# Learning Curves for SGD on Structured Features

immediate

August 5, 2021

## 1 Reviewer 1 (aars): Rating 6

### Main Concerns and Responses

1. The paper assumes a random feature model but random embedding is never used.
   Response: We study linear regression from arbitrary feature spaces. The random feature setting and wide neural networks are special cases which induce linearized models with certain feature covariance. We believe our theory used for these specific models to be of interest to understand how neural networks behave on real data in special limits.

2. Line 100 uses irreducible component while the next section in line 102 does not have an irreducible component.
   Response: We thank the reviewer for this observation. We will include the irreducible component below equation 5 and in equation 6.

3. Noisy Quadratic Model Zhang et al (2019) is another model used for understanding training dynamics in neural networks. Does the paper provide any insights beyond what is provided by NQM?
   Response: We thank the reviewer for bringing our attention to this paper which we will discuss in our paper. While the noisy quadratic model considers a similar setting as ours (as well as momentum and moving average schemes), the authors of that study assumed that the gradient covariance and the loss Hessian are codiagonalizable, which allows them to decouple their test loss dynamics across eigenspaces. We show in this paper that this is not possible in general and that the evolution of the loss across different eigenspaces are coupled through the $\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$ term in $\boldsymbol{A}$. We also solve for the expected test error for general features in terms of fourth moment structure. While the NQM paper found that optimal learning rates can exist, we identified that optimal batch sizes can exist as well which are problem dependent.

4. In figure 3, panels (h) and (i) suggest there is a bias between the empirical and predicted behavior. Would you mind commenting on that?
   Response: The log scale on the y-axis is amplifying the visual disparity between the theoretical average and the. Indeed, by Jensen's inequality, we have $\langle \log L_t \rangle \leq \log \langle L_t \rangle$. Visually, one perceives $\langle \log L_t \rangle$ by visually averaging over the fluctuations in the experimental curves. The theory curve, on the other hand shows $\log \langle L_t \rangle$. By plotting on a linear scale, we see that the two curves visually agree. I also averaged over a greater number of SGD trajectories and got even closer agreement.

5. It was unclear to me how the empirical distributions in Figure 4 have been fitted and why test loss scaling predictions have only been computed for a subset of the time steps. Would you mind

clarifying that?
Response: To compute the empirical distributions, I took 6000 random MNIST digits and computed their respective features. From the features $\boldsymbol{\psi}(\mathbf{x}^\mu)$ I computed the kernel gram matrix $\boldsymbol{K}$ with entries $K_{\mu,\nu} = K(\mathbf{x}^\mu, \mathbf{x}^\nu)$. Diagonlizing this matrix $\boldsymbol{K} = \sum_k \lambda_k \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top$, we can identify the eigenvalues $\lambda_k$ directly and the target power spectrum through projection onto the labels $v_k^2 \sim \left(\boldsymbol{\phi}_k^\top \boldsymbol{y}\right)^2$ where $\boldsymbol{y}$ are the labels.

## 2 Reviewer 2 (ht7v): Rating 6

- I am not very familiar with the literature on these methods, so it's hard for me to assess the originality. However, since the results consist of analyzing a regression model under certain Gaussian and then non-Gaussian assumptions on the regressors, it would not be surprising if a similar analysis already exists in the literature, albeit in a different context.
  Response: This problem has certainly been studied in prior work, although an exact non-asymptotic analysis of the average loss as a function of time has, to our knowledge, not been calculated. We provide exact expressions for arbitrary features in terms of second and fourth moment structure and also solve the Gaussian case exactly in terms of second moments. There are many works which study the asymptotic scalings of these algorithms, often using decaying learning rate or

- Section 2.3.2 is not very clear; it's hard to tell what the main points are. (This is in stark contrast to the rest of the paper.) Since it seems like the most practically relevant, the authors would do well to revise it for clarity and highlight the main points.

- The analysis ultimately amounts to an analysis of a linear regression model under squared-error loss. The connection to machine learning comes when the regressors are assumed to be feature maps from some generic machine learning method. The new/interesting/not-well-understood things are the feature maps! So it's not clear how significant/relevant these results are to current methods. If I'm missing that connection, the authors should make the connection clearer.

- The assumptions that is either learnable (or decomposes as $y = \boldsymbol{w}^* \boldsymbol{\psi} + y_\perp$) are very strong. I understand that they make the problem much easier to analyze, but seem quite unrealistic; they amount to already being Bayes optimal, up to a rescaling. All of the difficulty of the learning problem is already done at that point!
  Response: This is actually completely general.

## 3 Reviewer 3 (rpPh)

- However I have a strong concern about novelty. The actual learning problem under consideration is SGD on a linear model with a squared loss. By "linear", I mean that the prediction function is linear in the optimizing variable. As such, the derivation of the SGD learning trajectory is a simple exercise. Indeed all theoretical predictions here are exact (e.g. there is no asymptotic limit involved); this should give a hint at how easy the problem actually is. The nice prediction against real life data is therefore unsurprising.

- (To clarify, most significant works in the recent literature either study the learning trajectories in the nonlinear regime, or establish why the learning trajectory that is supposedly nonlinear becomes linearized and its implications, or study interesting unknown statistical properties that even a linear model may exhibit. Once linearization is assumed, with a squared loss, writing down the trajectory solution is not a difficult problem.)

- Another concern is the assumption that $y = w^* \cdot \boldsymbol{\psi} + y_\perp$ and that $y_\perp$ is independent of $\boldsymbol{\psi}$ in Section 2.5. This is a strong assumption.

- How is $w^*$ extracted from real data?

- Add a plot of another method; classification error during training.

## 4 Reviewer 4 (h3NJ)