* Top Level: Investigate why deep learning learns so well despite heavy parameterization (over fitting). Rank Bias + Spectral Bias to help better define Implicit Regularization
* For the ZerO Initialization Paper:
  1. Does identity mapping always imply uniform weighting?
     + Uniform dimensionality across each layer, so its easier to apply identity mapping
  2. If you could summarize, how does GLRL “help to explain the excellent generalization in gradient-based deep learning” (page 2)?
     + Has to do with Rank Bias
  3. What does the authors mean when they say psi is an “element-wise nonlinearity”? Is this definition common terminology?
     + Scatter-to-scatter function (non-linear by randomness)
  4. How does applying z\_l = psi(x\_l) + x\_{l-1} to certain nonlinearities result in W\_l = 0? Also, just to check, is it correct that when W\_l = 0, no signal is generated “from the residual branches” (page 4)?
     + Because W\_l = 0, the signal form nonlinearity branches goes to 0 so the equation becomes linear function.
  5. Why does the identity initialization after Definition 1 “restrict maximum network expressivity” (page 4)? How is maximum network expressivity normally defined?
     + Given 0 and 1, if initializing using standard identity to parallel output limiting what network can express. Expressivity is how many functions NN can express. Infinite layers, can express any function. X and Y cap how we express any function.
  6. Could you explain Figure 5?
     + A standard approach when trying to move batch normalization. Trying to remove batch normalization through batch statistic. Domain knowledge required to understand it… Want to learn without batch statistic
  7. At the end of page 8, when Figure 4 is referenced in the first paragraph, are the authors referencing the leftmost figure? Is the second paragraph essentially describing the center-left figure?
     + Yes
  8. At the end of page 8, in the first paragraph do the authors mean that after applying the partial identity initialization \hat{K}, skip connections were added with Hadamard transforms for channels that were initialized to zero?
  9. Are Kaiming and Xavier initialization methods commonly used and/or leading standards?
  10. What do the authors mean when they write, “the degradation is induced by the differences in architecture instead of ZerO initialization itself” (page 9)?
  11. In the first paragraph of Section 5, could you please explain what aspect of GD implicitly biases the model towards low complexity solutions?
  12. While I understand the appendices are still a work in progress, I still had a few questions about it:
      + Ongoing draft, see new paper draft
      + Could you please explain what Lemma 1 means and how it connects to satisfying the rank constraint in Theorem 3?
      + In Theorem 4, could you please explain how the derivative of each layer is a sum of rank-1 matrices?
      + What do each of the six subplots in Figure 8 represent (page 11)?
* For the Fourier Neural Operator Paper:
  1. Does the fact that “classical neural networks map between finite-dimensional spaces and can therefore only learn solutions tied to a specific discretization” (page 1) imply classical NN have the same issue as traditional solvers/methods?
  2. What do the authors mean by “let G^{\dagger}: A \to U be a (typically) non-linear map” (page 3)? Do they mean G^{\dagger} can be a linear map?
  3. What do the authors mean by “probability measure \mu supported on A” (page 3)? Do they mean that for probability measure \mu there is a nonzero probability density everywhere in the space?
     + Yes
  4. How do all four layers described in Figure 2b differ from one another if they are all based on equations (2) and (4) (pages 4 and 6)?
  5. I found the results for the Bayesian Inverse Problem and the spectral analysis sections (page 9) interesting as they were compared with traditional methods, which I know better.
* For the Spectral Bias Paper:
  1. How exactly is their view on implicit regularization “slightly shifted” (page 1)? Is the shift just that they specify what type of minima they consider?
     + Previous works were focusing on parameter space (norm bias (doesn’t really characterize implicit bias) -> rank bias (better characterization but still parameter space)). Spectral bias (input space instead).
     + No connection has been explicitly found. Previously no model has used fourier transform and doing linear transform over fourier space.
     + Strict principle in PDE: Low frequency components always have largest amplitudes/magnitudes.
       - Corresponds directly to linear mapping structure. Input space is ordered by significant component -> first column of mapped matrix will be similarly corresponded where each column is singular values. Thus, we can show rank bias follows similarly (and generalizes rank bias)
  2. Could you explain what they mean in footnote 3 by “In such cases, the Fourier transform is understood in the sense of tempered distributions acting on rapidly decaying smooth functions \phi as <\tilde{f}, \phi> = <f, \tilde{\phi}>?
     + Dimensionalities extensions (unsure either)
  3. While I have looked at Appendix C.2, could you still explain Lemma 2 (page 2)? In particular what they mean by “some n-codimensional face of P, D\_n: R^d \to C is in \Theta(1)(k \to \infty)” (page 2)?
  4. While I know a spectral norm what is, I was wondering why was spectral norm used in this case over the Frobenius norm in equation (7) on page 2?
  5. How do the plots in Figure 1 explain the gist that “the model prioritizes learning lower frequencies first” and “the spectral norm of weights increases as the model fits higher frequency” (page 4)?
  6. For Figure 2 for Experiment 1, is it normal for training to take 80000 iterations even with full-batch gradient descent (page 4)?
  7. Could you explain how “Figure 4b shows that the amplitude of high-frequency noise does not significantly affect the best validation score” (pages 4-5)? I could not really see how they draw this conclusion…
  8. I think I am just generally confused by the manifold section (Section 4).
     + What exactly is implied by Figure 7 (page 6)?
     + What do they mean by “a given signal defined on the manifold is easier to fit when the coordinate functions of the manifold embedding itself has high frequency components” (page 9)?
       - First claim is saying that behavior of NN no matter what the function, the NN will learn low frequency first (just behavior, no claim on speed)
       - Second claim is if first claim is true, and the target function has high frequency components, learns faster than just all low frequency manifold
         * If no high frequency component, then manifold will be low dimensional
         * If non-linearity applied to a concentrated low dimensional space, harder to spread than varied dimensionality
  9. Could you explain why they switch the vol definition from the Hausdorff measure in Appendix C.2.1 (page 18) to the Lebesgue measure in Appendix D (page 21)?
  10. Could you explain why “the smoothness of a KNN prediction function is not well studied” (page 22)?