A Concentration Inequality for Multiple Means Based on a Polynomial-Time Computable Metric

Vinícius M. Ribeiro[1], **André L. Vignatti**[1]

[1]Department of Computer Science – Federal University of Paraná (UFPR)
Curitiba – PR – Brazil

{vmribeiro,vignatti}@inf.ufpr.br

Abstract. *This paper introduces a novel concentration inequality for k empirical means that is independent of k. Unlike previous approaches that rely on NP-hard measures such as VC dimension or Rademacher Averages, we introduce the metric K_a, which is computable in polynomial time. This metric is obtained directly from the sample and represents the number of estimators effectively updated during the sampling process. By adapting uniform convergence proofs from statistical learning theory, we generalize these results to empirical means of real-valued variables without resorting to worst-case sample metrics like VC dimension or pseudodimension. Our approach contributes to the development of more efficient randomized algorithms with robust theoretical guarantees.*

1. Introduction

A fundamental aspect of randomized algorithm design is the simultaneous estimation of k values while bounding the probability p that at least one estimate falls outside the desired confidence interval. Classical approaches like the union bound [Mitzenmacher and Upfal 2017] produce a sample complexity that scales with k, which is problematic when k is large. VC dimension [Vapnik and Chervonenkis 2015], pseudodimension [Pollard 2012], and Rademacher Averages [Koltchinskii and Panchenko 2000] are tools from statistical learning theory that provide a way to bound p independently of k, and have been successfully used in various data and graph analysis algorithms for very large datasets [Preti et al. 2023, Riondato and Vandin 2020, Pellegrina and Vandin 2023].

One of the drawbacks of VC dimension is that it is restricted to empirical means over Bernoulli random variables. Pseudodimension is a generalization of the VC dimension that extends the support to real-valued variables. However, both metrics are distribution-free and assume worst-case sample sizes. In contrast, Rademacher Averages provide a data-dependent framework that yields tighter error bounds and enables the use of progressive sampling for a dynamic stopping criterion [Pellegrina and Vandin 2023, Cousins and Riondato 2020]. Still, although these tools provide robust theoretical guarantees, the exact calculation of VC dimension and Rademacher Averages is, in general, an NP-hard problem [Linial et al. 1991, Bartlett and Mendelson 2002].

In this work, we adapt the theorems based on VC dimension to overcome the aforementioned limitations. We introduce a new metric, K_a, which represents the total number of estimators effectively updated in a sample. The resulting bound is applied to estimators based on empirical means of real variables within the range $[0, 1]$. For bounded variables outside this range, the bound remains applicable by normalizing the variables and rescaling the confidence interval accordingly.

Because the confidence interval depends on K_a, it is calculated empirically from the sample rather than assuming the worst case. In addition to covering the broad scenarios of previous measures, this method offers the significant advantage of being computable in deterministic polynomial time.

2. Definitions

Consider a scenario with k estimators where sampling a single element from the population updates multiple estimators simultaneously. Formally, each element v from the sample S can be represented by a vector of size k containing the values to be updated in the estimators. We assume that this update is sparse; only a fraction of the estimators is modified for each new sampled element. Thus, the vast majority of the components in v maintain a default constant a.

Let $k_a(v)$ be the number of components in v that are different from a. For example, assuming $a = 0$ and $v = (1, 0, 3, 5, 0, 0, 15)$, we have $k_0(v) = 4$. The total number of estimators effectively updated, K_a, for a sample $S = (v_1, v_2, \ldots, v_n)$ is given by:

$$K_a = \sum_{i=1}^{n} k_a(v_i) \tag{1}$$

Since updates are sparse, $k_a(v)$ is bounded by some $t \ll k$. This is the key feature we exploit to bypass the dependency on k and, consequently, outperform classical approaches. Computing K_a is also very fast, as it only takes $O(nt)$ time.

3. Result

Theorem 1. *Let $Y_1, Y_2, \ldots, Y_{2n}$ be i.i.d. random vectors $Y_j = (X_1^{(j)}, X_2^{(j)}, \ldots, X_k^{(j)})$, with $0 \leq X_i^{(j)} \leq 1$. Let $A = (Y_1, Y_2, \ldots, Y_n)$ and $B = (Y_{n+1}, Y_{n+2}, \ldots, Y_{2n})$. Define the empirical means $Z_i^A = \frac{1}{n}\sum_{j=1}^{n} X_i^{(j)}$ and $Z_i^B = \frac{1}{n}\sum_{j=n+1}^{2n} X_i^{(j)}$, with expectation $\mathbb{E}Z_i = \mathbb{E}[Z_i^A] = \mathbb{E}[Z_i^B]$. Then, for $\delta \in (0, 1/2)$ and $K_a \geq 1$,*

$$\Pr\left(\bigcup_{i=1}^{k}\left\{|Z_i^{\bullet} - \mathbb{E}Z_i| \geq \sqrt{\frac{8}{n}\ln\left(\frac{2K_a}{(1-2e^{-4})\,\delta}\right)}\right\}\right) \leq \delta,$$

where K_a is defined according to Equation 1 and computed from $Y_1, \ldots, Y_{2n}$.

Proof. Our approach is inspired by PAC learning proofs for uniform convergence using the growth function and VC dimension [Anthony and Bartlett 2009, Thm. 4.3], [Mitzenmacher and Upfal 2017, Thm. 14.15]. We focus on calculating the probability involving A, but the same reasoning applies to B. To simplify the notation, we define $\epsilon = \sqrt{\frac{8}{n}\ln\left(\frac{2K_a}{(1-2e^{-4})\delta}\right)}$. The first step is to apply the symmetrization technique, i.e., we remove the $\mathbb{E}Z_i$ dependency by bounding the probability of interest by a probability involving only the two groups:

$$\Pr\left(\bigcup_{i=1}^{k}\left\{|Z_i^A - \mathbb{E}Z_i| \geq \epsilon\right\}\right) \leq \frac{1}{1-2e^{-4}}\Pr\left(\bigcup_{i=1}^{k}\left\{|Z_i^A - Z_i^B| \geq \frac{\epsilon}{2}\right\}\right) \tag{2}$$

By the triangle inequality, if $|Z_i^A - \mathbb{E}Z_i| \geq \epsilon$ and $|Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}$, then $|Z_i^A - Z_i^B| \geq \frac{\epsilon}{2}$. Defining the event $R = \bigcup_{i=1}^k \left\{|Z_i^A - \mathbb{E}Z_i| \geq \epsilon\right\}$, we obtain

$$\begin{aligned}
\frac{\Pr\left(\bigcup_{i=1}^k \left\{|Z_i^A - Z_i^B| \geq \frac{\epsilon}{2}\right\}\right)}{\Pr\left(R\right)} &\geq \frac{\Pr\left(\bigcup_{i=1}^k \left\{|Z_i^A - \mathbb{E}Z_i| \geq \epsilon,\ |Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}\right\}\right)}{\Pr\left(R\right)} \\
&= \Pr\left(\bigcup_{i=1}^k \left\{|Z_i^A - \mathbb{E}Z_i| \geq \epsilon,\ |Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}\right\} \;\middle|\; R\right) \\
&\geq \min_{1 \leq i \leq k} \Pr\left(|Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}\right).
\end{aligned}$$

The last inequality follows from the fact that if R occurs, there exists a particular Z_r^A such that $|Z_r^A - \mathbb{E}Z_r| \geq \epsilon$, and thus $\Pr\left(\bigcup_{i=1}^k \left\{|Z_i^A - \mathbb{E}Z_i| \geq \epsilon, |Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}\right\} \middle| R\right) \geq \Pr\left(|Z_r^B - \mathbb{E}Z_r| \leq \frac{\epsilon}{2}\right) \geq \min_{1 \leq i \leq k} \Pr\left(|Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}\right)$. Considering that $\ln\left(\frac{2K_a}{(1-2e^{-4})\delta}\right) \geq 1$ and applying Hoeffding's inequality,

$$\begin{aligned}
\frac{\Pr\left(\left\{\bigcup_{i=1}^k |Z_i^A - Z_i^B| \geq \frac{\epsilon}{2}\right\}\right)}{\Pr\left(R\right)} &\geq \min_{1 \leq i \leq k} \Pr\left(|Z_i^B - \mathbb{E}Z_i| \leq \frac{\epsilon}{2}\right) \\
&= 1 - \max_{1 \leq i \leq k} \Pr\left(|Z_i^B - \mathbb{E}Z_i| > \frac{\epsilon}{2}\right) \\
&\geq 1 - \max_{1 \leq i \leq k} \Pr\left(|Z_i^B - \mathbb{E}Z_i| > \frac{1}{2}\sqrt{\frac{8}{n}}\right) \\
&\geq 1 - 2\exp\left(-2n\left(\frac{1}{2}\sqrt{\frac{8}{n}}\right)^2\right) \\
&= 1 - 2e^{-4}.
\end{aligned}$$

This concludes the proof of Inequality 2.

Now, we rewrite the probability in terms of permutations between A and B. Let Γ_{2n} be the set of all permutations σ of $\{1, \ldots, 2n\}$, where either the i-th element is swapped with the $(i+n)$-th element, that is, $\sigma(i) = i + n$ and $\sigma(i+n) = i$, or the elements are left fixed, with $\sigma(i) = i$ and $\sigma(i+n) = i+n$. Let f be a function that applies the permutation $\sigma \in \Gamma_{2n}$ to the $Y_1, Y_2, \ldots, Y_{2n}$ random vectors (e.g. $f((Y_1, Y_2, Y_3, Y_4), (3, 4, 1, 2)) = (Y_3, Y_4, Y_1, Y_2)$), let $Q = \bigcup_{i=1}^k \left\{|Z_i^A - Z_i^B| \geq \frac{\epsilon}{2}\right\}$, and let $\mathcal{D}$ be the distribution where each Y_j is draw from. For any σ in Γ_{2n}, $\Pr_{z\sim\mathcal{D}^{2n}}\left(z \in Q\right) = \Pr_{z\sim\mathcal{D}^{2n}}(f(z,\sigma) \in Q)$ since permutations of z preserve its original distribution. In particular,

$$\begin{aligned}
\Pr_{z\sim\mathcal{D}^{2n}}\left(z \in Q\right) &= \frac{1}{|\Gamma_{2n}|} \sum_{\sigma \in \Gamma_{2n}} \Pr_{z\sim\mathcal{D}^{2n}}(f(z,\sigma) \in Q) \\
&= \frac{1}{|\Gamma_{2n}|} \sum_{\sigma \in \Gamma_{2n}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}}[\mathbf{1}_Q(f(z,\sigma))] \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}}\left[\frac{1}{|\Gamma_{2n}|} \sum_{\sigma \in \Gamma_{2n}} \mathbf{1}_Q(f(z,\sigma))\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}} \left[\Pr_{\sigma\sim\Gamma_{2n}} \left(f(z,\sigma) \in Q \right) \right] \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}} \left[\Pr_{\sigma\sim\Gamma_{2n}} \left(\bigcup_{i=1}^{k} \left\{ \left| \frac{1}{n} \sum_{j=1}^{n} X_i^{(\sigma(j))} - \frac{1}{n} \sum_{j=n+1}^{2n} X_i^{(\sigma(j))} \right| \geq \frac{\epsilon}{2} \right\} \right) \right] \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}} \left[\Pr_{\sigma\sim\Gamma_{2n}} \left(\bigcup_{i=1}^{k} \left\{ \left| \frac{1}{n} \sum_{j=1}^{n} \left(X_i^{(\sigma(j))} - X_i^{(\sigma(j+n))} \right) \right| \geq \frac{\epsilon}{2} \right\} \right) \right],
\end{aligned}
$$

where $\mathbf{1}_Q(x)$ is an indicator function equal to 1 if $x \in Q$ and 0 otherwise. Note that K_a is not a random variable for permutations $\sigma \sim \Gamma_{2n}$, as changing the order of the vectors $Y_i, \cdots, Y_{2n}$ does not change the sum over all vectors. Therefore, $\epsilon/2$ is a constant relative to the probability defined over Γ_{2n}. If two sequences $X_t^{(1)}, \cdots, X_t^{(2n)}$ and $X_r^{(1)}, \cdots, X_r^{(2n)}$ are equal, then the events $E_t = \left\{ \left| \frac{1}{n} \sum_{j=1}^{n} \left(X_t^{(\sigma(j))} - X_t^{(\sigma(j+n))} \right) \right| \geq \frac{\epsilon}{2} \right\}$ and $E_r = \left\{ \left| \frac{1}{n} \sum_{j=1}^{n} \left(X_r^{(\sigma(j))} - X_r^{(\sigma(j+n))} \right) \right| \geq \frac{\epsilon}{2} \right\}$, defined over Γ_{2n}, are also equal. Thus, the union does not need to range over all k sequences, but only over the set of distinct sequences $\mathcal{S}^*$. Furthermore, $|\mathcal{S}^*| \leq K_a$, since each Y_j can induce at most $k_a(v)$ new (distinct) sequences. By the union bound,

$$
\begin{aligned}
&\Pr_{\sigma\sim\Gamma_{2n}} \left(\bigcup_{i=1}^{k} \left\{ \left| \frac{1}{n} \sum_{j=1}^{n} \left(X_i^{(\sigma(j))} - X_i^{(\sigma(j+n))} \right) \right| \geq \frac{\epsilon}{2} \right\} \right) \\
&\qquad\qquad \leq K_a \max_{s\in\mathcal{S}^*} \Pr_{\sigma\sim\Gamma_{2n}} \left(\left| \frac{1}{n} \sum_{j=1}^{n} \left(X_s^{(\sigma(j))} - X_s^{(\sigma(j+n))} \right) \right| \geq \frac{\epsilon}{2} \right).
\end{aligned}
$$

Given the distribution of the permutations σ, the difference $X_s^{(\sigma(j))} - X_s^{(\sigma(j+n))}$ takes on signs ± 1 with equal probability. Using Rademacher variables $\beta_j \sim \{-1, +1\}$,

$$
\Pr_{\sigma\sim\Gamma_{2n}} \left(\left| \frac{1}{n} \sum_{j=1}^{n} \left(X_s^{(\sigma(j))} - X_s^{(\sigma(j+n))} \right) \right| \geq \frac{\epsilon}{2} \right) = \Pr_{\beta\sim\{\pm1\}^n} \left(\left| \frac{1}{n} \sum_{j=1}^{n} \left| X_s^{(j)} - X_s^{(j+n)} \right| \beta_j \right| \geq \frac{\epsilon}{2} \right),
$$

where each β_j is chosen uniformly and independently. By Hoeffding's inequality, this probability is upper bounded by $2\exp\left(-\frac{n\epsilon^2}{8}\right)$. Thus,

$$
\begin{aligned}
\Pr\left(\bigcup_{i=1}^{k} \left\{ |Z_i^A - Z_i^B| \geq \frac{\epsilon}{2} \right\} \right) &\leq \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}} \left[K_a \cdot 2\exp\left(-\frac{n\epsilon^2}{8} \right) \right] \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}} \left[K_a \cdot 2\exp\left(-\ln\left(\frac{2K_a}{(1-2e^{-4})\delta} \right) \right) \right] \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{D}^{2n}} \left[\left(1 - 2e^{-4}\right)\delta \right] \\
&= \left(1 - 2e^{-4}\right)\delta.
\end{aligned}
$$

Combining this result with Equation 2 concludes the proof. □

The variable ϵ is presented here in a simplified manner. A strictly formal approach, treating it as a random variable in the context of the proof, is compatible with the results obtained and will be explored in future work.

References

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Cousins, C. and Riondato, M. (2020). Sharp uniform convergence bounds through empirical centralization. *Advances in Neural Information Processing Systems*, 33:15123–15132.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer.

Linial, N., Mansour, Y., and Rivest, R. L. (1991). Results on learnability and the vapnik-chervonenkis dimension. *Information and Computation*, 90(1):33–49.

Mitzenmacher, M. and Upfal, E. (2017). *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press.

Pellegrina, L. and Vandin, F. (2023). Silvan: estimating betweenness centralities with progressive sampling and non-uniform rademacher bounds. *ACM Transactions on Knowledge Discovery from Data*, 18(3):1–55.

Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

Preti, G., De Francisci Morales, G., and Riondato, M. (2023). Maniacs: Approximate mining of frequent subgraph patterns through sampling. *ACM Transactions on Intelligent Systems and Technology*, 14(3):1–29.

Riondato, M. and Vandin, F. (2020). Misosoup: Mining interesting subgroups with sampling and pseudodimension. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(5):1–31.

Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer.