Minimal Projection Constants and Neural Networks

Ryan Malthaner, Weston Baines

Introducing Minimal Projection Constants

Definition

Given a Banach space $(X, ||\cdot||_X)$ and a closed linear subspace $Y \subset X$, an operator $P : X \to Y$ is called a *projection* if for every $y \in Y$, $P(y) = y$.

Using the operator norm on a given projection P i.e.

$$||P|| = \sup_{x \in X/0} \frac{||Px||_X}{||x||_X},$$

we can consider for the set of all projections $P : X \to Y$, $\min_P ||P|| = \lambda(Y)$ which is called *Minimal Projection Constant* for the subspace Y.

Relevant Background: Blatter and Cheney (1974)

- In general, minimal projection constants and formulas for them are exceedingly difficult to compute, however, certain cases have reasonable formulas i.e. trigonometric polynomials.
- Let $(f_1, f_2, ..., f_n) \in \mathbb{R}^n$ such that $||f||_1 = 1$.
- With respect to the ℓ_∞ (max) norm, we have for the hyperplane orthogonal to f,

$$\lambda(f^\perp) = \begin{cases} 1 & ||f||_\infty \geq 1/2 \\ 1 + \dfrac{1}{\sum_{i=1}^{n} \dfrac{|f_i|}{1 - 2|f_i|}} & ||f||_\infty < 1/2 \,. \end{cases} \tag{1}$$

- There exists another formula with respect to the ℓ_1 norm, though we omit it here.

Relevant Background: Lewicki (2000)

- In 2000, Lewicki extended these results to partial formula when the subspace has codimension 2, though the results seem to imply that classical techniques are insufficient to compute these values.

Definition 1.1. A subspace $Y \subset X_n$ is called *admissible* if $Y = \ker(f) \cap \ker(g)$, where f and g satisfy (1.1), $f_j, g_j < 1/2$ for any $j = 1, \ldots, n$ and the following conditions hold for $i = 3, \ldots, n$:
if

$$\det\begin{pmatrix} 2f_i - 1 & 2f_1 - 1 \\ 2g_i - 1 & 2g_2 - 1 \end{pmatrix} < 0$$

then

$$\det\begin{pmatrix} 2f_1 - 1 & 2f_i - 1 \\ 2g_2 - 1 & 2(g_2 + g_i) - 1 \end{pmatrix} < 0; \tag{1.2}$$

if

$$\det\begin{pmatrix} 2f_i - 1 & 2f_1 - 1 \\ 2g_i - 1 & 2g_2 - 1 \end{pmatrix} > 0$$

then

$$\det\begin{pmatrix} 2f_1 - 1 & 2(f_i + f_1) - 1 \\ 2g_2 - 1 & 2g_i - 1 \end{pmatrix} > 0. \tag{1.3}$$

Theorem 1.8. *Let Y be an admissible subspace of X_n. Let a, b be defined by (1.11) and (1.12). Suppose that*

$$f_3/g_3 > f_4/g_4 > \ldots > f_n/g_n. \tag{1.16}$$

If $a \geqslant b$, then $\lambda(Y, X_n) = a$ if

$$g_3(a-1)^{-1} > f_3\left(\sum_{i \in I_1} g_i/(1-2f_i)\right) + g_3\left(\sum_{i \in I_2} g_i/(1-2g_i)\right); \tag{1.17}$$

and

$$f_n(a-1)^{-1} > f_n\left(\sum_{i \in I_1} f_i/(1-2f_i)\right) + g_n\left(\sum_{i \in I_2} f_i/(1-2g_i)\right); \tag{1.18}$$

If $b > a$, then $\lambda(Y, X_n) = b$ if

$$f_n(b-1)^{-1} > g_n\left(\sum_{i \in I_2} f_i/(1-2g_i)\right) + f_n\left(\sum_{i \in I_1} f_i/(1-2f_i)\right); \tag{1.19}$$

$$g_3(b-1)^{-1} > g_3\left(\sum_{i \in I_2} g_i/(1-2g_i)\right) + f_3\left(\sum_{i \in I_1} g_i/(1-2f_i)\right). \tag{1.20}$$

Moreover, if (1.17), (1.18) or (1.19), (1.20) are satisfied, then the minimal projection is uniquely determined by the vectors $z, w \in R^n$, $f(z) = g(w) = 1$, $f(w) = g(z) = 0$ (compare with Lemma 0.4) of the form:

$$z_i = 0 \text{ for } i \in I_2, z_i = (a-1)/(1-2f_i), \text{ for } i \in I_1,$$

$$z_2 = -\left(\sum_{i \in I_1} g_i z_i\right)/g_2, \quad z_1 = \left(1 - \sum_{i \in I_1} f_i z_i\right)/f_1;$$

$$w_i = (a-1)/(1-2g_i), \text{ for } i \in I_2, w_i = 0, \text{ for } i \in I_1,$$

$$w_2 = (a - 1 - (1-2f_1)z_2)/(1-2g_2),$$

Theoretical Guarantees: ℓ_∞ Formulas

- In light of Universal Approximation Theorem by Cybenko, continuity is the only requirement to guarantee a function can be approximated by neural networks.
- Returning to the Blatter-Cheney Formula, we see this is a continuous, though not smooth, function of the input f.

$$\lambda(f^\perp) = \begin{cases} 1 & ||f||_\infty \geq 1/2 \\ 1 + \dfrac{1}{\sum_{i=1}^{n} \dfrac{|f_i|}{1-2|f_i|}} & ||f||_\infty < 1/2\,. \end{cases} \tag{2}$$

- This does **not** guarantee, however, that we can find a neural network that can perform this approximation.
- What about other cases where we don't have a formula?

Theoretical Guarantees: Continuity of Linear Programs

Theorem (1)

Consider the parametric linear program:

$$\begin{array}{lll} \textit{maximize} & c^T(t)x + d^T(t)q & \textit{subject to} \\ A(t)x + B(t)q = a(t), & C(t)x + D(t)q \leq b(t), & q \geq 0 \end{array}$$

Maximal value is continuous w.r.t. parameter vector t at t_0 provided that the primal and dual problem are each bounded and feasible at t_0 [Martin, 1975].

Theoretical Guarantees: Casting as a Linear Program

- The determination of minimal projection constants for certain finite dimensional spaces can be cast as a feasible linear programming problem as seen below [Foucart, 2016], hence the strong duality property for linear programs implies the boundedness and feasibility of the corresponding dual and primal problems.

$$\text{minimize}_{d,P}\ d \text{ subject to}$$
$$(U^T \bigotimes I_n)vec(P) = vec(U),$$
$$(\tilde{U}^T \bigotimes \tilde{U}^T)vec(P) = 0$$
$$\text{and to } \sum_{j=1}^{n} |P_{ij}| \leq d, \text{ for all } i \in [[1, n]]$$

- Thus Theorem 1 can be applied to guarantee that the minimal projection constant varies continuously with the subspace.

Architectures and Experimental Setup

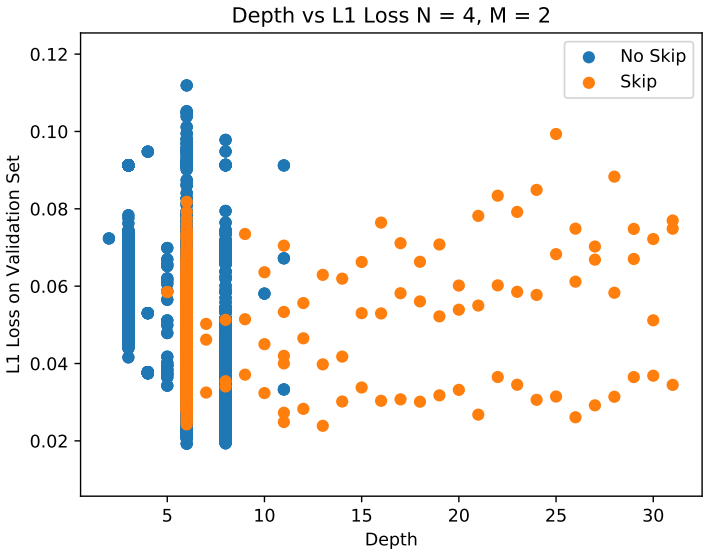
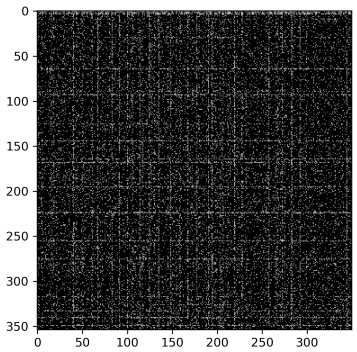


Figure: Best ℓ_1 losses achieved by networks of different depths. Note that due to early stopping several of the deeper networks without skip connections were terminated at high loss values, and thus are not shown in this plot.

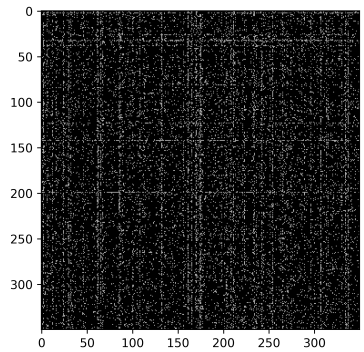
Parameter	Minimum	Step Size	Maximum
Depth	2	1	50*
Width	50	50	550
L. Rate	0.001	variable	.04

Table: Grid Search Parameters (* for depths greater than 3 hidden layers, a fixed hidden layer width of 100 is used)

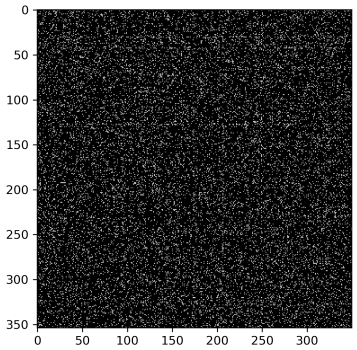
Visualization of Weight Matrices



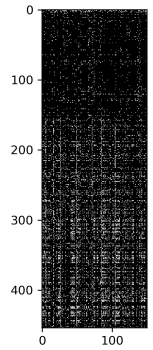
(a) Network with skip connections trained on scrambled data first weight matrix



(b) Trained network without skip connections first weight matrix



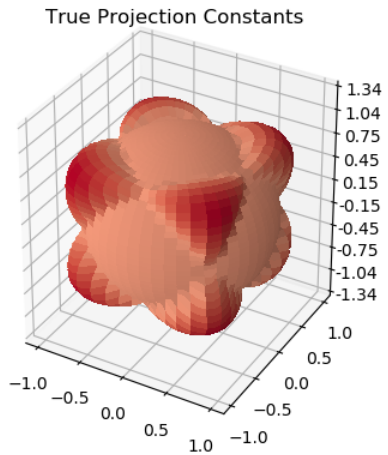
(c) Trained network with skip connections first weight matrix



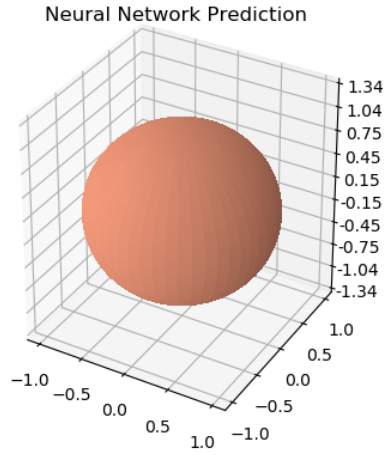
(d) Trained network with skip connections last weight matrix

Figure: gray-scale plots of weight matrices.

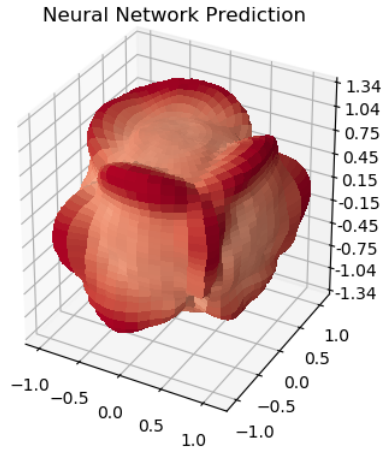
Visualization of Learned Representations



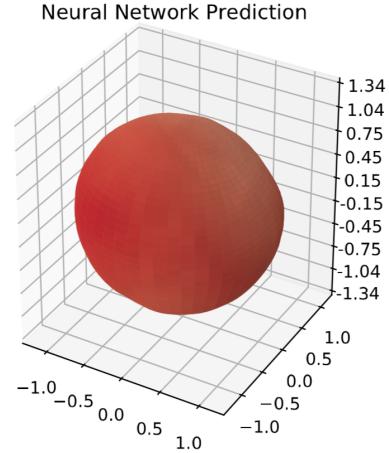
(a) True projection constants



(b) Neural Network without skip connections



(c) Neural Network with skip connections



(d) Neural Network trained with scrambled data

Figure: Predicted projection constants as subspace is varied over a 2-sphere. The distance of a point on the surface from the origin is the value of the projection constant for the associated subspace.

Conclusions and Future Work

- Conclusion: No skip connections performs better, but skip connections seem to be learning better and more meaningful representations.
- More computational power for better grid searching and cross validation (better hyper-parameter tuning)
- Investigate the relationship between functions whose points are the solutions to linear programming problems and neural networks using the Martin paper.
- Experiment more with differences between expressive power of neural networks with and without skip connections
- Visualizing activations may give more insight than visualizing weight matrices.
- All code for this project along with the datasets can be downloaded for free at https://github.com/RyanMalt/ProjectionConstants.

References

J. Blatter and E. W. Cheney. "Minimal projections on hyperplanes in sequence spaces". In: *Annali di Matematica Pura ed Applicata* 101.1 (Dec. 1974), pp. 215–227. ISSN: 1618-1891. DOI: 10.1007/BF02417105. URL: https://doi.org/10.1007/BF02417105.

François Chollet et al. *Keras*. https://keras.io. 2015.

G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals and Systems* 2.4 (Dec. 1989), pp. 303–314. ISSN: 1435-568X. DOI: 10.1007/BF02551274. URL: https://doi.org/10.1007/BF02551274.

Simon Foucart. "Computation of Minimal Projections and Extensions". In: *Numerical Functional Analysis and Optimization* 37.2 (2016), pp. 159–185. DOI: 10.1080/01630563.2015.1091014. eprint: https://doi.org/10.1080/01630563.2015.1091014. URL: https://doi.org/10.1080/01630563.2015.1091014.

P. Grohs et al. "A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations". In: *ArXiv e-prints* (Sept. 2018). arXiv: 1809.02362 [math.NA].

G. Lewicki. "Minimal Projections onto two dimensional subspaces of $\ell_\infty^{(4)}$". In: *Journal of Approximation Theory* 88 (1997), p. 92.

D. H. Martin. "On the continuity of the maximum in parametric linear programming". In: *Journal of Optimization Theory and Applications* 17.3 (Nov. 1975), pp. 205–210. ISSN: 1573-2878. DOI: 10.1007/BF00933875. URL: https://doi.org/10.1007/BF00933875.

Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

F. Metcalf. "Determination of a minimal projection from C[- 1, 1] onto the quadratics". In: *Numerical Functional Analysis and Optimization* 11 (1990), p. 1.

L. Skrzypek. "Chalmers - Metcalf operator and uniqueness of minimal projections". In: *Journal of Approximation Theory* 148 (2007), p. 71.

Dmitry Yarotsky. "Error bounds for approximations with deep ReLU networks". In: *CoRR* abs/1610.01145 (2016). arXiv: 1610.01145. URL: http://arxiv.org/abs/1610.01145.