**Associate Editor’s Comments:**

*This paper is about tuning hyperparameters in regression, a paradigmatic example of a machine learning algorithm.  The proposal is potentially interesting, but the benchmarking is against a straw man only: grid search.  There is insufficient evidence that practitioners should be using this versus existing procedures.  Minimally, I would expect the authors to benchmark against "Practical Bayesian optimization of machine learning algorithms" by Snoek et. al. (NIPS 2012), which is a very popular technique both in industry and academia due to the availability of software.  Do we reach better solutions in less time?  What about a simple benchmark based on calls to Nelder-Mead or something similar?  Moreover, this paper is far from the only candidate.  Please also do a more thorough literature search of recent work on the topic of hyperparameter optimization, starting with the above-mentioned paper as a seed.  It is a robust area of research, and the absence of a reference to Snoek et al is, on its own, a red flag that insufficient homework has been done*

## Response:

We appreciate the valuable feedback on the manuscript and have substantially updated the paper, particularly the Abstract, Introduction (Section 1), Examples (Section 2.4), Simulation Studies (Section 3), and Discussion (Section 5) accordingly. We performed a more thorough literature review and found a number of papers that also tune penalty parameters using a gradient-based approach. These gradient-based methods however are only applicable for smooth loss, whereas our method can be applied to a number of popular methods with non-smooth penalties (e.g. Lasso, trend-filtering, sparse-additive models, among others). For the simulation studies with non-smooth penalties, we now benchmark against the three gradient-free methods mentioned by the editor: grid search, Nelder-Mead, and the Bayesian optimization technique by Snoek et. al. (2012). We find that our method is much more efficient in finding models with low validation (and generalization) error, in fewer iterations. This advantage becomes particularly pronounced when there are twenty or more penalty parameters to tune. To better illustrate this fact, we have replaced one of the previous example regression problems with a nonparametric additive model with smoothness and sparsity penalties.