Double Descent on 2-layer Networks with ReLU Activation and Randomized Weights and Biases

Jacob Levine [1] Colin Lu [1] Matthew Tang [1]

[1]University of Illinois

Summary

We extend the work of Li et al. 2021 to include biases. We present a bound on the variance of 2-layer neural networks with the first layer's weights and biases sampled from zero-mean i.i.d. Gaussian RVs. We include an additional term in the bound to account for the additional variance that the biases bring. We note that with correct choices of parameters, our bound is asymptotically equivalent to the non-bias case.

Motivation

- Li et al. 2021 explore benign overfitting on 2-layer networks with ReLU activations, with the first layer having randomized weights. Method includes decomposition into bias and variance, shows that overparameterized networks $s \sim O(n^2)$ can achieve near-optimal performance in specific settings.
- Networks discussed do not include biases, but assumptions on zero training loss; may be limiting in low-dimensional cases. Assumption that $d = n^\alpha$ may abate this a little, but still can cause problems.
- One-dimensional input case:

$$\begin{aligned} g(x) &= \beta\sigma(\mathbf{w}x) \\ &= \sum_i \beta_i \mathbf{w}_i x[\mathbf{w}_i x > 0] \\ &= [x > 0]\left(\sum_{i:\mathbf{w}_i>0} \beta_i \mathbf{w}_i\right) x + [x < 0]\left(\sum_{i:\mathbf{w}_i<0} \beta_i \mathbf{w}_i\right) x \end{aligned}$$

  Thus, equivalent to picking slopes for lines when $x > 0$ and when $x < 0$. Cannot correctly predict $(x_1, y_1), (x_2, y_2)$ if $x_1, x_2 > 0$ and $\frac{y_1}{x_1} \neq \frac{y_2}{x_2}$
- High dimensional case: since $d + 1$ points define a hyperplane, random basis generated by column of $\mathbf{W}$ randomly separates use of hyperplanes. Thus, if $d + 1$ points in the same region, likely not able to get zero error
- Without biases, cannot classify points that are collinear with $\mathbf{0}$ in most cases.
- Huang et al. 2006 shows a significant improvement in representative power by allowing biases in the first layer. A network with randomized first layer weights and biases can for any continuous function, with high probability, act as a universal approximator.
- In the overfitting regime, we would like our network to have near perfect fit on training data. By allowing biases, we give our randomly initialized network the ability to represent a much larger set of functions, hence why we attempt to add biases into Li et al. 2021's results.

Bias Notes

- Sample biases as i.i.d. Gaussians: $\mathcal{N}(0, \sigma_b^2) \in \mathbb{R}^{s\times 1}$
- $\mathbf{z}_{x_i+\xi}$ is redefined to be $\sigma(\mathbf{W}(x_i + \xi_i) + \mathbf{b})$
- $D_x(\mathbf{W})$ is redefined as diag $\{[\mathbf{W}_1^\top x + b_1 > 0], \ldots, [\mathbf{W}_s^\top x + b_s > 0]\}$
- Setting is overparameterized: $s > n > d$

Theorem 1: Bounds on Variance

$$\begin{aligned} \mathbb{V}_R &= \mathbb{E}_x\left[\mathbb{E}_\epsilon\left[\mathbf{z}_x^\top(\mathbf{Z}_\xi^\top\mathbf{Z}_\xi)^\dagger\mathbf{Z}_\xi^\top\epsilon\epsilon^\top\mathbf{Z}_\xi(\mathbf{Z}_\xi^\top\mathbf{Z}_\xi)^\dagger\mathbf{z}_x\right]\right] \\ &= \sigma_0\left[\mathbf{z}_x^\top(\mathbf{Z}_\xi^\top\mathbf{Z}_\xi)^\dagger\mathbf{Z}_\xi^\top\mathbf{Z}_\xi(\mathbf{Z}_\xi^\top\mathbf{Z}_\xi)^\dagger\mathbf{z}_x\right] \\ &= \sigma_0\left[\mathrm{Tr}[\mathbf{z}_x^\top(\mathbf{Z}_\xi^\top\mathbf{Z}_\xi)^\dagger\mathbf{z}_x]\right] \\ &= \frac{\sigma_0}{n}\left[\mathrm{Tr}\left[\mathbf{z}_x^\top\left(\frac{1}{n}\mathbf{Z}_\xi^\top\mathbf{Z}_\xi\right)^\dagger\mathbf{z}_x\right]\right] \end{aligned}$$

Thus, need to bound the least positive eigenvalue $\mu_{\min}\left(\frac{1}{n}\mathbf{Z}_\xi^\top\mathbf{Z}_\xi\right)$ By Lemma 2, we know that probability at least $1 - 2e^{\frac{-s}{b_1}} - 2e^{\frac{-d}{b_1}}$,

$$\left\|\frac{1}{n}\mathbf{Z}_\xi^\top\mathbf{Z}_\xi - \mathbf{W}^\top\Sigma_\xi\mathbf{W}\right\|_2 = O\left(\sqrt{\frac{d}{n}}(d\sigma_w^2 + \sigma_w\sigma_b + \sigma_b^2)\right)$$

Li et al. proceed to lower bound the least positive eigenvalue by using lemma 2, and in our case this is

$$\begin{aligned} \mu_{\min}\left(\frac{1}{n}\mathbf{Z}_\xi^\top\mathbf{Z}_\xi\right) &\geq \mu_{\min}\left(\mathbf{W}\Sigma_\xi\mathbf{W}^\top\right) - b_2\left(\sqrt{\frac{d}{n}}(d\sigma_w^2 + \sigma_w\sigma_b + \sigma_b^2)\right) \\ &\geq \frac{1}{b_3}d\sigma_w^2 - b_2\left(\sqrt{\frac{d}{n}}(d\sigma_w^2 + \sigma_w\sigma_b + \sigma_b^2)\right) \geq \frac{1}{b_3}d\sigma_w^2. \end{aligned}$$

for some constant b_2, where the last step is leveraging their lemma 3 bounding $\mu_{\min}\left(\mathbf{W}\Sigma_\xi\mathbf{W}^\top\right)$ by the quantity shown. We remark that the value obtained by lemma 2 seems to disappear and is unnecessary aside from yielding a more favorable constant in the original paper, and is thus unclear why it is used in the first place. Still, we characterize it for the purposes of extending the paper. Thus,

$$\begin{aligned} \mathbb{V}_R &\leq b_4\frac{\sigma_0^2}{nd\sigma_w^2}\mathbb{E}_x\left[\mathrm{Tr}\left(\mathbf{z}_x^\top\mathbf{z}\right)\right] \\ &\leq b_4\frac{\sigma_0^2}{nd\sigma_w^2}s(\sigma_w^2\mathbb{E}_x\left[\|x\|_2^2\right] + 2\sigma_w\sigma_b\mathbb{E}_x\left[\|x\|_2\right] + \sigma_b^2) \end{aligned}$$

Note that in the original paper the σ_w^2 factors cancel, but we no longer are able to do so with the addition of extra terms. We conjecture that a nicer looking bound exists with further analysis.

Lemma 2: Asymptotic Bounds on Inner Variance Term

$$\begin{aligned} \frac{1}{n}\mathbf{Z}_\xi^\top\mathbf{Z}_\xi - \mathbf{W}\Sigma_\xi\mathbf{W} &= \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}\mathbf{z}_{x_i+\xi_i}^\top - \mathbf{W}\Sigma_\xi\mathbf{W} \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}\mathbf{z}_{x_i+\xi_i}^\top - \frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top \qquad (8) \\ &\quad + \frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top - \mathbf{W}\Sigma_\xi\mathbf{W} \qquad (9) \end{aligned}$$

Analysis of Eq. (8)

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}\mathbf{z}_{x_i+\xi_i}^\top - \frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}\mathbf{z}_{x_i+\xi_i}^\top - \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top \\ &\quad + \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top - \frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}(\mathbf{z}_{x_i+\xi_i} - (\mathbf{W}(x_i+\xi_i)+\mathbf{b}))^\top + \frac{1}{n}\sum_{i=1}^n(\mathbf{z}_{x_i+\xi_i} - (\mathbf{W}(x_i+\xi_i)+\mathbf{b}))(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}\mathbf{z}_{x_i+\xi_i}^\top + \frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top \end{aligned}$$

Similar to the technique in the paper, the last step follows because $\sigma(y) - y = \sigma(-y)$, and since $\mathbf{W}, \mathbf{b}$ are Gaussian with mean 0, $\mathbf{W}$ and $-\mathbf{W}$ have the same distribution, same with $\mathbf{b}$ and $-\mathbf{b}$. Now bound the norm

$$\begin{aligned} \|\text{Eq. (8)}\|_2 &\leq \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}\mathbf{z}_{x_i+\xi_i}^\top\right\|_2 + \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{z}_{x_i+\xi_i}(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top\right\|_2 \\ &= \left\|\frac{1}{n}\sum_{i=1}^n D_{x_i+\xi_i}(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top D_{x_i+\xi_i}\right\|_2 \\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^n D_{x_i+\xi_i}(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top\right\|_2 \\ &\leq 2\left\|\frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)+\mathbf{b})(\mathbf{W}(x_i+\xi_i)+\mathbf{b})^\top\right\|_2 \\ &= 2\left\|\frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)+\mathbf{b})((x_i+\xi_i)^\top\mathbf{W}^\top+\mathbf{b}^\top)\right\|_2 \\ &\leq 2\left\|\frac{1}{n}\sum_{i=1}^n(\mathbf{W}(x_i+\xi_i)(x_i+\xi_i)^\top\mathbf{W}^\top)\right\|_2 \\ &\quad + 2\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{W}(x_i+\xi_i)\mathbf{b}^\top + \mathbf{b}(x_i+\xi_i)^\top\mathbf{W}^\top\right\|_2 + 2\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{b}\mathbf{b}^\top\right\|_2 \\ &\leq 2\frac{d}{n}d\sigma_w^2 + 4\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{W}(x_i+\xi_i)\mathbf{b}^\top\right\|_2 + 2\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{b}\mathbf{b}^\top\right\|_2 \\ &\leq 2\frac{d}{n}d\sigma_w^2 + \frac{4}{n}\left\|\mathbf{W}(x_i+\xi_i)\mathbf{b}^\top\right\|_2 + \frac{2}{n}\left\|\mathbf{b}\mathbf{b}^\top\right\|_2 \\ &\leq 2\frac{d}{n}d\sigma_w^2 + \frac{4}{n}\|\mathbf{W}(x_i+\xi_i)\|_2\|\mathbf{b}\|_2 + \frac{2}{n}\|\mathbf{b}\|_2^2 \\ &\leq 2\frac{d}{n}d\sigma_w^2 + \frac{4}{n}\sqrt{d\sigma_w^2}\sqrt{d\sigma_b^2} + \frac{2}{n}(d\sigma_b^2) \end{aligned}$$

We remark that the last steps using the bound on $\|\mathbf{b}\mathbf{b}^\top\|_2$ by σ_b^2 are conjectured, chosen analogous to the σ_w^2 bound in the first term, though are unsure why authors are capable of deriving a bound on a Gaussian random variable without explicitly noting that the bound is with high probability. We can bound $\|\mathbf{W}(x_i+\xi_i)\|_2$ using the bound derived in the original paper for $\|\mathbf{W}(x_i+\xi_i)(\mathbf{W}(x_i+\xi_i))^\top\|_2$, since $\|\mathbf{A}\|_2 = \sqrt{\|\mathbf{A}\mathbf{A}^\top\|_2}$

Remark: We can obtain a similar bound on Eq. 9's norm using the same method, which yields a bound of $b_4\sqrt{\frac{d}{n}}d\sigma_w^2$ in the original paper for some constant b_4. While we have not yet calculated exactly the bound under the new setting, we conjecture similarly that we achieve $b_4(\sqrt{\frac{d}{n}}d\sigma_w^2 + 4\sqrt{\frac{d}{n}}\sigma_w\sigma_b + 2\sqrt{\frac{d}{n}}\sigma_b^2)$. Since $\frac{d}{n} < 1$, the square root bound is larger, and subsumes the terms from Eq. 8, and the overall bound should be

$$O\left(\sqrt{\frac{d}{n}}(d\sigma_w^2 + \sigma_w\sigma_b + \sigma_b^2)\right).$$

References

Li, Z., Z.-H. Zhou, and A. Gretton (June 2021). *Towards an Understanding of Benign Overfitting in Neural Networks*. arXiv:2106.03212 [cs, stat]. URL: http://arxiv.org/abs/2106.03212 (visited on 12/13/2022).

Huang, G.-B., L. Chen, and C.-K. Siew (July 2006). "Universal Approximation Using Incremental Constructive Feedforward Networks With Random Hidden Nodes". en. In: *IEEE Transactions on Neural Networks* 17.4, pp. 879–892. ISSN: 1045-9227. DOI: 10.1109/TNN.2006.875977. URL: http://ieeexplore.ieee.org/document/1650244/ (visited on 12/13/2022).