Supplementary material for Collaborative hyperparameter tuning

1. Supplement to Section 4

In Section 4 of the main paper, we present results on two benchmarks in terms of average ranking, since classification datasets may not be commensurable in terms of raw validation error. For the sake of completeness, we present here results in terms of average meta-test error. Meta-test error is defined slightly differently in our two experiments. We also present a PCA of our data in the MLP experiment.

1.1. A case study on AdaBoost

In this experiment, meta-test error is obtained by a 5-fold CV on the set of datasets. Figure 1 shows the average meta-test error as a function of the number of iterations. The curves are (obviously) similar in the beginning and at the end of the experiment, but between step 20 and 50, the speedup of reaching a given error level can be more than two-fold wrt. separate tuning, and more than three-fold wrt. random search.

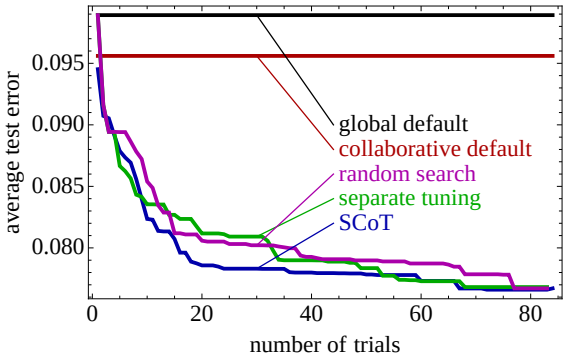

Figure 1. The average meta-test generalization error as a function of the number of trials. The curves are (obviously) similar in the beginning and at the end of the experiment, but in the middle of the experience, SCoT can reach a given average error twice as fast as separate tuning and three times as fast as random search.

1.2. A controlled experiment with MLPs

First, Figure 2 presents the PCA in $\mathbb{D}$ of the 20 datasets mentioned in Section 4.2.1, showing non-degeneracy but clustering, as expected. Second, unlike in the case study on AdaBoost, we did not perform meta-cross-validation, but rather acted as if we used SCoT to tune neural networks simultaneously on the 20 considered datasets. Methods that build models (collaborative default, separate tuning, and collaborative tuning in Section 4.1.2 of the main paper) started only after 10 random points had been evaluated on each dataset. The global default strategy was taken here to be the constant choice of the the best hyperparameters on average among these first 10 points.

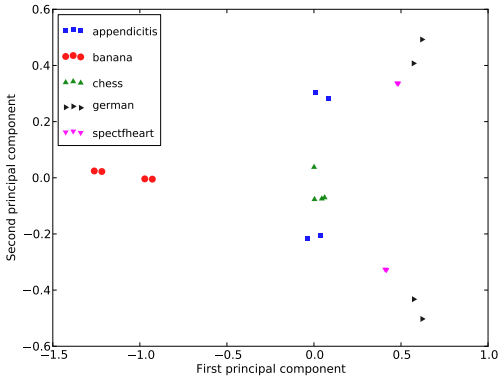

Figure 2. Projection of the 20 datasets used in the MLP experiment onto the first two principal components of the feature space. Similar markers are used to depict noisy versions of the same dataset, see Section 4.2.1 of the main paper.

Figure 3 depicts the results in terms of average obtained validation error. Soon after the initial 10 training points, SCoT clearly outperforms all other methods. Separate tuning comes second, but sharing information among problems obviously helps SMBO on this controlled benchmark.

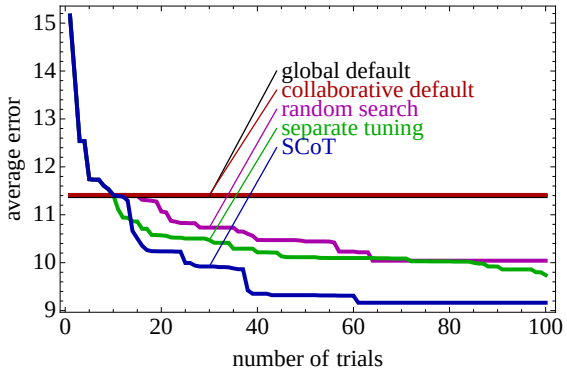

Figure 3. Results on the MLP benchmark in terms of the average validation error.