Antisymmetrization cancellations

Nilin

February 2022

We consider the numerical stability of the explicitly antisymmetrized function

$$Af(x) = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) f(x_\sigma), \qquad x_\sigma := x_{\sigma(1)}, \ldots, x_{\sigma(n)},$$

Specifically we may consider the case when f is given by a two-layer neural network

$$f(x) = \sum_{k=1}^{m} a_k \tau(w_k \cdot x) = \sum_{k=1}^{m} \tau(u_k \cdot x), \tag{1}$$

where τ is the ReLU activation function, $w_k \in \mathbb{R}^{nd}$, $a_k \in \mathbb{R}$, and $u_k = a_k w_k$.

We use the He initialization of (1) such that $a_k \sim \mathcal{N}(0, 2/m)$ and $w_k \sim \mathcal{N}(0, 2/(nd))^{\otimes nd}$ for $k = 1, \ldots, m$.

1 Correlations between x and $Af(x)$ for fixed f

In the following we sample $x \in \mathbb{R}^{nd}$ uniformly from the unit sphere.

1.1 Distance to degenerate subspaces

An antisymmetric function $g = Af$ is zero on each of the subspaces

$$V_{ij} = \{x \in \mathbb{R}^{nd} | x_i = x_j\}$$

of co-dimension d, for $i \neq j$. We investigate the relationship between $f(x)$ and the $L^{\infty,2}$-distance to the union of these subspaces,

$$\delta(x) = \min\{\|x_i - x_j\| : i \neq j\}.$$

For $d = 3$, $n = 5$ particles, $m = 1000$ layers, we get a correlation coefficient of

$$\mathrm{corr}(\delta^2(X), Af^2(X)) \approx 0.3$$

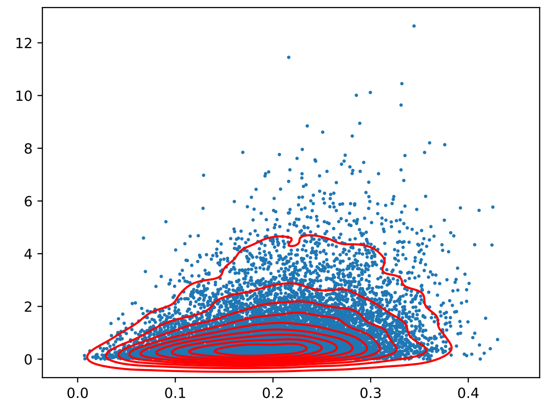

Figure 1: x axis: $\delta(X)$, y axis: $|Af(X)|$

2 Uncorrelated function values hypothesis

Suppose that the following holds:

Hypothesis 1. *The unsymmetrized function f is such that $f(x)$ carries little information about $f(x_\sigma)$ for a nontrivial permutation σ.*

Under hypothesis 1 we would expect that

$$Af(X) = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) f(X_\sigma) \approx_{\text{distribution}} \mathrm{RS}(\hat{X}, S) := \sum_{i=1}^{n!} S f(\hat{X}_i),$$

where S is a Rademacher R.V. and $\hat{X}$ is drawn from the same distribution as X. Here RS is for 'resample'.

2.1 Oscillating activation function

To illustrate the hypothesis we consider a chaotic f with activation function $\tilde{\tau}(x) = \sin(100x)$. Parameters $n = 8, d = 3, m = 10$. The distributions of the outputs from the antisymmetrized function Af (blue) and the resampled distributions (red) are

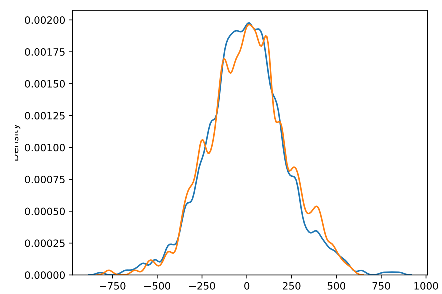

Figure 2: $Af(X)$ (blue) vs $\mathrm{RS}(X, S)$ (red) for oscillating activation. $n = 8$

Thus we see that the hypothesis of uncorrelated function values holds for an oscillatory activation function.

Here the raw (un-symmetrized) function values $f(X)$ have the distribution:

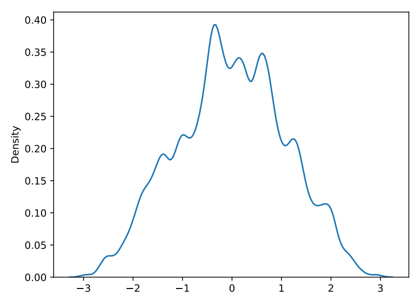

Figure 3: distribution of $f(X)$ for chaotic f

2.2 ReLU activation

For the ReLU activation function the distribution of $f(X)$ appears identical to a rescaling of the chaotic function.

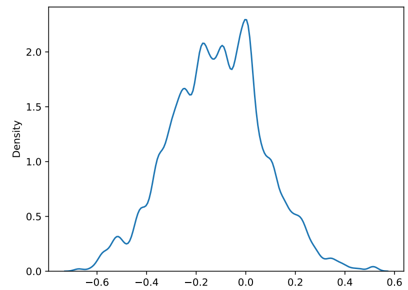

Figure 4: distribution of $f(X)$ for ReLU activation

However, the antisymmetrized function has much more cancellation with the ReLU activation:

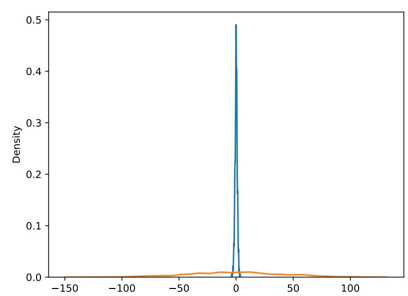

Figure 5: $Af(X)$ (blue) vs resampled $\mathrm{RS}(X, S)$ (red) for ReLU activation. $n = 8$

Thus the hypothesis of uncorrelated f values does not hold for (1) with the ReLU activation.

3 n-dependence of $\mathrm{Var}\, f(X)$ and $\mathrm{Var}\, Af(X)$

The He initialization yields the correct scaling of the unsymmetrized f when X is a standard Gaussian:

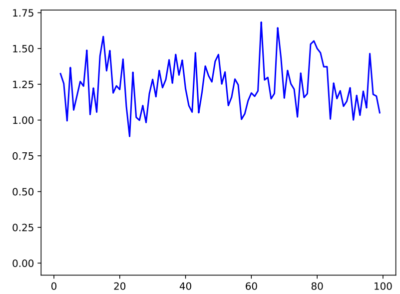

Figure 6: $\mathrm{Var}\, f(X)$ as a function of n. Each data point is the median variance of 20 networks sampled with the He initialization.

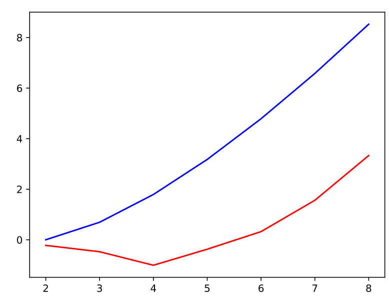

Figure 7: $\log \mathrm{Var}\, Af(X)$ (red) and the number of antisymmetrization terms $\log(n!)$ (blue) as a function of n. $d = 3, m = 10$. Each data point shows the median variance of 10 networks sampled with the He initialization.

Similar plots result when we sample X from the sphere of radius $\sqrt{nd}$.