View Reviews

Paper ID
166

Paper Title
Exact expressions for double descent and implicit regularization via surrogate random design

Reviewer #2

Questions

1. Rating (scale of 1-10)
Marginally below acceptance threshold

2. Review
SUMMARY
The paper investigates closed-form expressions for the Moore-Penrose estimator applied to well-specified unregularized linear regression, in the case of a surrogate (randomized determinant) design matrix proposed by the authors.

The authors establish exact non-asymptotic expressions for the mean squared error (MSE) of their estimator, which can be used to characterize its double descent behavior as a function of the ratio between the number of parameters and the number of training samples (Theorem 1). The authors relate the expected value of their estimator to the solution of ridge regression at the population level when the penalty term is tuned in such a way that the number of training samples cohincides with the classical notion of effective dimension in ridge regression (Theorem 3). The authors present asymptotic consistency results for Gaussian data that bounds the deviation of the surrogate-design MSE to standard-design MSE with a multiplicative factor that scales like O(1/d) (Theorem 4), conjecturing that the restriction to isotropic covariance matrices in the under-determined result can be relaxed (Conjectures 15 and 16). To derive these results, the authors introduce the notion of determinant preserving random matrices and perform various intermediate computations that can be of independent interest.

QUALITY
The paper is well written, and I could not find any typo in the main text.

SIGNIFICANCE
Apart from Theorem 4 (on the asymptotic consistency of surrogate design in the case of Gaussian distributions), all theoretical results established by the authors relate to the surrogate design matrix case. This is mainly a fictitious device used to obtain closed-form expressions and at times, especially (but not only) in the abstract and introduction, the authors do not seem to stress this point strongly enough. For instance, phrases like "This provides a precise characterization of the double descent phenomenon for perhaps the simplest and most ubiquitous regression problem" (Introduction) and "We observe that our theory aligns well with the empirical estimates, whereas previously, no such theory was available except for the special case [...]" (Section 1.1.) are misleading. On the one hand, most of the results obtained by the authors do not pertain to "the simplest and most ubiquitous regression problem", but they only pertain to a regression problem with a (rather fictitious, surely not ubiquitously used) surrogate random design. On the other hand, as the non-asymptotic theory developed by the authors only refers to the surrogate design, it is

not fair to claim that "[...] previously, no such theory was available [...]", as the previous theory was developed for the standard i.i.d. design. While the authors present asymptotic consistency results (Theorem 4) and numerical experiments (Figure 1) to describe the discrepancy of the surrogate theory they developed with respect to the real i.i.d. design matrix case, the authors do not seem to be always straightforward about this discrepancy.

As most of the papers on interpolation methods, also this paper does not seem to address the question of *why* someone would like to use an interpolating method, to begin with? In simple linear regression models, the double-descent phenomenon is known to disappear completely for optimally-tuned estimators (e.g. see the 2019 paper by Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”). As one of the key points of this submission is to derive exact non-asymptotic expression for a modified random design, connecting the properties of this estimator to properly-tuned ridge regression, I believe the authors were in a position to comment more explicitly (in the case of surrogate design) on what are the advantages/disadvantages of using non-regularized estimators that displays the double-descent phenomenon versus using an optimally-tuned ridge regression estimator that does not display the double-descent phenomenon. I quote from the 2019 paper "Harmless interpolation of noisy data in regression" by Muthukumar et al.: "While it is intellectually interesting that interpolation need not be harmful, it is practically always suboptimal to regularization with denoising." Can this remark be quantified exactly in the setting of surrogate design?

Other minor remarks:
- The authors seem to use the word "Homoscedastic noise" (Assumption 2) to indicate something other than just homoscedasticity. Homoscedasticity refers to the fact that the variance of the noise term does not depend on the feature vectors. However, in Assumption 2 the authors assume a special type of homoscedastic model, i.e. Gaussian noise, and Gaussianity seems to play a crucial role in many of the results developed by the authors. Can the authors clarify when/whether Gaussianity is required to achieve their results?

- In Figure 1, the captions, axis labels, and plot legend do not make it clear what refers to the surrogate design (the lines) and what refer to the standard i.i.d. design (the points). I would recommend that the authors clarify this distinction, as they do in the main text.

- When commenting on the findings of Figure 1, right before stating Theorem 4, the authors write "[...] particularly when μ is a multivariate Gaussian". This does not seem correct, as Figure 1 *only* concerns results on multivariate Gaussians, not *particularly*.

CONCLUSION
As the gap between the surrogate theory developed by the authors and the "real-world" theory used in practice is not properly addressed by non-asymptotic methods (note that the authors do not provide any discussion/results on the sharpness of the multiplicative weight O(1/d) in Theorem 4), the scope and significance of this paper is not as wide as the authors seem to claim, at least from the point of view of double-descent and implicit regularization. I believe the contribution of this paper lies more on the development of novel tools to investigate random designs, rather than in the non-asymptotic analysis of the double descent phenomenon and implicit regularization for the classical design.

Reviewer #3

Questions

1. Rating (scale of 1-10)
OK but not good enough; reject

2. Review
The paper analyzes the minimum norm interpolating predictor in squared loss linear regression. This is done by defining a surrogate problem, providing an exact analysis of the surrogate problem, and showing that at least in some cases the surrogate problem asymptotically captured the behavior of the true regression problem. The main concrete result regarding the actual regression problem of interest is an asymptotic expression for the mean squared error in the case of a well specified isotropic (ie iid features) regression model. The paper also provides an analysis of the surrogate model, deriving an expression for the mean squared parameter error (distance of learned predictor to population optimum), and the "bias" (the expectation of the learned predictor), and relating them in an interesting way to ridge regression---however these results are not rigorously related to the actual regression problem.

The paper studies a timely and important question where analysis is still lacking and welcome. It provides a new, fresh, and possibly useful perspective. It is overall well written and nicely guides the reader through the setup, issues and results, though it is perhaps over-promising and could be written more humbly (see some specific comments and suggestions below). The math is clean and crisp.

However, in terms of actual bottom-line results, the paper comes out rather thin. Most results are stated not for the actual regression model of interest, but for a made-up "surrogate" model involving a random number of dependent observations from a modified distribution. The specifics of this "surrogate" model are (quite rightfully) not even given up front when the results are stated and discussed but rather deferred to the later Section of the paper---this emphasizes that understanding the specific behavior of the surrogate model is not interesting on its own right, but only in so far as it helps us understand the true regression model. This is indeed done for the MSE analysis of the isotropic model, but not in any other case, leaving us with only a rather speculative prediction for the behavior of the regression model.

Since the analysis of the MSE in the isotropic case is not new, as far as I understand (see below), it seems the paper should be evaluated as a speculative paper providing a heuristic prediction and conjecturing that it is asymptotically correct. I would have preferred if the paper was more explicitly written this way. In particular, I would have liked to see much more of a discussion and arguments, both analytic and empirical, for why this prediction might be accurate.

Figure 1 does provide some evidence toward this, but falls short of a comprehensive empirical evaluation of the prediction.

In this regard, I also find the title of the paper misleading: the result here is not an exact expression VIA surrogate design, but rather an exact expression FOR the surrogate design, and more importantly it is "(a prediction for) an asymptotic expression for double descent and implicit regularization". Contrasting to prior work by saying prior work is only asymptotic is also unfair because of this.

Indeed, the paper is lacking in direct comparisons to prior work. The authors mention several recent analysis of the minimum norm interpolating predictor (first paragraph of Section 2 on pages 6--7), but does not actually compare and contrast to them. In particular:

- the isotropic case, where you do provide a rigorous asymptotic analysis, has been similarly analyzed e.g. by Hastie et al 2019. In the isotopic case the parameter error (which you analyze) is also the same as the prediction error, and so there is no different there. You should directly compare your asymptotic result to theirs, verify that they agree, and highlight any ways in which they differ.

- Hastie et al also provide some sort of analysis for the non-isotropic case. Despite differences in the way the result is stated, and on bounded the parameter vs prediction errors, it would still be appropriate to compare and contrast.

- Are there any other predictions made in any of these papers? How do your predictions compare? How do your predictions align with prior empirical results?

A difference between the analysis in this submission and some prior analysis is in the quantities being analyzed, where I found the quantities being analyzed here somewhat less relevant (though still interesting) for interpolation learning:

- The submission analyzes the parameter error, as opposed to the predicting error, which is arguably more relevant for interpolation learning. It seems the authors try to present this as an advantage, and this is of course a matter of perspective. But is this an inherent limitation of the approach? Can you also analyze the prediction error of the surrogate model? Do you also expect it to asymptotically capture the behavior in the true regression model for non-isotropic and perhaps non-well-specified models?

- Theorem 3 in the submission looks at an interesting quantity: the asymptotic bias as captured by the expectation of the minimum norm predictor (in the surrogate model---this is not connected to the true model). The norm of this expectation is lower than that of the population optimal predictor, as shown in Figure 1b. But its useful to contrast this with the norm of the actual minimum norm predictor, which is larger that of the population optimum, as has been studied before.

Overall, I am torn about the paper. It seems there is a new and potentially useful approach here, but the paper is not satisfying in the ways discussed above.

Reviewer #4

Questions

1. Rating (scale of 1-10)
Marginally above acceptance threshold

2. Review
The paper derives an analytic expression for the excess risk of the pseudo-inverse estimator, which is non-vacuous even for n < d (overparametrized regime). In particular, this expression takes different form for underparametrized (n <= d), matching (n = d), and overparametrized regimes, and demonstrates the double descent behaviour observed experimentally in OLS and other learning models. Analytic expressions for the excess risk of OLS have been explored before, however, many of them were developed for so called weak feature models, where one observes only a subset of features generated by the statistical model (such as Belkin et al. 2019c). In the model considered in this paper, one observes all the features. As the paper mentions, indeed, a number of works have considered exactly the same setting, such as Hastie et al. 2019 and Bartlett et al. 2019. Unfortunately, the paper doesn't compare obtained results to these earlier works. Therefore, it's difficult to argue about the contribution in terms of the final identities.
On the other hand, the proof seems to be rather ingenious and tools developed deserve attention on their own. In particular, instead of showing an excess risk bound for the pseudo-inverse of a data matrix, the proof works with the surrogate data matrix sampled from a carefully constructed (rescaled) version of the marginal density. For such construct density we have interesting identities, which do not hold for the original matrix (e.g. most notably E[det(X^T X)] = det(E[X^T X])). These identities are the core of the proof, while eventually the original excess risk is related to the excess risk of the pseudo-inverse of the surrogate.
Finally, an analytic form of the excess risk shows clear dependence on the effective ridge dimension (commonly encountered in the analysis of the ridge regression), with the analogue of the ridge coefficient being an implicit distribution-dependent quantity. This demonstrates the effect of the implicit regularization in the pseudo-inverse (which is not surprising since we have a minimum norm solution).
I'm inclined towards accepting the paper -- it is very well written, clear, results seems interesting, and tools are very interesting. Sadly, the paper doesn't position itself critically among the recent literature (that is not just mentioning, but actually comparing). Even though, some of these might derive bounds, it would be interesting to understand how much do we gain through these new analytic results.