Discussion of "Random projection ensemble classification" by Cannings and Samworth

We congratulate the authors for their thoughtful article on high-dimensional classification. Supervised classification is a rather challenging task when the number of features p is comparable or much larger than the sample size n. In this article, the authors propose to apply an arbitrary base classifier based on random projections of the feature vectors and use a data-driven approach to aggregate these results. Specifically, they divide the random projections into non-overlapping blocks, select the projection that gives the smallest estimated test error, and aggregate the classifiers based on these selected random projections. By doing so, they show that the test error of their classifier can be controlled by terms that are independent of p (Theorem 5).

We note that Theorem 5 depends on Assumption (A.2), which requires the distribution function of the estimated test error of a random-projection based classifier to be close to the minimum estimated test error over all random projections. It implicitly assumes that the constants β_0, β and ρ involved in that condition are independent of p. When these constants do depend on p however, the upper bound for the test error of their classifier will be dependent of p. It will be helpful if the authors could elaborate more on this condition. For example, which values will β_0, β and ρ take if we use linear discriminant analysis, quadratic discriminant analysis or the k-nearest neighbor as the base classifiers.

The authors provided an upper bound on the test error of their proposed classifier. From a theoretical perspective, it might be more interesting to study the asymptotic distribution of the test error of their random projection ensemble classifier. Does the asymptotic distribution of the test error exists or not? If it exists, what are the conditions on n, p, B_1 and B_2? In practice, it is also more useful to study statistical inference of the test error of the proposed classifier, so that a researcher can use it to test hypothesis or construct confidence intervals.

Although the authors mainly focus on the problem of classification, their methods can be extended to handle other high-dimensional problems as well. Recently, there's increasing interest in developing individualized optimal treatment regimes to capture patient's heterogeneous response to treatment in causal inference. Zhao et al. (2012) showed that estimating the optimal treatment regimes is equivalent to a weighted classification problem. When the number of covariates is much larger than the sample size, classical methods such as Q-learning (Watkins and Dayan, 1992) and A-learning (Murphy, 2003) will fail. However, we can apply Q-learning

or A-learning based on a series of random projections of the covariate space, diving these random projections into non-overlapping blocks, and aggregate these estimated optimal treatment regimes on the selected projections that give the largest estimate of the value function (the mean response under a given treatment regime) within each block. Then, similar to Theorem 5, we can provide an upper bound on the difference of the value function under the optimal treatment regime and that under our estimated optimal treatment regime.

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). *Estimating individualized treatment rules using outcome weighted learning*, Journal of the American Statistical Association, 107, 1106-1118.

Watkins, C. J., & Dayan, P. (1992). *Q-learning*, Machine learning, 8, 279-292.

Murphy, S. A. (2003). *Optimal dynamic treatment regimes*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65, 331-355.