A COMBINATORIAL CONJECTURE FROM PAC-BAYESIAN MACHINE LEARNING

MALIK YOUNSI AND ALEXANDRE LACASSE

Abstract. We present a proof of a combinatorial conjecture from the second author’s Ph.D. thesis [9]. The proof relies on binomial and multinomial sums identities. We also discuss the relevance of the conjecture in the context of PAC-Bayesian machine learning.

1. PAC-Bayes bounds in machine learning.

In machine learning, ensemble methods are used to build stronger models by combining several base models. In the setting of classification problems, the model, called classifier, can be combined using a weighted majority vote of the base classifiers. Examples of learning algorithms built using such classifiers include Bagging [1] and Boosting [4].

The PAC-Bayes theorem, introduced by McAllester in the 1990’s ([13], [14], [15]), provides bounds on the expectation of the risk of the base classifiers among a majority vote classifier. These bounds are computed from the empirical risk of the majority classifier made during the training phase of a learning algorithm, and can then be used to provide guarantees on existing algorithms or to design new ones.

The version of the PAC-Bayes theorem in [7] involves a function \( \xi : \mathbb{N} \to \mathbb{R} \) defined by

\[
\xi(m) := \sum_{k=0}^{m} \binom{m}{k} \left( \frac{k}{m} \right)^k \left( 1 - \frac{k}{m} \right)^{m-k}
\]

and improves upon other versions, such as the ones in [11] and [12] for instance, which provide the bounds \( m + 1 \) and \( 2\sqrt{m} \) respectively using approximations. The version in [7], on the other hand, is based on a direct calculation in order to obtain \( \xi(m) \).

Bounds on the risk of the majority vote itself can be deduced from the bounds given by the PAC-Bayes theorem. Such bounds, however, turn out to be greater than the bounds on the base voters. In order to circumvent this issue and obtain better bounds, a new version of the PAC-Bayes theorem was given in [10]. This new theorem can be seen as a two-dimensional version, since it gives bounds simultaneously on the expectation of the joint errors and on the joint success of pairs of voters in a majority vote. As shown in [10], this approach does indeed lead to better bounds on the risk of the majority vote. The method of [10] is based on

Date: June 4, 2020.

Key words and phrases. Combinatorial identities, Binomial sums, multinomial sums, PAC-Bayesian machine learning.

The first author was supported by NSF Grant DMS-1758295.
techniques from [7] and involves another combinatorial sum function $\xi_2 : \mathbb{N} \to \mathbb{R}$ defined by

$$\xi_2(m) := \sum_{j=0}^{m} \sum_{k=0}^{m-j} \binom{m}{j} \binom{m-j}{k} \binom{j}{m-k} \binom{k}{m} (1 - \frac{j}{m} - \frac{k}{m})^{m-j-k}.$$  

The function $\xi_2$ was also used in [8] and [9].

2. A COMBINATORIAL CONJECTURE.

In [9], the second author posed the following conjecture, based on numerical evidence.

Conjecture 2.1 (Lacasse [9]). For every $m \in \mathbb{N}$, we have

$$\xi_2(m) = \xi(m) + m.$$  

In this paper, we present a proof of Conjecture 2.1 due to the first author that appeared in a an unpublished manuscript on the arXiv [24]. The proof uses binomial and multinomial sums identities in order to obtain simpler expressions for the functions $\xi$ and $\xi_2$. It has been cited in several publications related to machine learning, such as [3], [6], [8], [16], [17], [18], [19], [22]. We also mention that although other proofs of Conjecture 2.1 were subsequently obtained (see e.g. [2], [5], [20], [23]), the proof we present remains, as far as we know, the only one providing equivalent expressions for the functions $\xi$ and $\xi_2$ that are simpler and more convenient from a numerical perspective. Furthermore, none of the aforementioned papers [2], [5], [20], [23] discuss the context and relevance of Conjecture 2.1 in the theory of machine learning.

3. PROOF OF CONJECTURE 2.1

In this section, we present the proof of Conjecture 2.1 from [24].

3.1. An equivalent formulation. It will be more convenient to express the conjecture in terms of binomial and multinomial type sums. We therefore introduce the following functions:

$$\gamma(m) := m^m \xi(m) = \sum_{k=0}^{m} \binom{m}{k} k^k (m-k)^{m-k} \quad (m \in \mathbb{N})$$

and

$$\gamma_2(m) := m^m \xi_2(m) = \sum_{j=0}^{m} \sum_{k=0}^{m-j} \binom{m}{j} \binom{m-j}{k} j^j k^k (m-j-k)^{m-j-k} \quad (m \in \mathbb{N}).$$

Note that Conjecture 2.1 is equivalent to the identity

$$\gamma_2(m) - \gamma(m) = m^{m+1} \quad (m \in \mathbb{N}).$$

It turns out that the functions $\gamma$ and $\gamma_2$ have considerably simpler expressions.

Lemma 3.1. For $m \in \mathbb{N}$, we have

$$\gamma(m) = \sum_{j=0}^{m} m^j \frac{m!}{j!}.$$
and
\[ \gamma_2(m) = \sum_{j=0}^{m} m^{m-j} \binom{m}{j} (j+1)! . \]

Assuming Lemma 3.1, we can now prove Conjecture 2.1.

Proof. The proof consists of an elementary calculation. We have
\[
\gamma_2(m) - \gamma(m) = \sum_{j=0}^{m} m^{m-j} \binom{m}{j} (j+1)! - \sum_{j=0}^{m} m^j \frac{m!}{j!} \\
= \sum_{k=0}^{m} m^k \binom{m}{m-k} (m-k+1)! - \sum_{j=0}^{m} m^j \frac{m!}{j!} \\
= \sum_{k=0}^{m} \frac{m^k m!}{k!} (m-k+1) - \sum_{j=0}^{m} m^j \frac{m!}{j!} \\
= \sum_{k=0}^{m} \frac{m^k m!}{k!} (m-k) \\
= m \sum_{k=0}^{m} \frac{m^k m!}{k!} - \sum_{k=0}^{m} \frac{km^k m!}{k!} \\
= \sum_{j=1}^{m+1} \frac{m^j m!}{(j-1)!} - \sum_{k=1}^{m} \frac{m^k m!}{(k-1)!} \\
= m^{m+1} 
\]
as required. □

3.2. Proof of Lemma 3.1. For the proof of Lemma 3.1 we need a binomial sum identity first proved by Abel as well as its generalization to the multinomial case by Hurwitz.

More precisely, for \( m \in \mathbb{N}, x, y \in \mathbb{R}, p, q \in \mathbb{Z} \), define
\[
A_m(x, y; p, q) := \sum_{k=0}^{m} \binom{m}{k} (x+k)^{k+p}(y+m-k)^{m-k+q}. 
\]

Note that the special case \( p = 0, q = -1, y \neq 0 \) corresponds to the classical Abel binomial theorem:
\[
A_m(x, y; 0, -1) = \frac{1}{y} (x + y + m)^m. 
\]

Moreover, we have
\[
A_m(0, 0; 0, 0) = \gamma(m) \quad (m \in \mathbb{N}). 
\]

The other function \( \gamma_2 \) can be expressed in terms of a multinomial version of \( A_m \). More precisely, for \( x_1, \ldots, x_n \in \mathbb{R} \) and \( p_1, \ldots, p_n \in \mathbb{Z} \), define
\[
B_m(x_1, \ldots, x_n; p_1, \ldots, p_n) := \sum_{k_1! \cdots k_n!} \frac{m!}{k_1! \cdots k_n!} \prod_{j=1}^{n} (x_j + k_j)^{k_j+p_j}, 
\]
where the sum is taken over all non-negative integers $k_1, \ldots, k_n$ such that $k_1 + \cdots + k_n = m$.

A simple calculation then shows that

\begin{equation}
B_m(0, 0, 0; 0, 0, 0) = \gamma_2(m) \quad (m \in \mathbb{N}).
\end{equation}

We can now proceed with the proof of Lemma 3.1.

**Proof.** In [21, p.21], we find the following identity:

\[ A_m(x, y; 0, 0) = \sum_{k=0}^{m} \binom{m}{k} k! (x + y + m)^{m-k} \]

Setting $x = y = 0$ gives

\[ \gamma(m) = A_m(0, 0; 0, 0) = \sum_{k=0}^{m} \binom{m}{k} m^{m-k} = \sum_{j=0}^{m} \frac{m!}{j!} m^j, \]

where we used Equation (1). This proves the first equality.

For the second equality, we use [21, Equation (35), p.25]:

\[ B_m(x_1, \ldots, x_n; 0, \ldots, 0) = \sum_{k=0}^{m} \binom{m}{k} (x_1 + \cdots + x_n + m)^{m-k} \alpha_k(n-1) \]

where

\[ \alpha_k(r) := \frac{(r + k - 1)!}{(r-1)!}. \]

Note that if $n = 3$, then $\alpha_k(n-1) = (k+1)!$. Setting $x_1 = x_2 = x_3 = 0$ then gives

\[ \gamma_2(m) = B_m(0, 0, 0; 0, 0, 0) = \sum_{k=0}^{m} \binom{m}{k} m^{m-k}(k+1)!, \]

where we used Equation (2). This completes the proof of the lemma. \Box

3.3. **Conclusion.** We now summarize what we have proved in the following theorem.

**Theorem 3.2.** For $m \in \mathbb{N}$, define

\[ \xi(m) := \sum_{k=0}^{m} \binom{m}{k} \left( \frac{k}{m} \right)^k \left( 1 - \frac{k}{m} \right)^{m-k} \]

and

\[ \xi_2(m) := \sum_{j=0}^{m} \sum_{k=0}^{m-j} \binom{m}{j} \binom{m-j}{k} \left( \frac{j}{m} \right)^j \left( \frac{k}{m} \right)^k \left( 1 - \frac{j}{m} - \frac{k}{m} \right)^{m-j-k}. \]

Then we have

\[ \xi(m) = \frac{1}{m^m} \sum_{j=0}^{m} j! \left( \frac{m}{j} \right)^j (m^j + 1) \quad (m \in \mathbb{N}) \]

and

\[ \xi_2(m) = \frac{1}{m^m} \sum_{j=0}^{m} m^{m-j} \left( \frac{m}{j} \right)^j (j+1)! \quad (m \in \mathbb{N}). \]
Furthermore,

\[ \xi_2(m) = \xi(m) + m \quad (m \in \mathbb{N}). \]

**Remark.** Theorem 3.2 not only proves Conjecture 2.1 but also gives simpler expressions for the functions \( \xi \) and \( \xi_2 \) that are more convenient from a numerical perspective. As discussed in Section 1, this might be of interest for the computation of PAC-Bayes bounds in machine learning.

**References**

1. L. Breiman, Bagging predictors, Mach. Learn., 24 (1996), 123–140.
2. W.Y.C. Chen, J.F.F. Peng, H.R.L. Yang, Decomposition of triply rooted trees, Electron. J. Combin., 20 (2013), 10pp.
3. E.B. Fox, D.B. Dunson, Bayesian Nonparametric Covariance Regression, Journal of Machine Learning Research, 16 (2015), 2501–2542.
4. Y. Freund and R.E. Schapire, A decision-theoretic generalization of on-line learning and application to boosting, In Proceedings of the Second European Conference on Computational Learning Theory, EuroCOLT 95, 1995, 23–37.
5. I.M. Gessel, Lagrange inversion, J. Combin. Theory Ser. A, 144 (2016), 212–249.
6. P. Germain, Généralisations de la théorie PAC-bayésienne pour l’apprentissage inductif, l’apprentissage transductif et l’adaptation de domaine, Ph.D. Thesis, Université Laval, 2015.
7. P. Germain, A. Lacasse, F. Laviolette, M. Marchand, PAC-Bayesian Learning of Linear Classifiers, In Proceedings of the 26th Annual International Conference on Machine Learning, ICML 09, 2009, 353–360.
8. P. Germain, A. Lacasse, F. Laviolette, M. Marchand, J.-F. Roy, Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm, Journal of Machine Learning Research, 16 (2015), 3501–3542.
9. A. Lacasse, Bornes PAC-Bayes et algorithmes d’apprentissage, Ph.D. Thesis, Université Laval, 2010.
10. A. Lacasse, F. Laviolette, M. Marchand, P. Germain, N. Usunier, PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier, In Advances in Neural Information Processing Systems 19, 2009, 769–776.
11. J. Langford, Tutorial on Practical Prediction Theory for Classification, J. Mach. Learn. Res., 6 (2005), 273–306.
12. A. Maurer, A note on the PAC bayesian theorem, CoRR, 2004.
13. D.A. McAllester, Some pac-bayesian theorems, In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 98, 1998, 230–234.
14. D.A. McAllester, Some PAC-Bayesian theorems, Machine Learning, 37 (1999), 355–363.
15. D.A. McAllester, Pac-bayesian stochastic model selection, Mach. Learn., 51 (2003), 5–21.
16. L. Oneto, Model selection and error estimation without the agonizing pain, WIREs Data Mining and Knowledge Discovery, 8 (2018).
17. L. Oneto, PAC-Bayes Theory, Model Selection and Error Estimation in a Nutshell, (2019), 75–86.
18. L. Oneto, The Five W of MS and EE, Model Selection and Error Estimation in a Nutshell, (2019), 5–11.
19. L. Oneto, D. Anguita, S. Ridella, PAC-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis, Pattern Recognition Letters, 80 (2016), 200–207.
20. H. Prodinger, An identity conjectured by Lacasse via the tree function, Electron. J. Combin., 20 (2013), 3pp.
21. J. Riordan, Combinatorial Identities, Robert E. Krieger Publishing Co., New York, 1968.
22. D.M. Smith, G. Smith, Tight Bounds on Information Leakage from Repeated Independent Runs, 2017 IEEE 30th Computer Security Foundations Symposium (CSF).
23. Y. Sun, A simple proof of an identity of Lacasse, Electron. J. Combin., 20 (2013), 3pp.
24. M. Younsi, Proof of a combinatorial conjecture coming from the PAC-Bayesian machine learning theory, arXiv:1209.0824
Department of Mathematics, University of Hawaii Manoa, Honolulu, HI 96822, USA. 
E-mail address: malik.younsi@gmail.com

Coveo Solutions Inc., Québec, Québec G1W 2K7, Canada. 
E-mail address: alacasse@coveo.com