Author Response – Paper ID 12191 (Bayes Consistency vs. $\mathcal{H}$-Consistency)

Reviewer #1

Thanks for all the suggestions (on broader impact, phrasing in Definition 7, and other improvements), and the additional references! We will incorporate all these suggestions and include these references in the final version if accepted.

Re. Footnote 5: The essence of the definition we have given (in Definition 7) is the same as that of Long and Servedio's definition, but technically, it is slightly more general in that we allow $\mathcal{F}$ to be any set of scoring function vectors $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$, while in the original definition $\mathcal{F}$ contained scoring function vectors $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$ whose component scoring functions $f_1, \ldots, f_n : \mathcal{X} \to \mathbb{R}$ all came from some fixed class of real-valued functions $\mathcal{F}_0 \subset \{f : \mathcal{X} \to \mathbb{R}\}$. We will modify the wording in Footnote 5 to clarify this.

Reviewer #2

Thanks for your comments – we are glad you enjoyed reading the paper!

Re. the realizable $\mathcal{H}$-consistency setting: We would like to clarify some aspects of this setting. (1) In general, we agree that $\mathcal{H}$-realizability can be a strong assumption on the distribution, and that universal Bayes consistency may be more desirable to achieve in practice. Nevertheless, we believe it is helpful to understand what is and what is not possible in the $\mathcal{H}$-realizable setting (part of the reason that there has been interest in this setting is that it is related to the classical PAC learning setting studied in computational learning theory). (2) It is true that there are other "all-in-one" approaches that are consistent for broader classes of distributions. Our goal here, however, is not to contrast one-vs-all algorithms with all-in-one algorithms; our main interest has been to examine the strengths of the various surrogate losses involved, and to emphasize that just because a particular surrogate loss does not produce expected results when minimized over a 'simple' scoring function class, it should not be immediately discarded or labeled ineffective, since it may be the case that it needs to be coupled with a different scoring function class in order to obtain the desired results. (3) Obtaining $\mathcal{H}$-consistent algorithms for general distributions (i.e. distributions that are not necessarily $\mathcal{H}$-realizable) is in general a computationally difficult problem. This is known from agnostic PAC learning theory; e.g. even finding an optimal (in 0-1 sense) *binary* linear classifier for non-linearly separable data is NP-hard. Note also that in the general non-$\mathcal{H}$-realizable/agnostic setting, $\mathcal{H}$-consistency becomes different from Bayes consistency.

Re. performance of hinge vs. logistic loss: For binary classification, the hinge loss comes with better regret transfer bounds (Bartlett et al., *JASA*, 2006), which could provide a partial explanation. It could be interesting to conduct a similar exploration in the multiclass case as well.

Reviewer #3

Re. experiments: Your question on how the new scoring class makes a difference is already answered in our experiments. In particular, in Figure 3 and Table 2, please compare the results for $\psi_{\text{OvA,log/hinge}}$ with $\mathcal{F}_{\text{lin}}$ and with $\mathcal{F}_{\text{spwlin}}$.

Re. clarity: We are sorry that you found the ordering hard to follow. We have assumed some familiarity with the overall concepts in the introduction. We will try to re-order things somewhat in the final version if accepted. Note that the union $\cup_{m=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^m$ simply allows the training sample to be of *any* size $m \geq 1$.

Re. additional comments: (1) $\mathcal{F}_{\text{spwlin}}$ is a class of vectors of non-linear scoring functions with shared parameters; $\mathcal{F}_{\text{lin}}$ is the class of vectors of linear scoring functions (see Figure 1, rows 2 and 3). There is no direct connection between them. (2) Lemma 1 is proved in the supplementary material (the proof is not hard, but to our knowledge the result is new). A similar result can be shown for more general function classes as well (see lines 222—228). (3) For non-$\mathcal{H}$-realizable distributions, achieving $\mathcal{H}$-consistency is generally NP-hard. Please see also our response to Reviewer #2 above.

Reviewer #5

Thanks for your comments – we are glad you liked the paper!

Re. notion of $\mathcal{H}$-consistency and relation to Long and Servedio's definition: Please see our response to Reviewer #1 above (comment on Footnote 5).

Re. other supervised learning settings such as multi-label learning: The question of consistency in general is certainly of interest in other supervised learning problems, and indeed there has been much work on understanding Bayes consistency for such problems in recent years (e.g. Duchi et al., ICML 2010; Gao & Zhou, COLT 2011; and many others in recent years). In all these cases, the target loss of interest is different from the 0-1 loss. For $\mathcal{H}$-consistency, when a distribution is $\mathcal{H}$-realizable (or simply realizable), it turns out the Bayes optimal model for all losses (that are zero on the diagonal and positive elsewhere) is the same as the Bayes optimal model for the 0-1 loss, and so one could in principle directly apply our results in such settings as well. But it could be interesting to consider other surrogate losses more commonly used for such problems, and other function classes $\mathcal{H}$ that may be more natural for them.