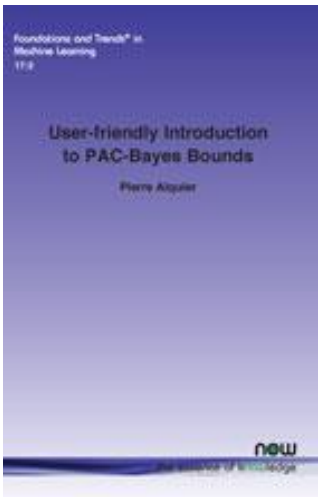
User-friendly Introduction to PAC-Bayes Bounds
Pierre ALQUIER
Foundations and Trends® in Machine Learning 17-2



Le livre de Pierre Alquier s'inscrit dans la riche collection *Foundations and Trends®*, série de monographies plutôt courtes d'introduction à des domaines assez circonscrits d'économie et d'ingénierie, à mi-chemin entre le manuel universitaire et la revue scientifique. Le thème de *Machine Learning* abordé ici, la recherche de majorants de l'erreur de généralisation, joue évidemment un rôle majeur dans l'élaboration des algorithmes.

De quoi parle-t-on ?

Tout d'abord, l'acronyme PAC est un raccourci pour l'anglais *Probably Approximately Correct*, terminologie introduite dans les années 1980 et traduisant une façon de quantifier la performance d'apprentissage [1]. Une bonne formule vaut mieux qu'un long discours ; une borne PAC typique se présentera sous la forme suivante (voir les notations et quelques rappels en annexe) :

$$\mathbb{P}_S\left(R(\hat{\theta}) \leq r(\hat{\theta}) + C\sqrt{\frac{log\frac{M}{\delta}}{2n}}\right) \geq 1-\delta$$

Ce type de résultat permet en particulier d'exercer un certain niveau de contrôle sur l'erreur de généralisation d'un modèle, paramétré par $\hat{\theta}$ estimé à partir d'un échantillon de données., En prenant par exemple $\delta = 0{,}05$, soit avec une probabilité d'au moins 95%, le risque d'erreur (R) sur de nouvelles données s'éloignera au maximum de la borne $C\sqrt{\frac{log\frac{M}{\delta}}{2n}}$, par rapport au risque (r) calculé sur un échantillon de taille n. Cette expression fait apparaître deux constantes, reflets d'autant d'hypothèses utilisées pour l'établir :

- La fonction de perte est bornée par une constante C.
- L'espace des paramètres est fini et de cardinal M.

On notera par ailleurs une précision évoluant en $\frac{1}{\sqrt{n}}$ avec la taille de l'échantillon, situation habituelle en matière d'échantillonnage, qu'on aimerait améliorer.

L'approche PAC-Bayes objet du livre est apparue une dizaine d'années plus tard et permet d'élargir ce cadre restreint, en particulier avec la possibilité de considérer des espaces de paramètres infinis, non discrets et éventuellement non bornés.

Une inspiration bayésienne

Comme on peut le deviner, cette voie adopte une logique de statistique bayésienne avec la manipulation de distributions de probabilité sur les paramètres à estimer. Il s'agit de véhiculer une connaissance *a priori* sur le problème, que l'on confronte aux données observées et transforme en estimateurs *a posteriori* (sans pour autant reposer nécessairement sur le théorème de Bayes). Une des motivations provient de l'agrégation ou de la « randomisation » de modèles : par exemple, en apprentissage supervisé, des prédicteurs agrégés sont obtenus en affectant certains poids aux résultats d'un ensemble de prédicteurs, c'est-à-dire selon une certaine distribution de probabilité. Les prédicteurs randomisés, eux, sont issus d'un échantillonnage dans un ensemble de prédicteurs de base, selon une distribution de probabilité prédéfinie. Dans les deux cas, la démarche s'appuie sur une loi de probabilité sur l'ensemble des prédicteurs et les bornes PAC-Bayes contribuent à en comprendre l'erreur de généralisation.

Au cours des dernières décennies, le sujet a fait l'objet de nombreux travaux de recherche dans différentes directions, avec un intérêt particulier ces dernières années en raison des progrès accomplis dans l'application aux réseaux de neurones, et plus généralement en *Machine Learning*. L'abondance des travaux a rendu le domaine quelque peu intimidant quand on souhaite se familiariser rapidement avec le sujet. Fidèle aux intentions du titre de l'ouvrage, l'auteur guide en douceur le lecteur dans le vaste corpus de recherche, au fil d'une complexité progressive. Il fournit une introduction conviviale qui éclaire la théorie de base et pointe vers les publications les plus pertinentes pour approfondir un aspect particulier.

L'objectif premier n'est pas tant de présenter les résultats connus les plus précis ou de plus grande portée, que de combler un relatif déficit d'exposés intermédiaires entre les textes d'introduction au domaine et les articles à la pointe de la recherche. La priorité est alors donnée aux énoncés dont la démonstration met en lumière les idées et techniques probabilistes jouant un rôle-clé et omniprésent dans les raisonnements : le plus souvent des inégalités de concentration plus ou moins fines, mais aussi de simples « trucs » astucieux basés par exemple sur l'inégalité de Boole[1].

Un ingrédient-clé et différents points de vue

La majeure partie des résultats dérive donc de l'application d'inégalités dites de concentration, qui bornent la probabilité qu'une variable aléatoire s'écarte d'une certaine valeur, son espérance le plus souvent. Contrairement au théorème central limite, qui est asymptotique, elles conduisent à des propriétés valides pour des échantillons finis. Les plus connues et basiques, l'inégalité de Markov et celle de Bienaymé-Tchebychev, se rencontrent généralement dès les premiers cours de probabilité. La formule exhibée plus haut repose sur un outil plus élaboré et très utilisé en apprentissage machine, l'inégalité de Hoeffding.

De multiples variantes sont aussi utilisées, dont l'inégalité de Chernoff (combinaison d'exponentielle et d'inégalité Markov) également très sollicitée, ou encore celles de Bernstein, de Plinsker [2]. Une des raisons est que la quantification de l'erreur de généralisation peut prendre différentes formes selon le contexte et la façon d'appréhender l'aléa. Plutôt qu'une

[1] Ou l'anglais *Union Bound* : la probabilité de la réunion d'une famille finie ou dénombrable d'évènements est majorée par la somme des probabilités de chaque évènement.

formulation en probabilité, comme exprimée plus haut, on peut tout aussi bien rechercher une majoration de l'erreur en espérance et parler alors d'*Expectedly Approximately Correct*. Dans le même ordre d'idées, si on reprend l'exemple des estimateurs randomisés ou agrégés issus d'une loi de probabilité P sur le paramètre θ, il est possible de considérer différentes quantités (voir les notations en annexe) :

- Le risque d'un estimateur randomisé $\bar{\theta}$ tiré dans la loi P : $R(\bar{\theta})$.
- L'espérance correspondante : $\mathbb{E}_P R(\theta)$.
- Le risque d'un estimateur agrégé selon P : $R(\mathbb{E}_P f_\theta)$.

Dans une autre optique s'opère la distinction entre :

- Borne ou risque empirique, numériquement calculable à partir des données, par exemple du type : $\mathbb{P}_S\left(R\left(\hat{\theta}_{ERM}\right) \leq 0{,}1\right) \geq 0{,}95$. Remarquons qu'on dispose alors d'une borne majorante explicite, mais sans savoir comment la situer par rapport à son minimum théorique. Un des intérêts pratiques majeurs de cette quantité est d'être utilisable comme objectif d'optimisation dans un algorithme d'apprentissage.
- Excès de risque ou oracle : $\mathbb{P}_S\left(R\left(\hat{\theta}_{ERM}\right) \leq \inf_\theta R(\theta) + e_n(\delta)\right) \geq 1 - \delta$, qui exprime un surplus de risque $e_n(\delta)$ par rapport au minimum théorique $\inf_\theta R(\theta)$, mais ce dernier n'est pas numériquement calculable puisque R est inconnu.

Les deux notions présentent leur propre utilité, tout en étant complémentaires et s'alimentant l'une et l'autre dans les deux sens.

Amélioration et généralisation

Après l'exposé de concepts et résultats de base, l'ouvrage aborde les principaux axes d'amélioration et de généralisation. Un des objectifs prioritaires est naturellement la recherche de bornes majorantes les plus petites possibles. A l'extrême, il n'est pas rare en effet avec cette approche d'aboutir à une borne « vide d'information », c'est-à-dire une inégalité triviale. Dans le cas, par exemple, d'un problème de classification, avec une fonction de perte 0-1 (taux d'erreur) $R(\theta)$ est par définition à valeur dans l'intervalle $[0{,}1]$ et toute borne plus grande que 1 est dénuée de contenu.

Une autre direction effleurée plus haut s'attache à améliorer le lien de dépendance avec la taille n de l'échantillon de données. Dans certaines situations, on peut accélérer la décroissance des bornes avec une relation en $\frac{1}{n}$ au lieu de $\frac{1}{\sqrt{n}}$. C'est notamment le cas en apprentissage profond. Il est également montré comment légitimer une distribution a priori dépendant des données d'échantillon.

Les possibilités de généralisation sont par ailleurs explorées, avec différentes tentatives de s'affranchir des hypothèses restrictives souvent adoptées dans les travaux d'origine : échantillon de données indépendantes et identiquement distribuées (i.i.d.), fonction de perte bornée. Dans le premier cas, le recours à des notions de dépendance faible permet de mobiliser des outils techniquement plus sophistiqués, applicables principalement à des observations de type séries temporelles ou champs aléatoires.

Liens avec d'autres approches

La dernière partie du livre donne un aperçu des interactions fructueuses avec d'autres aspects de statistique et d'apprentissage machine : Inférence bayésienne, minimisation du risque empirique, apprentissage séquentiel ou incrémental (*Online Learning*), agrégation d'estimateurs. Elle met particulièrement en avant combien certaines formes d'équivalence

avec des concepts de théorie de l'information sont longtemps passées assez inaperçues. On peut y entrevoir en passant une manifestation [3] :

- des liens étroits entre différents terrains de la grande dimension : physique statistique, thermodynamique, réseaux de neurones, etc. ;
- du rôle médiateur de la théorie de l'information et de son point de vue mathématique sur la notion d'entropie et ses dérivées, révélées ainsi pertinentes dans d'autres contextes que celui de la physique où elles ont émergé.

L'exposé mathématique dans cette fin d'ouvrage est beaucoup plus heuristique que dans les chapitres précédents et vise essentiellement à élargir le cadre et mettre en lumière les perspectives, en renvoyant vers de nombreuses références bibliographiques.

ANNEXE

Le rapide exposé qui suit, schématique et assez informel, n'a d'autre objectif que préciser quelques notations et termes utilisés dans le corps du texte.

Dans l'ouvrage, le cadre de référence est celui de l'apprentissage supervisé paramétrique. On dispose d'un échantillon S de couples de *données* $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ pour $i = 1, \dots, n$. On suppose qu'il existe une loi de probabilité P (inconnue) sur l'ensemble produit $\mathcal{X} \times \mathcal{Y}$, dont les données sont tirées de façon indépendante (hypothèse i.i.d. pour indépendantes et identiquement distribuées). On cherche à identifier une relation fonctionnelle de la forme $f: \mathcal{X} \longrightarrow \mathcal{Y}$, qui permet de « prédire au mieux » la valeur de y connaissant x : $y = f(x)$. Dans un modèle paramétrique, la recherche est circonscrite à un ensemble de *paramètres* Θ et à des *prédicteurs* $f_\theta: \mathcal{X} \longrightarrow \mathcal{Y}$, paramétrés par $\theta \in \Theta$.

La quête d'un meilleur *estimateur* $\hat{\theta}$ des paramètres repose sur le choix d'une *fonction de perte* l définie sur $\mathcal{Y} \times \mathcal{Y}$, qualifiant l'écart entre valeur prédite et vraie valeur : $l(f(x), y)$. Pour un prédicteur donné f_θ, on peut alors s'intéresser au risque d'*erreur de généralisation* sur de nouvelles données, vu en espérance par rapport à la loi P :

$$R(\theta) = R(f_\theta) = \mathbb{E}_{(X,Y)\rightsquigarrow P} l(f_\theta(X), Y)$$

La loi P étant inconnue, cette quantité n'est pas accessible, mais à défaut on peut calculer un *risque empirique* avec les observations dont on dispose :

$$r(\theta) = r(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

Un échantillon étant par hypothèse issu de tirages dans la loi P, on a : $\mathbb{E}_S r(\theta) = R(\theta)$.
$\mathbb{E}_S$ et $\mathbb{P}_S$ désignent respectivement une espérance et une probabilité calculées par rapport à la loi de l'échantillon, i.e. celle de n couples de variables aléatoires (X_i, Y_i) i.i.d. de loi P.

Dans ce contexte, un estimateur naturel du jeu de paramètres est celui qui minimise le risque empirique, noté $\hat{\theta}_{ERM}$ pour *Empirical Risk Minimizer* (si tant est qu'il existe et soit unique).

Références

[1] Papier d'origine : Leslie Valiant (1984), A theory of the learnable, *Communications of the ACM* 27(11). dl.acm.org/doi/10.1145/1968.1972

[2] Une référence sur le sujet : Stéphane Boucheron, Gabor Lugosi, Pascal Massart (2013), *Concentration inequalities*, Oxford University Press.

[3] Le lecteur intéressé peut se plonger dans le cours annuel 2022-2023 de Stéphane MALLAT au Collège de France (chaire Science des données). Les vidéos et notes de cours sont librement accessibles en ligne.

Mots-clés : Machine Learning, apprentissage PAC, erreur de généralisation, inégalités de concentration.

Pierre ALQUIER (ENSAE 2003) est titulaire d'un doctorat et d'une HDR de l'Université Pierre et Marie Curie. Il a occupé différents postes d'enseignant et de chercheur à Dauphine, Paris VII, University College Dublin, à L'ENSAE et à l'institut RIKEN au Japon. Il est aujourd'hui professeur à l'ESSEC.