Referee report on “Interpolation of High-Dimensional Data”

by Thomas C. H. Lux et al.

Given an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ with $d \gg 1$, and a set of scattered and noise-free evaluations of f, $(x^{(i)}, f(x^{(i)})$ for $i = 1, \ldots, M$, the manuscript under review focuses on approximating f by interpolation methods, as opposed to regression methods for the same task. Some discussion is also provided for vector-valued / functional problems, i.e. $f : \mathbb{R}^d \to \mathbb{R}^k$ or $f : \mathbb{R}^d \to C^0([a, b])$. The authors do not propose novel methods, but rather compare a handful of existing methods and propose a new error bound for multi-variate piecewise linear interpolation over d-dimensional simplexes.

While the article is well-written and the reading flows nicely, I unfortunately cannot recommend its publication in its present form, for the following reasons:

- The survey is missing some methods which are quite relevant in the high-dimensional approximation community:
  - regression based on radial basis functions (which could however be part of the support vector regressor method considered in the paper);
  - regression/interpolation based on sparse grids (for scattered and structured data, respectively; see e.g. H.-J. Bungartz and M. Griebel: Sparse grids, Acta Numerica, 2004, or D. Pflüger, B. Peherstorfer and H.-J. Bungartz: Spatially adaptive sparse grids for high-dimensional data-driven problems, Journal of Complexity, 2010);
  - interpolation over e.g. Padua points, Approximate Fekete points and Discrete Leja points (although one could argue that these are not “high-dimensional enough”, see e.g. L. Bos, S. De Marchi, M. Vianello: Trivariate Polynomial Approximation on Lissajous Curves, IMA Journal of Numerical Analysis, 2016).
- There is no numerical evidence about the error bound proved in Section 5. It is sharp or suboptimal? I would also expect some convergence plot, whereby one sees that the interpolation error gets smaller if the number of data points gets larger. The relation between number of points and error is not obvious from the error bound, which is over a single simplex (rather over the entire triangulation).
- The numerical tests focus only on “statistical datasets”. I would rather recommend that the authors first test the methods under investigation on interpolating functions with an analytic expression, like the suit of test functions proposed in Volker Barthelmann, Erich Novak and Klaus Ritter: High dimensional polynomial interpolation on sparse grids, Advances in Computational Mathematics, 2000.
- while I appreciated using the Kolmogorov-Smirnov (KS) distance and the corresponding KS test to measure errors in the “function-valued interpolation setting”, these are only valid if the function to be approximated is CDF (more precisely, the KS measure is simply the L^∞ norm so can be applied to the difference between any two functions, while the KS test only applies for CDFs). In case the function to be approximated is not a CDF, alternatives should be discussed. For instance, cross-validation errors could be computed by L^∞ or L^2 norm and then compactly visualized by box-plots (right now the box-plots visualize the KS statistic, not the KS measure)
- More on the KS distance and test: I assume that most of the readers of Numerical Algorithms have a strong numerical analysis background but probably a weaker backgound in statistics, so terms like null hypothesis, p-value, and box-plots cannot be given for granted. For the same reason, most of the discussion that right now is in the captions of Figures should be moved to the main text and further elaborated.