Rejoinder: Making V.-C. bounds accurate

Léon Bottou

I am very grateful to my colleagues Olivier Catoni and Vladimir Vovk because their insightful comments add considerable value to my article.

Olivier elegantly points out out how similar conclusions can be achieved with a PAC-Bayesian approach. He convinced me to try filling my knowledge gap by reading parts of his excellent monograph [2]. The introductory material of [1] also provides a broad overview of the connections between PAC-Bayesian bounds and more traditional empirical process bounds. I find instructive to observe how the same fundamental phenomena can be discussed from a purely combinatorial viewpoint (as in my text) or from a purely probabilistic approach (as in Olivier's comment.)

Besides providing a beautiful connection between sample compression bounds and conformal prediction, Vladimir raises two issues that I should have discussed much more precisely in the first place. The first issue focuses on the level of data dependence for learning bounds. Four successive data dependence levels make the bounds potentially more accurate and also less useful for predicting the risk because they depend on quantities that have not been observed at the time of the prediction. Since combinatorial bounds belong to the last category ("*data super-dependence*"), they are not very useful to predict the expected risk. The second issue raises questions about the exact difference between the exchangeability assumption and the i.i.d. assumption. These two issues are in fact intimately connected.

De Finetti's theorem characterizes exchangeable sample distributions as *mixtures* of i.i.d. distributions. Such mixtures are usually not i.i.d. distributions themselves. Consider for instance a sample of k real numbers drawn from a equal mixture of normal distributions centered in two distinct points $x, y \in \mathbb{R}$. The expected sample mean is of course $(x+y)/2$. However, regardless of k, one half of the samples has an empirical mean close to x and the other half has an empirical mean close to y. We have exchangeability but the law of large numbers does not apply.

Such a situation is far from unrealistic. Every data collection campaign is in practice corrupted by uncontrolled variables that can be viewed as latent mixture

Léon Bottou
Microsoft Research, 641 Avenue of the Americas, New York, NY, e-mail: leon@bottou.org

variables. Despite this, the combinatorial error bounds accurately describe what can be observed when one splits the data into training set and testing set. One cannot expect these same bounds to predict the expected error because it is impossible to construct such a prediction without additional assumption (such as independence assumptions). This is why, in practice, gathering representative data consistently remains the hardest part of building a machine learning application.

Finally, I find instructive to question whether predicting the expected risk is the true purpose of learning bounds. Under i.i.d. assumptions, the most accurate "*inductively data-dependent*" way to estimate the expected risk almost always consists of holding out testing data. Held out data affords considerably better confidence intervals; they easily compensate what is lost by reducing the training set size. In fact, it is easy to see that one can match the best learning bounds by holding out a fraction of examples inversely proportional to $\log \mathrm{Card}\, \Omega_{\mathscr{A}}(S)$.

Let us nevertheless imagine a training set so small that we cannot afford to save a few testing examples, and let us also ignore the fact that the resulting learning system will probably perform too poorly to be of any use. Rather than using a learning bound, the practitioner would be wise to use a k-fold cross-validation approach and average the predictions of the k learning systems. Under the appropriate convexity conditions, this ensemble should perform at least as well as the average of the errors estimated on each fold.

Why then are we devoting considerable efforts to construct more accurate learning bounds? The history of our field provides an easy answer: building more accurate learning bounds forces us to describe new phenomena and acquire new insights. These insights are often useful to inspire and to characterize new learning algorithms. Consider for instance the under-dispersion of the error vectors (figure 7.2). If our data super-dependent learning bound cannot be accurate without taking this effect into account, we can expect that accurate risk bounds or efficient learning algorithms should somehow to take this phenomenon into account.

References

1. Audibert, J.Y., Bousquet, O.: PAC-Bayesian generic chaining. In: S. Thrun, L. Saul, B. Schölkopf (eds.) Advances in Neural Information Processing Systems 16, pp. 1125–1132. MIT Press (2004)
2. Catoni, O.: PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, *IMS Lecture Note Monograph Series*, vol. 56. Institute of Mathematical Statistics (2007)