position(TN)	23

Table evaluating the risk metric. Equal number of True and False positions.

7 Conclusion

In this report, we explored the PAC-Bayes framework as well as the various associated bounds and applications. In particular, we analysed the application of the theorem to Gaussian Processes in a classification setting. We adapted this bound for evaluating the risk associated with localisation

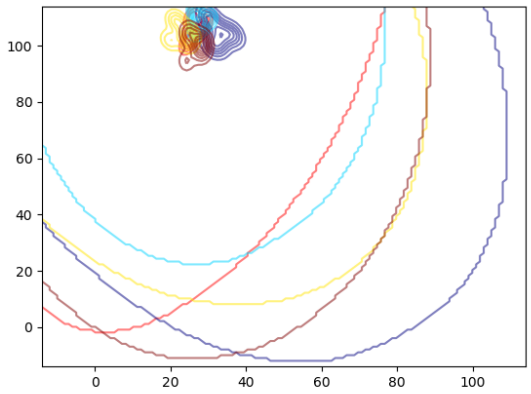

Figure 1: Position distribution for a few different control samples

under noisy controls, and find that the risk bound has a bias towards detecting risk events, which is also intuitive since this is an upper bound.

References

[1] David McAllester. A PAC-Bayesian Tutorial with A Dropout Bound. 2013.

[2] Benjamin Guedj. A Primer on PAC-Bayesian Learning. jan 2019.

[3] David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds. oct 2018.

[4] Matthias Seeger. *Gaussian processes for machine learning.*, volume 14. 2004.

[5] Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3(2):233–269, 2003.