Deep Mean Field Theory: Variance and Width Variation by Layer as Methods to Control Gradient Explosion

Anonymous authors
Paper under double-blind review

Abstract

A recent line of work has studied the statistical properties of neural networks to great success from a *mean field theory* perspective, making and verifying very precise predictions of neural network behavior and test time performance. In this paper, we build upon these previous works to explore two methods for taming the behaviors of random residual networks (with only fully connected layers and no batchnorm).

The first method is *width variation (WV)*, i.e. varying the widths of layers as a function of depth. We show that width decay reduces gradient explosion without affecting the mean forward dynamics of the random network.

The second method is *variance variation (VV)*, i.e. changing the initialization variances of weights and biases over depth. We show VV, used appropriately, can reduce gradient explosion of tanh and ReLU resnets from $\exp(\Theta(\sqrt{L}))$ and $\exp(\Theta(L))$ respectively to constant $\Theta(1)$. A complete phase-diagram is derived for how variance decay affects different dynamics, such as those of gradient and activation norms. In particular, we show the existence of many phase transitions where these dynamics switch between exponential, polynomial, logarithmic, and even constant behaviors.

Using the obtained mean field theory, we are able to track surprisingly well how VV at initialization time affects training and test time performance on MNIST after a set number of epochs: the level set of test/train set accuracies coincide with the level sets of certain gradient norms.

1 Introduction

Deep mean field theory studies how random neural networks behave with increasing depth, in the limit as widths go to infinity. Some of the landmark results include precise predictions of norms of activation vectors across depth in feedforward (Poole et al., 2016) and residual networks [1] (Yang and Schoenholz, 2017), correlations between corresponding activation vectors computed from two input vectors (Poole et al., 2016; Yang and Schoenholz, 2017), and gradient norms (Schoenholz et al., 2016; Yang and Schoenholz, 2017). The latter is used to remarkably predict test time performance given only initialization variances (Schoenholz et al., 2016; Yang and Schoenholz, 2017). Overall, a program emerged toward building a mean field theory for state-of-the-art neural architectures as used in the wild, so as to provide optimal initialization parameters quickly for any neural network user.

In this paper, we contribute to this program by studying how width variation (WV), as practiced commonly, can change the behavior of quantities mentioned above, with gradient norm being of central concern. We find that WV can dramatically reduce gradient explosion without affecting the mean dynamics of forward computation, such as the activation norms, although possibly increasing deviation from the mean in the process (Section 4).

We also study a second method, variance variation (VV), for mean field dynamics of a random neural network (Section 5). In this paper, we focus on its application to tanh and

[1] without batchnorm and with only fully connected layers

ReLU residual networks, where we show that VV can dramatically ameliorate gradient explosion, and in the case of ReLU resnet, activation explosion [2].

Previous works (Poole et al., 2016; Schoenholz et al., 2016; Yang and Schoenholz, 2017) have focused on how network architecture and activation functions affect the dynamics of mean field quantities, subject to the constraint that initialization variances and widths are constant across layers. In each combination of (architecture, activation), the mean field dynamics remain the same asymptotically regardless of the variances. For example, tanh feedforward networks have $\exp(\Theta(l))$ forward and backward dynamics, while tanh residual networks have $\mathsf{poly}(l)$ forward and $\exp(\Theta(\sqrt{l}))$ backward dynamics. Such asymptotics were considered characteristics of the (architecture, activation) combination (Yang and Schoenholz, 2017). We show by counterexample that *this perception is erroneous*. In fact, as discussed above, WV can control the gradient dynamics arbitrarily and VV can control forward and backward dynamics jointly, all without changing the network architecture or activation. To the best of our knowledge, this is the first time initialization schemes for reducing gradient explosion or vanishing have been proposed that vary variance and/or width across layers.

Using the derived mean field theories of VV, we make stunning predictions of test time performances of grids of resnets.

2 Preliminaries

Notations and Settings. We adopt the notations of Yang and Schoenholz (2017) and review them briefly. Consider a vanilla feedforward neural network of L layers, with each layer l having $N^{(l)}$ neurons; here layer 0 is the input layer. Let $x^{(0)} = (x_1^{(0)}, \ldots, x_{N^{(0)}}^{(0)})$ be the input vector to the network, and let $x^{(l)}$ for $l > 0$ be the activation of layer l. Then a neural network is given by the equations $x_i^{(l)} = \phi(h_i^{(l)}), h_i^{(l)} = \sum_{j=1}^{N^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)} + b_i^{(l)}$ where (i) $h^{(l)}$ is the pre-activation at layer l, (ii) $w^{(l)}$ is the weight matrix, (iii) $b^{(l)}$ is the bias vector, and (iv) ϕ is a nonlinearity, for example tanh or ReLU, which is applied coordinatewise to its input. To lighten up notation, we suppress the explicit layer numbers l and write $x_i = \phi(h_i), h_i = \sum_{j=1}^{N} w_{ij}\underline{x}_j + b_i$ where $\bullet$ implicitly denotes $\bullet^{(l)}$, and $\underline{\bullet}$ denotes $\bullet^{(l-1)}$ (and analogously, $\overline{\bullet}$ denotes $\bullet^{(l+1)}$). When the widths are constant $N^{(l)} = N^{(m)}, \forall m, l$, residual network (He et al., 2015; 2016) adds an *identity connection* or *skip shortcut* that "jumps ahead" every couple layers. We adopt one of the simplified residual architectures defined in Yang and Schoenholz (2017) for ease of analysis [3], in which every *residual block* is given by

$$x_i = \sum_{j=1}^{M} v_{ij}\phi(h_j) + a_i + \underline{x}_i, \qquad h_i = \sum_{j=1}^{\underline{N}} w_{ij}\underline{x}_j + b_i.$$

where $M^{(l)}$ is the width of the "hidden layer" of the residual block, $(v_{ij}^{(l)})_{i,j=1}^{N^{(l)},M^{(l)}}$ is a new set of weights and $(a_i^{(l)})_{i=1}^{N^{(l)}}$ is a new set of biases for every layer l. If we were to change the width of a residual network, as is done in practice, we need to insert *"projection" residual blocks* (He et al., 2015; 2016) every couple layers. We assume the following simplified projection residual block in this paper, for the ease of presentation [4]:

$$x_i = \sum_{j=1}^{M} v_{ij}\phi(h_j) + y_i + a_i, \qquad h_i = \sum_{j=1}^{N} w_{ij}\underline{x}_j + b_i, \qquad y_i = \sum_{j=1}^{\underline{N}} \pi_{ij}\underline{x}_j.$$

[2] this is the phenomenon where in deep ReLU resnet, the activation vectors blow up exponentially with depth, invalidating the computation with `infs`; see Yang and Schoenholz (2017)

[3] It is called *full residual network* by Yang and Schoenholz (2017), but in this paper, we will simply assume this architecture whenever we say *residual network*.

[4] If the reader is disappointed by our simplifications, he or she is encouraged to consult He et al. (2016), which empirically explored many variations of residual blocks

where $y = (y_i)_{i=1}^{N^{(l)}}$ is a "projection" of $x = (x_i)_{i=1}^{N^{(l-1)}}$ [5], and $(\pi_{ij})_{i,j=1}^{N^{(l)},N^{(l-1)}}$ is the "projection" matrix. Note that we only consider fully-connected affine layers instead of convolutional layers.

Deep mean field theory is interested in the "average behavior" of these network when the weights and biases, $w_{ij}^{(l)}, b_i^{(l)}, \pi_{ij}^{(l)}, v_{ij}^{(l)}$, and $a_i^{(l)}$ are sampled i.i.d. from Gaussian distributions. Following standard practice, they respectively have standard deviations $\sigma_w^{(l)}/\sqrt{N^{(l-1)}}, \sigma_b^{(l)}, \sigma_\pi^{(l)}/\sqrt{N^{(l-1)}}, \sigma_v^{(l)}/\sqrt{M^{(l)}}$, and $\sigma_a^{(l)}$; here we normalize the variance of weight parameters so that, for example, the variance of each h_i is σ_w^2, assuming each $\underline{x}_j$ is fixed. While previous works have all focused on fixing $\sigma_\bullet^{(l)}$ to be constant across depth, in this paper we are interested in studying varying $\sigma_\bullet^{(l)}$. In particular, other than $\sigma_\pi^{(l)}$ which we fix at 1 across depth (so that "projection" doesn't act like an "expansion" or "contraction"), we let $\sigma_\bullet^{(l)} = \sigma_\bullet l^{-\beta_\bullet}$ for each $\bullet = v, w, a, b$, where $\sigma_\bullet, \beta_\bullet$ are constant w.r.t. l. Hereafter the bar notation $\underline{\bullet}, \bullet, \overline{\bullet}$ do not apply to σs, so that, by σ_a, for example, we always mean the constant σ_a.

Statistical Assumptions. We make several key "mean field" assumptions, which were formulated in its entirety first in Yang and Schoenholz (2017) (though Axiom 2(a) has been stated first by Schoenholz et al. (2016)). While these assumptions may be mathematically unsatisfying, identifying them and discovering that they lead to highly precise prediction is in fact one of the most important contributions of deep mean field theory.

Axiom 1 (Symmetry of activations and gradients). *(a) We assume $\langle (h_i^{(l)})^2 \rangle = \langle (h_j^{(l)})^2 \rangle$ and $\langle (x_i^{(0)})^2 \rangle = \langle (x_j^{(0)})^2 \rangle$ for any i, j, l. (b) We also assume that the gradient $\partial E/\partial x_i^{(l)}$ with respect to the loss function E satisfies $\langle (\partial E/\partial x_i^{(l)})^2 \rangle = \langle (\partial E/\partial x_j^{(l)})^2 \rangle$ for any i, j, l.*

Axiom 2 (Gradient independence). *(a) We assume the we use a different set of weights for backpropagation than those used to compute the network outputs, but sampled i.i.d. from the same distributions. (b) For any loss function E, we assume that the gradient at layer l, $\partial E/\partial x_i^{(l)}$, is independent from all activations $h_j^{(l)}$ and $x_j^{(l-1)}$ from the previous layer.*

One can see that Axiom 1(a) is satisfied if the input $x^{(0)} \in \{\pm 1\}^N$ and Axiom 1(b) is satisfied if Axiom 2 below is true and the gradient at the last layer $\partial E/\partial x L \in \{\pm 1\}^N$. Axiom 2(a) was first made in Schoenholz et al. (2016) for computing the mean field theory of gradients for feedforward tanh networks. This is similar to the practice of feedback alignment Lillicrap et al. (2014).

Mean Field Quantities. Now we define the central quantities studied in this paper.

Definition 2.1. Fix an input $x^{(0)}$. Define the **length quantities** $q^{(l)} := \langle (h_1^{(l)})^2 \rangle$ and $p^{(l)} := \langle (x_1^{(l)})^2 \rangle$ for $l > 0$ and $p^{(0)} = \|x^{(0)}\|^2/N$. Here the expectations $\langle \bullet \rangle$ are taken over all random initialization of weights and biases for all layers l, as $N \to \infty$ (large width limit).

Note that in our definition, the index 1 does not matter by Axiom 1.

Definition 2.2. Fix two inputs $x^{(0)}$ and $x^{(0)\prime}$. We write $\bullet'$ to denote a quantity $\bullet$ with respect to the input $x^{(0)'}$. Then define **the correlation quantities** $\gamma^{(l)} := \langle h_1^{(l)} h_1^{(l)\prime} \rangle$ and $\lambda^{(l)} := \langle x_1^{(l)} x_1^{(l)\prime} \rangle$ for $l > 0$ and $\gamma^{(0)} = x^{(0)} \cdot x^{(0)\prime}/N$, where the expectations $\langle \bullet \rangle$ are taken over all random initialization of weights and biases for all layers l, as $N \to \infty$ (large width limit). Again, here the index 1 does not matter by Axiom 1. Additionally, define **the cosine distance quantities** $\mathfrak{c}^{(l)} := \gamma^{(l)}/\sqrt{p^{(l)} p^{(l)\prime}}$ and $c^{(l)} := \lambda^{(l)}/\sqrt{q^{(l)} q^{(l)\prime}}$.

Definition 2.3. Fix an input $x^{(0)}$ and a gradient vector $(\partial E/\partial x_i^{(L)})_i$ of some loss function E with respect to the last layer $x^{(L)}$. Then define **the gradient quantities** $\daleth^{(l)} := \langle (\partial E/\partial x_1^{(l)})^2 \rangle, \chi_\bullet^{(l)} := \langle (\partial E/\partial \bullet_1^{(l)})^2 \rangle$ for $\bullet = a, b$, and $\chi_\bullet^{(l)} := \langle (\partial E/\partial \bullet_{11}^{(l)})^2 \rangle$ for $\bullet = w, v$.

[5] "projection" is in quotes because the dimension $N^{(l)}$ can be less than or greater than $N^{(l-1)}$; but we will follow the traditional wording

Table 1: Mean Field Dynamics with VV and WV.

Forward (Thm A.1)	Backward (Thm A.2)
$q = \sigma_w^2 l^{-\beta_w} \underline{p} + \sigma_b^2 l^{-\beta_b}$	$\daleth/\gimel = \begin{cases} (N/\underline{N})(\sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w} \mathrm{V}\dot{\phi}(0,q)+1) \\ (\sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w} \mathrm{V}\dot{\phi}(0,q)+1) \end{cases}$
$\Delta p = \sigma_v^2 l^{-\beta_v} \mathrm{V}\phi(0,q) + \sigma_a^2 l^{-\beta_a}$	$\chi_b = (N/M)\sigma_v^2 l^{-\beta_v} \gimel \mathrm{V}\dot{\phi}(0,q)$
$\lambda = \sigma_w^2 l^{-\beta_w} \lambda + \sigma_b^2 l^{-\beta_b}$	$\chi_w = (N/M)\sigma_v^2 l^{-\beta_v} \gimel \mathrm{V}\dot{\phi}(0,q)\underline{p}$
$\Delta\gamma = \sigma_v^2 l^{-\beta_v} \mathrm{W}\phi(0,q,q,\lambda) + \sigma_a^2 l^{-\beta_a}.$	$\chi_v = \gimel \mathrm{V}\phi(0,q) \quad \chi_a = \gimel \quad \chi_\pi = \gimel \underline{p}.$

The two cases for $\daleth/\gimel$ are respectively for a projection and a normal residual block, under the assumption that $\sigma_\pi = 1$.

Here the expectations are taken with Axiom 2 in mind, over both random initialization of forward and backward weights and biases, as $N \to \infty$ (large width limit). Again, the index 1 or 11 does not matter by Axiom 1.

Asymptotic notations. The expressions $f = O(g) \iff g = \Omega(f)$ have their typical meanings, and $f = \Theta(g)$ iff $f = O(g), g = O(f)$. We take $f(x) = \tilde{O}(g(x)) \iff g(x) = \tilde{\Omega}(f(x))$ to mean $f(x) = O(g \log^k x)$ for some $k \in \mathbb{Z}$ (this is slightly different from the standard usage of $\tilde{O}$), and $f = \tilde{\Theta}(g) \iff f = \tilde{O}(g)$ & $g = \tilde{O}(f)$. All asymptotic notations are sign-less, i.e. can indicate either positive or negative quantities, unless stated otherwise.

3 Theory

We recall integral transforms from Yang and Schoenholz (2017):

Definition 3.1. Define the transforms V and W by $\mathrm{V}\phi(\mu,\rho) := \mathrm{E}[\phi(z)^2 : z \sim \mathcal{N}(\mu,\rho)]$ and $\mathrm{W}\phi(\mu,\rho,\rho',\nu) := \mathrm{E}[\phi(z)\phi(z') : (z,z') \sim \mathcal{N}(\mu, \begin{pmatrix} \rho & \nu \\ \nu & \rho' \end{pmatrix})]$.

Table 1 summarizes the equations governing the dynamics of various mean field quantities when we vary variances and/or widths; they are proved similarly to how corresponding recurrences were proved in Yang and Schoenholz (2017); Schoenholz et al. (2016).

4 Width Variation

Width variation is first passingly mentioned in Schoenholz et al. (2016) as a potential way to guide gradient dynamics for feedforward networks. We develop a complete mean field theory of WV for residual networks.

Via Table 1 and Thm A.2, we see that width variation (WV) has two kinds of effects on the mean gradient norm: when compared to no width variation, it can multiply the squared gradient norm of biases b_i or weights w_{ij} by N/M (which doesn't "stack", i.e. does not affect the squared gradient norm of lower layers), or it can multiply the squared gradient norm of x_i by $N/\underline{N}$ (which "stacks", in the sense above, through the dynamics of $\gimel$). We will focus on the latter "stacking" effect and assume $N^{(l)} = M^{(l)}$

Suppose from layer l to layer m, $\gimel^{(m)}$ rises to r times $\gimel^{(l)}$. If we vary the width so that $N^{(m-1)}$ is $rN^{(m)}$, then this gradient expansion is canceled, and $\gimel^{(m-1)} = (N^{(m)}/N^{(m-1)})(\sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w} \mathrm{V}\dot{\phi}(0,q^{(m)})+1)\gimel^{(m)} = (\sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w} \mathrm{V}\dot{\phi}(0,q^{(m)})+1)\gimel^{(l)}$ so that it is as if we restarted backpropagation at layer m.

Remarkably, changing the width does not change the mean field forward dynamics (for example the recurrences for p, q, γ, λ remain the same) (Thm A.1). But, as always, if we reduce the width as part of WV (say, keeping $N^{(0)}$ the same but reducing the widths of later

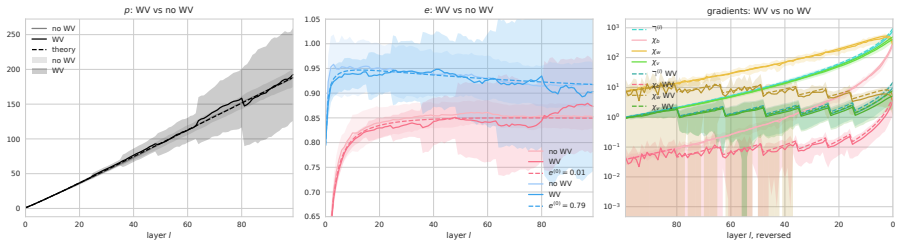

Figure 1: Using width decay to control the gradient explosion of a tanh residual network. Shades indicate standard deviation (taken in normal scale, but possibly displayed in log scale), while solid lines indicate the corresponding mean. The left two plots show that mean forward dynamics are more or less preserved, albeit variance explodes toward the deeper layers, when width decay is applied. The last plot show that the gradient dynamics is essentially suppressed to be a constant compared to the $\exp(\sqrt{L})$ dynamics of tanh resnet without width decay. Dashed lines indicate theoretical estimates in all three plot. The empirical data is generated from random residual networks with 100 layers and $N^{(0)} = 2048$, and we half the widths at layers $l = m^2$ for $m = 4, 5, \ldots, 9$.

layers), the variance of the sampled dynamics will also increase; if we increase the width as part of WV (say, keeping $N^{(L)}$ the same but increasing the widths of earlier layers), the variance of the sampled dynamics will decrease.

We can apply this theory of WV to tanh residual networks ($\phi = \tanh$ in Table 1) without VV. By Yang and Schoenholz (2017), tanh residual networks with all $\beta_\bullet = 0$ have gradient dynamics $\log(\daleth^{(m)}/\daleth^{(l)}) = \mathcal{A}(\sqrt{l} - \sqrt{m}) + \Theta(\log l - \log m)$ where $\mathcal{A} = \frac{4}{3}\sqrt{\frac{2}{\pi}}\frac{\sigma_v^2\sigma_w}{\sqrt{\sigma_v^2+\sigma_a^2}}$. If we place projection blocks projecting $N^{(l-1)} \mapsto N^{(l)}/\exp(\mathcal{A})$ at layer $l = n^2$, for $n = 1, 2, \ldots$, then the gradient norms would be bounded (above and below) across layers, as reasoned above. Indeed, this is what we see in Fig. 1. The rightmost subfigure compares, with log scale y-axis, the gradient dynamics with no WV to that with WV as described above, and we see that our theory tracks the mean gradient dynamics remarkably precisely for both the WV and the no WV cases, and indeed, WV effectively cap the gradient norm for $l \geq 16$ (where WV is applied). The left two figures show the forward dynamics of p and $\mathfrak{e}$, and we see that the WV does not affect the mean dynamics as predicted by theory. However, we also see dramatic increase in deviation from the mean dynamics at every projection layer in the forward case, while the deviation from the mean is constant across depth, but nevertheless is much larger than that in the case with no WV.

Therefore, width variation is unique in a few ways among all the techniques discussed in the mean field networks literature so far, including variance decay as studied below, adding skip connections, or changing activation functions: 1. It can ameliorate or suppress altogether gradient explosion (or vanishing) problems without affecting the mean forward dynamics of $p, q, \lambda, \gamma, c, \mathfrak{e}$. 2. To do so, it has to choose a trade off from the following spectrum: At one end, we truncate neurons from the original network (say, keeping $N^{(0)}$ the same), so that we have fewer parameters, faster training, but larger deviations from the mean dynamics. At the other, we add neuron to the original network (say, keeping $N^{(L)}$ the same), so that we have more parameters, slower training, and smaller deviations from the mean dynamics.

5 Variance Variation

As discuss in Section 4, WV is essentially a post-hoc technique to tweak an existing gradient dynamic without changing the forward dynamics. Thus in this section we assume all widths are constant, $N^{(l)} = N^{(m)} = M^{(n)}$ for any m, l, n, so that WV can be applied as a "touch-up" if necessary.

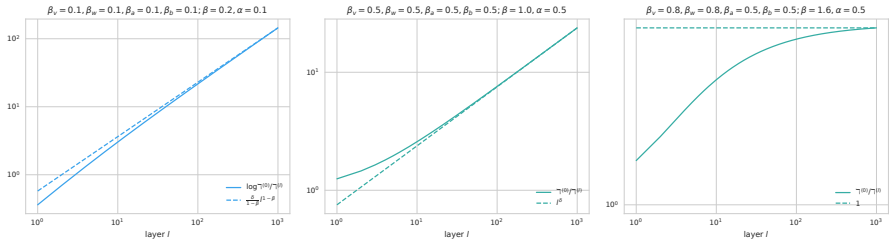

Figure 2: We verify the asymptotic characterizations of $\daleth$ given by Thm A.8. Here solid lines indicate computations done via the recurrences Table 1, while dashed lines indicate asymptotics proved in Thm A.8. To facilitate visualization, all green dashed lines are shifted vertically to match the end point of the corresponding solid lines. From left to right: **(a)** log-log plot of the *log* of $\daleth^{(0)}/\daleth^{(l)}$, against the predicted leading term $\frac{\delta}{1-\beta} l^{1-\beta}$, when $\beta = .2$; thus $\daleth^{(0)}/\daleth^{(l)}$ is superpolynomial. As in Thm A.8, $\delta = \frac{1}{2}\sigma_v^2\sigma_w^2$. **(b)** log-log plot of $\daleth^{(0)}/\daleth^{(l)}$ when $\beta = 1$, showing it is polynomial. **(c)** $\daleth^{(0)}/\daleth^{(l)}$ is bounded above when $\beta = 1.6$.

5.1 Residual Network with ReLU

It was shown in Yang and Schoenholz (2017) that, with no variance decay, in an ReLU resnet, both the mean squared activation norm (p) and the mean squared gradient norm ($\daleth$) explode exponentially with depth, and this causes training to fail for even 100 layer networks. We show that this problem is in fact extremely easy to fix, requiring no architectural changes at all, only that $\beta_\bullet$ be increased from 0 so that the randomization variances decay across depth (Thms A.6 and A.8).

Gradient quantities. The main driver of this gradient mollification is $\beta := \beta_v + \beta_w$. When $\beta \in [0, 1)$, the gradient norm varies like $\daleth^{(0)}/\daleth^{(l)} = \exp(\Theta(l^{1-\beta}))$; when $\beta = 0$ this recapitulate the exponential behavior derived in Yang and Schoenholz (2017). When $\beta = 1$, it experiences a sharp phase transition, where now $\daleth^{(0)}/\daleth^{(l)} = \mathsf{poly}(l)$. As β becomes larger than 1, $\daleth^{(0)}/\daleth^{(l)} = \Theta(1)$, all bounded! Fig. 2 verifies this result empirically, and in fact show that our computed asymptotic expansions in Thm A.8 are highly accurate predictors. It is both interesting and important to note that the gradient norms for actual trainable parameters, such as χ_w and χ_v, are affected differently by β. In fact, $\chi_w^{(1)}/\chi_w^{(l)}$ is *bounded*(!) with l when $\beta < 1$ (the $\beta = 0$ case is already observed in Yang and Schoenholz (2017)) but phase transitions to $\mathsf{poly}(l)$ for $\beta \geq 1$, while $\chi_v^{(1)}/\chi_v^{(l)}$ is already $\mathsf{poly}(l)$ when $\beta < 1$, and remains so as β is increased to > 1.. Curiously, χ_v turns out to be more predictive of performance than $\daleth$ in the $\beta > 1$ regime, in a surprising way; see Section 6.1.

Length quantities. Similarly, β is the primary conduit for mollifying the behavior of squared activation norms p and q (Thm A.6). Like the gradient dynamics, $p = \exp(\Theta(l^{1-\beta}))$ when $\beta < 1$; when $\beta = 0$ this recapitulates the results of Yang and Schoenholz (2017). As β rise to 1 and above, p experiences a phase transition into polynomial dynamics, but unlike the case of $\daleth$, it is not constant when $\beta > 1$. Instead, a different parameter, $\alpha := \min(\beta_v + \beta_b, \beta_a)$ drives the asymptotics of p in the $\beta > 1$ regime. When $\alpha \in [0, 1)$ is small, p grows like $\Theta(l^{1-\alpha})$. The instant α hits 1, p is just logarithmic, $p = \Theta(\log l)$. As α shoots past 1, p becomes constant. Thus the dynamics of p is governed by a zigzag through $(\beta_\bullet)_{\bullet=v,w,a}$ space. Fig. 3 goes through each of the five cases discussed above and verifies that our asymptotics are correct.

Cosine distance. $\mathfrak{c} = \gamma/p$ measures how well the input space geometry (angles, in this case) is preserved as the input space is propagated through each layer. Its dynamics are much simpler than those of p and $\daleth$ above. If $\mathfrak{c}^{(0)} = 1$, then $\mathfrak{c}^{(l)} = 1$ for all l trivially. If $\mathfrak{c}^{(0)} < 1$, then we have one of the following two cases, • If $\beta \leq 1$ or $\alpha \leq 1$, then $\mathfrak{c} \to 1$ irrespective of initial data $p^{(0)}$ and $\gamma^{(0)}$. • If $\beta > 1$ and $\alpha > 1$, then $\mathfrak{c}$ converges to a fixed point $\mathfrak{c}^* < 1$ which depends on the initial data $p^{(0)}$ and $\gamma^{(0)}$, at a rate of $\Theta(l^{1-\alpha})$. Thus

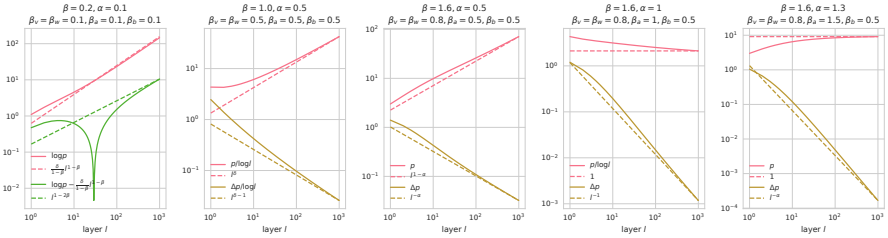

Figure 3: We verify the asymptotic characterizations of p given by Thm A.6. Here solid lines indicate computations done via the recurrences Table 1, while dashed lines indicate asymptotics proved in Thm A.6. In all but the leftmost plot, we show both p and $\Delta p = p - \underline{p}$ (possibly adjusted for log factors) and their asymptotics. To facilitate visualization, all dashed lines except the red one in the leftmost plot are shifted vertically to match the end point of the corresponding solid lines. In the leftmost plot, the red lines are respectively $\log p$ (solid) and the leading term $\frac{\delta}{1-\beta} l^{1-\beta}$ in its asymptotic expansion (dashed). The green lines in the same plot are the remainder (solid) and its polynomial asymptotic form, ignoring the leading coefficient (dashed). From left to right gives one example for each of the 5 cases discussed in the main text: **(a)** $\beta < 1$ and $p = \exp(\mathsf{poly}(l))$. **(b)** $\beta = 1$ and p phase transitions into polynomial dynamics. **(c)** $\beta > 1, \alpha < 1$ and $p = \mathsf{poly}(l)$. **(d)** $\beta > 1, \alpha = 1$ and $p = \Theta(\log l)$. **(e)** $\beta > 1, \alpha > 1$ and p becomes bounded but $\Delta p = \Theta(l^{-\alpha})$.

ReLU *very much* likes to collapse the input space into a single point ($\mathfrak{c} = 1$ means every two input vectors get mapped to the same output vector), and the only way to prevent this from happening is to make the $\beta_\bullet$s so high, that higher layers barely modifies the computation done by lower layers at all. Indeed, the second condition $\beta > 1$ and $\alpha > 1$ ensures that $p = \Theta(1)$ as discussed above (Thm A.6), so that as $l \to \infty$, layer l's residual adds to $x^{(l-1)}$ only a vector of vanishing size compared to the size of $x^{(l-1)}$.

5.2 Residual Network with Tanh

While ReLU resnet depends heavily on $\beta = \beta_v + \beta_w$ and $\alpha = \min(\beta_v + \beta_b, \beta_a)$, $\hat{\mu}_p := \min(\beta_v, \beta_a)$ and $\hat{\nu}_p := 1 - \beta_w + \hat{\mu}_p - 2\beta_v$ are the key quantity determining the dynamics in the case of tanh resnet, with β playing a minor role. We will study tanh resnets in the setting where $q \to \infty$; otherwise, p is bounded, meaning that higher layers become essentially unused by the neural network (similar to the discussion made in Section 5.1(Cosine distance) above). In this setting, it can be shown that $\hat{\mu}_p \leq 1$ (Thm A.3).

Gradient quantities. By Thm A.4, as long as $\hat{\mu}_p$ stays below 1, the asymptotics of $\daleth$ is entirely governed by $\hat{\nu}_p$, which is 1 when all $\beta_\bullet$s are 0 and most of the time decreases as $\beta_\bullet$s are increased. When $\hat{\nu}_p > 0$, $\daleth^{(0)}/\daleth^{(l)} = \exp(\Theta(l^{\frac{1}{2}\hat{\nu}_p}))$; the results of Yang and Schoenholz (2017) are recovered by setting all $\beta_\bullet$s to 0, thus $\hat{\nu}_p = 1$ and $\daleth^{(0)}/\daleth^{(l)} = \exp(\Theta(l^{\frac{1}{2}}))$. When $\hat{\nu}_p$ hits 0, $\daleth^{(0)}/\daleth^{(l)}$ becomes polynomial, and as $\hat{\nu}_p$ dips below 0, gradient expansion becomes bounded.

If $\hat{\mu}_p = 1$, $\daleth^{(0)}/\daleth^{(l)}$ is automatically suppressed to be subpolynomial. The only minor phase transition here is going from $\beta = 1$ to $\beta > 1$ (and β cannot be less than 1 by our assumption that $q \to \infty$). In the former case, the gradient expansion is $\exp(\Theta(\sqrt{\log l}))$, while in the latter case it is bounded.

Length quantities have simpler asymptotics determined by $\hat{\mu}_p$: either $\hat{\mu}_p < 1$ and $p = \Theta(l^{1-\hat{\mu}_p})$, or $\hat{\mu}_p = 1$ and $p = \Theta(\log l)$ (Thm A.3). **Cosine distance,** unlike the case of ReLU resnets, can be controlled effectively by β_a and β_v (Thm A.5). When $\beta_a > \beta_v$, the magnitude of $a_i^{(l)}$ drops much more quickly with depth than that of $v_{ij}^{(l)}$, so that higher layers experience the chaotic phase (Schoenholz et al., 2016; Yang and Schoenholz, 2017), driving $\mathfrak{c}^{(l)}$ toward the limit point $\mathfrak{c}^* = 0$. On the other end, when $\beta_a < \beta_v$, $v_{ij}^{(l)}$ vanishes in comparison to $a_i^{(l)}$ with large l, so that the higher layers experience the stability phase (Schoenholz et al., 2016; Yang and Schoenholz, 2017), collapsing all inputs to the same output vector, sending

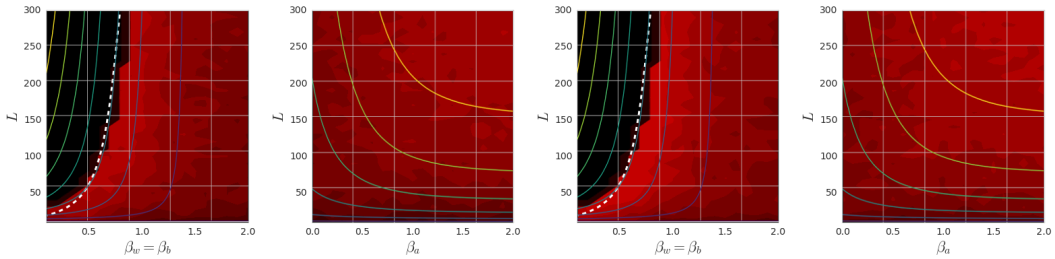

Figure 4: Zigzagging through $\beta_\bullet$-space with ReLU resnet. In each pair, plots indicate performance as $\beta_w = \beta_b$ is varied (the **zig**, left) and, with $\beta_w = \beta_b$ fixed at 2, as β_a is varied (the **zag**, right). The left pair corresponds to test set performance; the right pair corresponds to training set performance. Brighter color means higher accuracy, for both the heatmaps and overlaid contour plots.

$\mathfrak{e}^{(l)} \to 1$. Only when $\beta_a = \beta_v$ could the fixed point $\mathfrak{e}^*$ be controlled explicitly by σ_v and σ_a, with $\mathfrak{e}^*$ given by the equation $\mathfrak{e}^* = \frac{\sigma_v^2 \frac{2}{\pi} \arcsin(\mathfrak{e}^*) + \sigma_a^2}{\sigma_v^2 + \sigma_a^2}$ as in the case with no VV (Yang and Schoenholz, 2017).

6 Experiments

6.1 ReLU Residual Network: Zigzagging through $\beta_\bullet$-space

As explained in Section 5.1, a zigzag of parameters control various asymptotics of ReLU resnet: "zigging" β from 0 to > 1, and then "zagging" α from 0 to > 1. To investigate how this zigzag affects neural network training and performance, we trained a grid of ReLU residual networks, pinning all $\sigma_\bullet$s at 1, but varying the $\beta_\bullet$s as follows: 1. The zig: fix $\beta_v = \beta_a = 0$; fix $\beta_w = \beta_b$ and increase both from 0 to 2 (making $\beta = \beta_w + \beta_v$ go from 0 to 2 as well); 2. The zag: fix $\beta_v = \beta_a = 0, \beta_w = \beta_b = 2$; increase β_a from 0 to 2 (increasing α from 0 to 2 as well). For each setting of the hyperparameters, we train a network on MNIST with those hyperparameters for 30 epochs. We then report the accuracy that the network achieved on the training and test sets [6]. Fig. 4 shows the result, with the zig on the left and the zag on the right. In the zig, we have overlaid a contour plot of $p - \gamma$, which is almost identical to the contour plots of p and $\daleth^{(0)}/\daleth^{(l)}$, and for simplicity's sake, in the following we assume that they are exactly identical. This contour decreases to the right. The dashed line is given by $\frac{\delta}{1-\beta} l^{1-\beta} = \mathcal{C}$ for $\mathcal{C} \approx 10$. In the zag, we have overlaid a contour plot of $\chi_v^{(1)}/\chi_v^{(l)}$. This contour increases to the upper right.

Discussion. First notice that training fails in the upper left corner of the zag. This happens because of numerical instability caused by exploding p and $\daleth^{(0)}/\daleth^{(l)}$, which grow like $\exp(\frac{\delta}{1-\beta} l^{1-\beta} + o(l^{1-\beta}))$. Indeed, one of the contour lines of p traces out almost exactly where training fails. The dashed line is a level set of the dominant term of asymptotic expansion of $\log p$, and we see it agrees with the contour of p very well.

Second, observe that performance actually dips in the direction where $\daleth^{(0)}/\daleth^{(l)}$ decreases, quite counterintuitively. This can be explained by noting that gradient against weights, χ_w and χ_v, in fact respectively remain bounded and polynomial in l (and changes rather mildly with β); gradient against biases do experience the same behavior as $\daleth$, but in general they are much less important than the weights, as parameters go. In addition, the performance is also dipping in the direction where $p - \gamma$ decreases (exponentially) in (β, L)-space. This is the quantity that essentially underlies the exponential expressivity (as told from a extrinsic curvature perspective) result of Poole et al. (2016); as $p - \gamma$ decreases dramatically, it gets harder and harder for a linear functional in the final layer to tell apart two input vectors. This exponential loss in expressivity dominates the effect on performance more than a polynomial reduction in gradient explosion.

[6] Thus we are investigating a question of training speed rather than one of final generalization accuracy

Third, it is remarkable that in the zag regime, the level curves of $\chi_v^{(1)}/\chi_v^{(l)}$ accurately predict the contour of the test set performance, in such a counterintuitive way that greater gradient explosion $\chi_v^{(1)}/\chi_v^{(l)}$ correlates with better performance. It is currently unclear if in certain situations like this, larger gradient expansion is actually beneficial, or if there is a quantity yet undiscovered which has the same level curves and can explain away this seeming paradox (like how $p - \gamma$ explains away $\daleth^{(0)}/\daleth^{(l)}$ above, in the zig regime).

7 Conclusion

In this paper, we derived the mean field theory of width and variance variation and showed that they are powerful methods to control forward (VV) and backward (VV + WV) dynamics. We proved that even with a fixed architecture and activation function, the mean field dynamics of a residual neural network can still be manipulated at will by these two methods. Extraordinarily, the mean field theory we developed allowed us to accurately predict the performances of trained MNIST models on test sets, but one puzzling aspect remains where test set performance seems to increase as gradient explosion worsens.

Open Problems. We solved a small part, width variation, of the program to construct mean field theories of state-of-the-art neural networks used in practice. Many open problems still remain, and the most important of them include but is not limited to 1. batchnorm, 2. convolution layers, and 3. recurrent layers. In addition, more work is needed to mathematically justify our "physical" assumptions Axiom 1 and Axiom 2 to a "math" problem. We hope readers will take note and contribute toward deep mean field theory.

References

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *arXiv:1602.05897 [cs, stat]*, February 2016. URL http://arxiv.org/abs/1602.05897. arXiv: 1602.05897.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv: 1512.03385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. *arXiv:1603.05027 [cs]*, March 2016. URL http://arxiv.org/abs/1603.05027. arXiv: 1603.05027.

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random feedback weights support learning in deep neural networks. *arXiv:1411.0247 [cs, q-bio]*, November 2014. URL http://arxiv.org/abs/1411.0247. arXiv: 1411.0247.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *arXiv:1606.05340 [cond-mat, stat]*, June 2016. URL http://arxiv.org/abs/1606.05340. arXiv: 1606.05340.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. *arXiv:1606.05336 [cs, stat]*, June 2016. URL http://arxiv.org/abs/1606.05336. arXiv: 1606.05336.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. *arXiv:1611.01232 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.01232. arXiv: 1611.01232.

Greg Yang and Samuel S. Schoenholz. Meanfield Residual Network: On the Edge of Chaos. In *Advances in neural information processing systems*, 2017.

Appendices

A Main Theorems

A.1 Recurrences

Yang and Schoenholz (2017) gave recurrences for mean field quantities $p, q, \gamma, \lambda, \daleth$ under the assumption of constant initialization variances across depth. The proofs there carry over straightforwardly when variance varies from layer to layer. Schoenholz et al. (2016) also derived backward dynamics for when the width of the network is not constant. Generalizing to the residual network case requires some careful justifications of independences, so we provide proof for gradient dynamics; but we omit the proof for the forward dynamics as it is not affected by nonconstant width and is almost identical to the constant variance case.

Theorem A.1. *For any nonlinearity ϕ in an FRN, regardless of whether widths vary across layers,*

$$\begin{aligned}
q^{(l)} &= \sigma_w^2 l^{-\beta_w} p^{(l-1)} + \sigma_b^2 l^{-\beta_b} \\
p^{(l)} &= \sigma_v^2 l^{-\beta_v} \mathrm{V}\phi(0, q^{(l)}) + \sigma_a^2 l^{-\beta_a} + p^{(l-1)} \\
\lambda^{(l)} &= \sigma_w^2 l^{-\beta_w} \lambda^{(l-1)} + \sigma_b^2 l^{-\beta_b} \\
\gamma^{(l)} &= \sigma_v^2 l^{-\beta_v} \mathrm{W}\phi(0, q^{(l)}, q^{(l)\prime}, \lambda^{(l)}) + \sigma_a^2 l^{-\beta_a} + \gamma^{(l-1)}.
\end{aligned}$$

Theorem A.2. *Suppose a random residual network receives a fixed gradient vector $\left(\partial E/\partial x_i^{(L)}\right)_{i=1}^{N^{(l)}}$ with respect to some cost function E, at its last layer. For any nonlinearity ϕ in an FRN, under Axiom 1 and Axiom 2, whenever $\dot{\phi}(\zeta)^2$ has finite variance for any Gaussian variable ζ,*

$$\begin{aligned}
\beth &= \begin{cases} ({}^N/_{\underline{N}})(\sigma_v^2 \sigma_w^2 l^{-\beta_v - \beta_w} \mathrm{V}\dot{\phi}(0, q) + \sigma_\pi^2)\daleth & \textit{if layer } l \textit{ has projection connection} \\ (\sigma_v^2 \sigma_w^2 l^{-\beta_v - \beta_w} \mathrm{V}\dot{\phi}(0, q) + 1)\daleth & \textit{if layer } l \textit{ has identity connection} \end{cases} \\
\chi_b &= ({}^N/_M)\sigma_v^2 l^{-\beta_v} \daleth \mathrm{V}\dot{\phi}(0, q) \\
\chi_w &= ({}^N/_M)\sigma_v^2 l^{-\beta_v} \daleth \mathrm{V}\dot{\phi}(0, q)\underline{p} \\
\chi_v &= \daleth \mathrm{V}\phi(0, q) \\
\chi_a &= \daleth \\
\chi_\pi &= \daleth \underline{p}.
\end{aligned}$$

Proof. We will show the proof for the projection connection case; the identity connection case is similar but easier.

Write $\beth_i^{(l)} = \frac{\partial E}{\partial x_i^{(l)}}$. We have the following derivative computations:

$$\frac{\partial x_i}{\partial \underline{x}_j} = \pi_{ij} + \sum_k v_{ik} \dot{\phi}(h_k) \frac{\partial h_k}{\partial \underline{x}_j}, \quad \frac{\partial x_i}{\partial h_j} = v_{ij} \dot{\phi}(h_j), \quad \frac{\partial h_i}{\partial \underline{x}_j} = w_{ij}, \quad \frac{\partial h_i}{\partial w_{ij}} = \underline{x}_j, \quad \frac{\partial h_i}{\partial b_i} = 1,$$
$$\frac{\partial x_i}{\partial v_{ik}} = \phi(h_k), \quad \frac{\partial x_i}{\partial a_i} = 1, \quad \frac{\partial x_i}{\partial \pi_{ij}} = \underline{x}_j$$

Then

$$\begin{aligned}
\underline{\beth}_j &= \sum_{i=1}^N \beth_i (\pi_{ij} + \sum_{k=1}^M v_{ik} \dot{\phi}(h_k) \frac{\partial h_k}{\partial \underline{x}_j}) \\
&= \sum_{i=1}^N \beth_i (\pi_{ij} + \sum_{k=1}^M v_{ik} \dot{\phi}(h_k) w_{kj})
\end{aligned}$$

Thus,

$$
\begin{aligned}
\langle \beth_j^2 \rangle &= \langle [\sum_{i=1}^{N} \beth_i (\pi_{ij} + \sum_{k=1}^{M} v_{ik} \dot{\phi}(h_k) w_{kj})]^2 \rangle \\
&= \langle \sum_{i=1}^{N} \beth_i^2 [(\pi_{ij} + \sum_{k=1}^{M} v_{ik} \dot{\phi}(h_k) w_{kj})]^2 \rangle \\
&\quad + 2 \sum_{i<i'} \langle \beth_i \beth_{i'} (\pi_{ij} + \sum_{k=1}^{M} v_{ik} \dot{\phi}(h_k) w_{kj})(\pi_{i'j} + \sum_{k=1}^{M} v_{i'k} \dot{\phi}(h_k) w_{kj}) \rangle \\
&= N \langle \beth_1^2 \rangle (\langle \pi_{11}^2 \rangle + \langle [\sum_{k=1}^{M} v_{1k} \dot{\phi}(h_k) w_{kj}]^2 \rangle) \\
&\quad + 2 \sum_{i<i'} \langle \beth_i \beth_{i'} \rangle \langle (\pi_{ij} + \sum_{k=1}^{M} v_{ik} \dot{\phi}(h_k) w_{kj})(\pi_{i'j} + \sum_{k=1}^{M} v_{i'k} \dot{\phi}(h_k) w_{kj}) \rangle
\end{aligned}
$$

where in the second equality, we expanded algebraically, and in the third equality, we use the symmetry assumption Axiom 1 and the independence assumption Axiom 2. Now, $\langle [\sum_{k=1}^{M} v_{1k} \dot{\phi}(h_k) w_{kj}]^2 \rangle = \sum_{k=1}^{M} \langle v_{1k}^2 \dot{\phi}(h_k)^2 w_{kj}^2 \rangle = \sum_{k=1}^{M} \frac{\sigma_v^2 l^{-\beta_v}}{M} \mathrm{V}\dot{\phi}(0,q) \frac{\sigma_w^2 l^{-\beta_w}}{\underline{N}} = \sigma_v^2 l^{-\beta_v} \mathrm{V}\dot{\phi}(0,q) \frac{\sigma_w^2 l^{-\beta_w}}{\underline{N}}$ by the independence of $\{v_{ik}\}_{i,k} \cup \{h_k\}_k \cup \{w_{kj}\}_{k,j}$ (by our independence assumptions Axiom 2). Similarly, because $\{\pi_{ij}\}_j \cup \{\pi_{i'j}\}_j \cup \{v_{ik}\}_k \cup \{v_{i'k}\}_k$ for $i \neq i'$ is mutually independent by our assumptions, one can easily see that $\langle (\pi_{ij} + \sum_{k=1}^{M} v_{ik} \dot{\phi}(h_k) w_{kj})(\pi_{i'j} + \sum_{k=1}^{M} v_{i'k} \dot{\phi}(h_k) w_{kj}) \rangle = 0$.

Therefore,

$$
\begin{aligned}
\beth &= N \daleth (\frac{\sigma_\pi^2}{\underline{N}} + \sigma_v^2 l^{-\beta_v} \mathrm{V}\dot{\phi}(0,q) \frac{\sigma_w^2 l^{-\beta_w}}{\underline{N}}) \\
&= \frac{N}{\underline{N}} \daleth (\sigma_\pi^2 + \sigma_v^2 \sigma_w^2 l^{-\beta_v - \beta_w} \mathrm{V}\dot{\phi}(0,q)).
\end{aligned}
$$

For the other gradients, we have (where we apply Axiom 2 implicitly)

$$
\begin{aligned}
\frac{\partial E}{\partial b_j} &= \sum_{i=1}^{N} \frac{\partial E}{\partial x_i} \frac{\partial x_i}{\partial h_j} \frac{\partial h_j}{\partial b_j} = \sum_{i=1}^{N} \beth_i v_{ij} \dot{\phi}(h_j) \\
\chi_b &= \langle [\sum_{i=1}^{N} \beth_i v_{ij} \dot{\phi}(h_j)]^2 \rangle = N \daleth \frac{\sigma_v^2 l^{-\beta_v}}{M} \mathrm{V}\dot{\phi}(0,q) = \frac{N}{M} \sigma_v^2 l^{-\beta_v} \daleth \mathrm{V}\dot{\phi}(0,q); \\
\frac{\partial E}{\partial w_{ji}} &= \sum_{i=1}^{N} \frac{\partial E}{\partial x_i} \frac{\partial x_i}{\partial h_j} \frac{\partial h_j}{\partial w_{ji}} = \sum_{i=1}^{N} \beth_i v_{ij} \dot{\phi}(h_j) \underline{x}_j \\
\chi_w &= \langle \sum_{i=1}^{N} \beth_i v_{ij} \dot{\phi}(h_j) \underline{x}_j \rangle = N \daleth \frac{\sigma_v^2 l^{-\beta_v}}{M} \mathrm{V}\dot{\phi}(0,q) \underline{p} = \frac{N}{M} \sigma_v^2 l^{-\beta_v} \daleth \mathrm{V}\dot{\phi}(0,q) \underline{p} \\
\frac{\partial E}{\partial v_{ik}} &= \frac{\partial E}{\partial x_i} \frac{\partial x_i}{\partial v_{ik}} = \beth_i \phi(h_k) \implies \chi_v = \langle [\beth_i \phi(h_k)]^2 \rangle = \daleth \mathrm{V}\phi(0,q) \\
\frac{\partial E}{\partial a_i} &= \frac{\partial E}{\partial x_i} \frac{\partial x_i}{\partial a_i} = \frac{\partial E}{\partial x_i} \implies \chi_a = \daleth \\
\frac{\partial E}{\partial \pi_{ij}} &= \frac{\partial E}{\partial x_i} \frac{\partial x_i}{\partial \pi_{ij}} = \beth_i \underline{x}_j \implies \chi_\pi = \daleth \underline{p}
\end{aligned}
$$

□

A.2 Residual Networks with Tanh

In this section we derive the asymptotics of various mean field quantities for tanh resnet. The main proof technique is to bound the dynamics in question with known dynamics of difference equations (as in Yang and Schoenholz (2017)).

Let $\hat{\mu}_p := \min(\beta_v, \beta_a)$.

Theorem A.3. *Suppose $\phi = \tanh$, and $q^{(l)} \to \infty$ as $l \to \infty$.*

1. *If $\hat{\mu}_p < 1$, then $1 > \beta_w + \hat{\mu}_p$ and*
$$p^{(l)} = \Theta(l^{1-\hat{\mu}_p})$$
$$q^{(l)} = \Theta(l^{1-\beta_w-\hat{\mu}_p}).$$
More specifically,
$$(p^{(l)}, q^{(l)}) \sim \begin{cases} (\sigma_v^2 l^{1-\beta_v}, \sigma_v^2\sigma_w^2 l^{1-\beta_v-\beta_w}) & \text{if } \beta_v < \beta_a \\ (\sigma_a^2 l^{1-\beta_a}, \sigma_a^2\sigma_w^2 l^{1-\beta_a-\beta_w}) & \text{if } \beta_a < \beta_v \\ ((\sigma_a^2+\sigma_v^2) l^{1-\beta_v}, (\sigma_a^2+\sigma_v^2)\sigma_w^2 l^{1-\beta_v-\beta_w}) & \text{if } \beta_a = \beta_v. \end{cases}$$

2. *If $\hat{\mu}_p = 1$, then $\beta_w = 0$, and $p^{(l)} = \Theta(\log l) = q^{(l)}$. More specifically,*
$$(p^{(l)}, q^{(l)}) \sim \begin{cases} (\sigma_v^2, \sigma_w^2\sigma_v^2 \log l) & \text{if } \beta_v < \beta_a \\ (\sigma_a^2 \log l, \sigma_w^2\sigma_a^2 \log l) & \text{if } \beta_v > \beta_a \\ ((\sigma_a^2+\sigma_v^2)\log l, \sigma_w^2(\sigma_a^2+\sigma_v^2)\log l) & \text{if } \beta_v = \beta_a. \end{cases}$$

3. *$\hat{\mu}_p$ cannot be greater than 1.*

Proof. **Claim 1.** We have $\mathrm{V}\phi(0,q) = 1 - \sqrt{\frac{2}{\pi}} q^{-1/2} + \Theta(q^{-3/2})$ by Lemma B.5. Thus
$$\begin{aligned} p - \underline{p} &= \sigma_v^2 l^{-\beta_v}(1 - \sqrt{2/\pi} q^{-1/2} + \Theta(q^{-3/2})) + \sigma_a^2 l^{-\beta_a} \\ &= \sigma_v^2 l^{-\beta_v} + \sigma_a^2 l^{-\beta_a} - o(l^{-\beta_v}) \end{aligned}$$
$$p \sim \begin{cases} \sigma_v^2 l^{1-\beta_v} & \text{if } \beta_v < \beta_a \\ \sigma_a^2 l^{1-\beta_a} & \text{if } \beta_a < \beta_v \\ (\sigma_a^2+\sigma_v^2) l^{1-\beta_v} & \text{if } \beta_a = \beta_v \end{cases}$$
where we used the assumption $1 - \hat{\mu}_p > 0$. Thus $p^{(l)} = \Theta(l^{1-\hat{\mu}_p})$ and goes to infinity with l.

Then $q = \sigma_w^2 l^{-\beta_w} \underline{p} + \sigma_b^2 l^{-\beta_b}$. Because we assumed $q \to \infty$, the first term necessarily dominates the second, and $1 - \hat{\mu}_p - \beta_w > 0$. The possible asymptotics of q are then
$$q \sim \begin{cases} \sigma_v^2\sigma_w^2 l^{1-\beta_v-\beta_w} & \text{if } \beta_v < \beta_a \\ \sigma_a^2\sigma_w^2 l^{1-\beta_a-\beta_w} & \text{if } \beta_v > \beta_a \\ (\sigma_a^2+\sigma_v^2)\sigma_w^2 l^{1-\beta_v-\beta_w} & \text{if } \beta_a = \beta_v \end{cases}$$
which gives $q^{(l)} = \Theta(l^{1-\beta_w-\hat{\mu}_p})$.

Claim 2. If $\hat{\mu}_p = 1$, then
$$p \sim \begin{cases} \sigma_v^2 \log l & \text{if } \beta_v < \beta_a \\ \sigma_a^2 \log l & \text{if } \beta_v > \beta_a \\ (\sigma_a^2+\sigma_v^2)\log l & \text{if } \beta_v = \beta_a. \end{cases}$$
Then for q to go to infinity, β_w has to be 0, so that $q = \Theta(\log l)$ as well, and
$$q \sim \begin{cases} \sigma_w^2\sigma_v^2 \log l & \text{if } \beta_v < \beta_a \\ \sigma_w^2\sigma_a^2 \log l & \text{if } \beta_v > \beta_a \\ \sigma_w^2(\sigma_a^2+\sigma_v^2)\log l & \text{if } \beta_v = \beta_a. \end{cases}$$

Claim 3. If $\hat{\mu}_p > 1$, then $p = \Theta(1) \implies q = \Theta(1)$, a contradiction. □

Theorem A.4. *Suppose $\phi = \tanh$, and $q^{(l)} \to \infty$ with l. Then $\log(\daleth^{(0)}/\daleth^{(l)})$ is*

If $\hat{\mu}_p < 1$

If $1 - \beta_w + \hat{\mu}_p - 2\beta_v > 0$
$\Theta(l^{\frac{1}{2}(1-\beta_w+\hat{\mu}_p-2\beta_v)})$
If $1 - \beta_w + \hat{\mu}_p - 2\beta_v = 0$
$\Theta(\log l)$
If $1 - \beta_w + \hat{\mu}_p - 2\beta_v < 0$
$\Theta(1)$

If $\hat{\mu}_p = 1$

If $\beta_v + \beta_w < 1$
This case cannot happen
If $\beta_v + \beta_w = 1$
$\Theta(\sqrt{\log l})$
If $\beta_v + \beta_w > 1$
$\Theta(1)$

Proof. By Thm A.2 and Lemma B.4, we have

$$\begin{aligned}\log(\beth/\daleth) &= \log(1 + \sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w}\mathrm{V}\dot{\phi}(0,q)) \\ &= \log(1 + \sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w}(Cq^{-1/2} + \Theta(q^{-3/2}))) \\ &= \sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w} Cq^{-1/2} + \Theta(l^{-2\beta_v-2\beta_w})\end{aligned}$$

where $C = \frac{2}{3}\sqrt{\frac{2}{\pi}}$. Since $q^{(l)}$ has different growth rates depending on the hyperparameters, we need to consider different cases:

- If $\hat{\mu}_p < 1$, then Thm A.3 implies $\beta_w + \hat{\mu}_p < 1$ and $q = \Theta(l^{1-\beta_w-\hat{\mu}_p})$. So $l^{-\beta_v-\beta_w}q^{-1/2} = l^{-\beta_v-\beta_w}\Theta(l^{-\frac{1}{2}(1-\beta_w-\hat{\mu}_p)}) = \Theta(l^{\frac{1}{2}(-1-\beta_w+\hat{\mu}_p-2\beta_v)})$.
  - If $1 - \beta_w + \hat{\mu}_p - 2\beta_v > 0$, $\log \daleth^{(0)}/\daleth^{(l)} = \sum_{m=1}^{l}\Theta(m^{\frac{1}{2}(-1-\beta_w+\hat{\mu}_p-2\beta_v)}) = \Theta(l^{\frac{1}{2}(1-\beta_w+\hat{\mu}_p-2\beta_v)})$.
  - If $1 - \beta_w + \hat{\mu}_p - 2\beta_v = 0$, then $l^{-\beta_v-\beta_w}q^{-1/2} = \Theta(l^{-1})$. Thus $\log \daleth^{(0)}/\daleth^{(l)} = \sum_{m=1}^{l}\Theta(m^{-1}) = \Theta(\log l)$.
  - If $1-\beta_w+\hat{\mu}_p-2\beta_v < 0$, then $\frac{1}{2}(-1 - \beta_w + \hat{\mu}_p - 2\beta_v) > 1$, so that $\log \daleth^{(0)}/\daleth^{(l)} = \Theta(1)$.
- If $\hat{\mu}_p = 1$, then by Thm A.3, $\beta_w = 0$ and $q = \Theta(\log l)$. So $l^{-\beta_v-\beta_w}q^{-1/2} = \Theta(l^{-\beta_v-\beta_w}/\sqrt{\log l})$.
  - If $\beta_v + \beta_w < 1$, then $\beta_v < 1 \implies \hat{\mu}_p < 1$, contradiction.
  - If $\beta_v + \beta_w = 1$, then $\log \daleth^{(0)}/\daleth^{(l)} = \sum_{m=1}^{l}\Theta(l^{-1}/\sqrt{\log l}) = \Theta(\sqrt{\log l})$.
  - If $\beta_v + \beta_w > 1$, then $\log \daleth^{(0)}/\daleth^{(l)} = \Theta(1)$.

□

Theorem A.5. *Suppose $\phi = \tanh$ and $q \to \infty$. If $\mathfrak{e}^{(0)} < 1$, then $\mathfrak{e}^{(l)}$ converges to a fixed point $\mathfrak{e}^*$, given by the equations*

$$\mathfrak{e}^* = \begin{cases} 0 & \text{if } \beta_a > \beta_v \\ \frac{\sigma_v^2\frac{2}{\pi}\arcsin(\mathfrak{e}^*)+\sigma_a^2}{\sigma_v^2+\sigma_a^2} & \text{if } \beta_a = \beta_v \\ 1 & \text{if } \beta_a < \beta_v. \end{cases}$$

Note that in the case $\beta_a = \beta_v$, we recover the fixed point of tanh residual network without variance decay.

Proof. We have

$$\begin{aligned}
\sigma_v^2 l^{-\beta_v} \mathrm{W}\phi(0,q,q,cq) + \sigma_a^2 l^{-\beta_a} &= \gamma - \underline{\gamma} \\
&= \mathfrak{e}p - \underline{\mathfrak{e}}\underline{p} \\
&= \mathfrak{e}p - \underline{\mathfrak{e}}p + \underline{\mathfrak{e}}p - \underline{e}e\underline{p} \\
&= (\mathfrak{e} - \underline{\mathfrak{e}})p + \underline{\mathfrak{e}}(p - \underline{p}) \\
&= (\underline{\mathfrak{e}} + (\mathfrak{e} - \underline{\mathfrak{e}})\frac{p}{p - \underline{p}})(p - \underline{p})
\end{aligned}$$

$$\frac{\sigma_v^2 l^{-\beta_v} \mathrm{W}\phi(0,q,q,cq) + \sigma_a^2 l^{-\beta_a}}{\sigma_v^2 l^{-\beta_v} \mathrm{V}\phi(0,q) + \sigma_a^2 l^{-\beta_a}} - \underline{\mathfrak{e}} = (\mathfrak{e} - \underline{\mathfrak{e}})\frac{p}{p - \underline{p}}$$

Using Lemma B.1, **??** and Lemma B.5, we can see that the LHS is monotonic (increasing or decreasing) for large enough l. Therefore $\mathfrak{e}^{(l)}$ is a bounded monotonic sequence for large enough l, a fortiori it has a fixed point.

If we express $\mathfrak{e} = \mathfrak{e}^* + \epsilon$, $\gamma^{(l)} = (\mathfrak{e}^* + \epsilon^{(l)})p^{(l)}$, then

$$\frac{\sigma_v^2 l^{-\beta_v} \mathrm{W}\phi(0,q,q,cq) + \sigma_a^2 l^{-\beta_a}}{\sigma_v^2 l^{-\beta_v} \mathrm{V}\phi(0,q) + \sigma_a^2 l^{-\beta_a}} = \mathfrak{e}^* + \underline{\epsilon} + (\epsilon - \underline{\epsilon})\frac{p}{p - \underline{p}}$$

RHS It's easy to verify via Thm A.3 that either $\underline{p}/(p - \underline{p}) = \Theta(l \log l)$ (when $\hat{\mu}_p = 1$) or $\underline{p}/(p - \underline{p}) = \Theta(l)$ (all other cases). If $\epsilon - \underline{\epsilon} = \Omega((l \log l)^{-1})$, then $\epsilon = \Omega(\log\log l)$ (by Euler-MacLaurin formula), which would be a contradiction to $\epsilon = o(1)$. Thus $(\epsilon - \underline{\epsilon})\frac{\underline{p}}{p - \underline{p}} = o(1)$. [7] Hence the RHS goes to 0 with l.

LHS If $\beta_v > \beta_a$, then the LHS converges to 1. So $\mathfrak{e}^* = 1$.

If $\beta_v < \beta_a$, then the LHS $\sim \frac{\mathrm{W}\phi(0,q,q,cq)}{\mathrm{V}\phi(0,q)}$. As $l \to \infty$, $c \sim \mathfrak{e} \to \mathfrak{e}^*$, and $\mathrm{W}(0,q,q,\mathfrak{e}^* q) \to \frac{2}{\pi}\arcsin(\mathfrak{e}^*)$, and $\mathrm{V}(0,q) \to 1$. Therefore, $\mathfrak{e}^* = \frac{2}{\pi}\arcsin(\mathfrak{e}^*)$, for which there are 2 solutions, 0, the stable fixed point, and 1, the unstable fixed point. In particular, for all $\mathfrak{e}^{(0)} < 1$, $\mathfrak{e}^{(l)} \to 0$.

If $\beta_v = \beta_a$, then taking limits $l \to \infty$, we get

$$\frac{\sigma_v^2 \frac{2}{\pi}\arcsin(\mathfrak{e}^*) + \sigma_a^2}{\sigma_v^2 + \sigma_a^2} = \mathfrak{e}^*.$$

□

A.3 Residual Network with ReLU

The asymptotics of ReLU resnet depends on the following values:

$$\beta = \beta_v + \beta_b, \qquad \alpha = \min(\beta_v + \beta_b, \beta_a), \qquad \delta = \frac{1}{2}\sigma_v^2\sigma_w^2$$

Theorem A.6. *Let $\phi = \mathrm{ReLU}$. Then with $\delta = \frac{1}{2}\sigma_v^2\sigma_w^2$ and $\alpha = \min(\beta_v + \beta_b, \beta_a) = \beta_v + \min(\beta_b, \beta_a - \beta_v)$, we have the following asymptotics of p and q:*

- *If $\beta_v + \beta_w > 1$*
  - *If $\alpha \in [0, 1)$*

[7] In fact, since ϵ must be $o(\log\log\cdots\log l)$ for any chain of logs, $\epsilon - \underline{\epsilon} = o((l \log l \log\log l \cdots \log^{(k)} l)^{-1})$ for any k, where $\log^{(k)}$ is k-wise composition of log; so $(\epsilon - \underline{\epsilon})\frac{\underline{p}}{p - \underline{p}} = o((\log\log l \cdots \log^{(k)} l)^{-1})$ for any k.

    - $p = \Theta(l^{1-\alpha})$ *and* $q = \Theta(l^{\max(1-\alpha-\beta_w, -\beta_b)})$.
  - *If* $\alpha = 1$
    - $p = \Theta(\log l)$ *and* $q = \begin{cases} \Theta(l^{-\beta_w} \log l) & \text{if } \beta_w \le \beta_b \\ \Theta(l^{-\beta_b}) & \text{otherwise.} \end{cases}$
  - *If* $\alpha > 1$
    - $p = \Theta(1), p - \underline{p} = \Theta(l^{-\alpha})$ *and* $q = \Theta(l^{-\min(\beta_w, \beta_b)})$.
- *If* $\beta_v + \beta_w = 1$
  - *If* $\delta \neq 1 - \alpha$
    - $p = \Theta(l^{\max(\delta, 1-\alpha)})$, *and* $q = \Theta(l^{\max(\max(\delta,1-\alpha)-\beta_w, -\beta_b)})$.
  - *Otherwise*
    - $p = \Theta(l^{\delta} \log l)$ *and* $q = \begin{cases} \Theta(l^{\delta-\beta_w} \log l) & \text{if } \beta_b \ge \beta_w - \delta \\ \Theta(l^{-\beta_b}) & \text{otherwise} \end{cases}$
- *If* $\beta_v + \beta_w < 1$
  - $p, q = \exp(\frac{\delta}{1-\beta} l^{1-\beta} + \tilde{\Theta}(l^{\max(0,1-2\beta)}))$. *In particular,* $p = \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + R(l; \delta, \beta_v + \beta_w) + O(1))$, R *is the same* R *as in Lemma B.7, depending on only* l, δ, *and* $\beta_v + \beta_w$, *with*

$$R = \begin{cases} \Theta(l^{1-2\beta}) & \text{if } \beta < 1/2 \\ \Theta(\log l) & \text{if } \beta = 1/2 \\ \Theta(1) & \text{if } \beta > 1/2. \end{cases}$$

    and $q = \exp(\frac{\delta}{1-\beta} l^{1-\beta_v-\beta_w} + R(l; \delta, \beta_v + \beta_w) - \beta_w \log l + O(1))$.

In particular, $p, q = \mathsf{poly}(l)$ *if* $\beta_v + \beta_w \ge 1$.

Proof. Since $\mathrm{V}\phi(0, q) = \frac{1}{2}q$, we have

$$\begin{aligned} p - \underline{p} &= \sigma_v^2 l^{-\beta_v} \mathrm{V}\phi(0, q) + \sigma_a^2 l^{-\beta_a} \\ &= \frac{1}{2}\sigma_v^2 l^{-\beta_v}(\sigma_w^2 l^{-\beta_w} \underline{p} + \sigma_b^2 l^{-\beta_b}) + \sigma_a^2 l^{-\beta_a} \\ &= \frac{1}{2}\sigma_v^2 \sigma_w^2 l^{-\beta_v-\beta_w} \underline{p} + \frac{1}{2}\sigma_v^2 \sigma_b^2 l^{-\beta_v-\beta_b} + \sigma_a^2 l^{-\beta_a}. \end{aligned}$$

We apply Lemma B.8 with $\beta = \beta_v + \beta_w$, $\delta = \frac{1}{2}\sigma_v^2\sigma_w^2$, and $\alpha = \min(\beta_v + \beta_b, \beta_a) = \beta_v + \min(\beta_b, \beta_a - \beta_v)$ to upper and lower bound our dynamics here, with appropriately chosen C. With a little bit of case work, we obtain the desired results. □

Theorem A.7. *Let* $\phi = \mathrm{ReLU}$. *Suppose* $\mathfrak{e}^{(0)} < 1$.

- *If* $\beta_v + \beta_w \le 1$ *or* $\alpha \le 1$, *then* $\lim_{l\to\infty} \mathfrak{e}^{(l)} = 1$.
- *If* $\beta_v + \beta_w > 1$ *and* $\alpha > 1$, *then* $\mathfrak{e}^{(l)}$ *converges to a fixed point* $\mathfrak{e}^* < 1$, *dependent on the initial data* $p^{(0)}$ *and* $\gamma^{(0)}$, *at a rate of* $\Theta(l^{-\alpha+1})$ *where* $\alpha = \min(\beta_v + \beta_b, \beta_a)$.

Proof. By Cho and Saul (2009); Yang and Schoenholz (2017), we have $\mathrm{W}(0, q, q, cq) = \mathrm{V}\phi(0, q)\mathbb{J}_1(c) = \frac{1}{2}q \cdot \frac{1}{\pi}(\sqrt{1-c^2} + (\pi - \arccos(c))c)$. As in the proof of Thm A.5, we have

$$\frac{\sigma_v^2 l^{-\beta_v} q/2 \mathbb{J}_1(c) + \sigma_a^2 l^{-\beta_a}}{\sigma_v^2 l^{-\beta_v} q/2 + \sigma_a^2 l^{-\beta_a}} - \underline{\mathfrak{e}} = (\mathfrak{e} - \underline{\mathfrak{e}})\frac{p}{p - \underline{p}}$$

But $\frac{\sigma_v^2 l^{-\beta_v} q/2 \mathbb{J}_1(c) + \sigma_a^2 l^{-\beta_a}}{\sigma_v^2 l^{-\beta_v} q/2 + \sigma_a^2 l^{-\beta_a}} \ge \mathbb{J}_1(c) \ge c \ge \underline{\mathfrak{e}}$, where the last inequality holds for all large l, by Lemma B.1. So the LHS is nonnegative for large l, which implies $\mathfrak{e}^{(l)}$ is a bounded monotonically increasing sequence for large enough ls and thus has a limit $\mathfrak{e}^*$.

Writing $\mathfrak{e} = \mathfrak{e}^* + \epsilon$, we have

$$\mathfrak{e}^* + \underline{\epsilon} + (\epsilon - \underline{\epsilon})\frac{p}{p - \underline{p}} = \frac{\sigma_v^2 l^{-\beta_v} q/2 \mathbb{J}_1(c) + \sigma_a^2 l^{-\beta_a}}{\sigma_v^2 l^{-\beta_v} q/2 + \sigma_a^2 l^{-\beta_a}} \tag{$\star$}$$

Suppose $\beta_v + \beta_w \leq 1$ or $\alpha \leq 1$.

LHS. By Thm A.6, $p = \omega(1)$ (assuming $\sigma_v, \sigma_w > 0$), so that $p/(p - \underline{p}) = \tilde{O}(l)$. As in the proof of Thm A.5, $\epsilon - \underline{\epsilon}$ cannot be $\tilde{\Omega}(l^{-1})$, or else $\epsilon \to \infty$. Thus $(\epsilon - \underline{\epsilon})p/(p - \underline{p}) = o(1)$, and the LHS in the limit $l \to \infty$ becomes $\mathfrak{e}^*$.

RHS. In all cases, we will find $\mathfrak{e}^* = 1$.

If $q = o(l^{\beta_v - \beta_a})$, in the RHS $\sigma_a^2 l^{-\beta_a}$ dominates, and the RHS converges to 1.

If $q = \omega(l^{\beta_v - \beta_a})$, then in the RHS $\frac{1}{2}\sigma_v^2 l^{-\beta_v} q$ dominates. By Lemma B.1:

- If $p = \omega(l^{\beta_w - \beta_b})$, then $c \sim \mathfrak{e} \to \mathfrak{e}^*$, so that in the limit $l \to \infty$, $\mathfrak{e}^* = \mathbb{J}_1(\mathfrak{e}^*)$. This equation has solution $\mathfrak{e}^* = 1$.
- If $p = o(l^{\beta_w - \beta_b})$, then $c \to 1$, so that $\mathfrak{e}^* = \mathbb{J}_1(1) = 1$.
- If $p \sim Cl^{\beta_w - \beta_b}$, then $c \to \frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2}$, and we have an equation $\mathfrak{e}^* = \mathbb{J}_1(\frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2})$. Note that since $\mathfrak{e}^* \in [0, 1]$, $\frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2} = \frac{\sigma_w^2 C}{\sigma_w^2 C + \sigma_b^2} \cdot \mathfrak{e}^* + \frac{\sigma_b^2}{\sigma_w^2 C + \sigma_b^2} \cdot 1 \geq \mathfrak{e}^*$, with equality iff $\mathfrak{e}^* = 1$. Since $\mathbb{J}_1$ is monotonic, $\mathfrak{e}^* = \mathbb{J}_1(\frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2})$ iff the equality condition above is satisifed, i.e. $\mathfrak{e}^* = 1$.

If $q \sim Dl^{\beta_w - \beta_b}$ for some D, then the RHS is

$$\frac{\frac{1}{2}\sigma_v^2 D\mathbb{J}_1(c) + \sigma_a^2}{\frac{1}{2}\sigma_v^2 D + \sigma_a^2} + o(1)$$

by the same logic as in the proof of Lemma B.1. As in the above case, Lemma B.1 yields the following results:

- If $p = \omega(l^{\beta_w - \beta_b})$, then $c \sim \mathfrak{e} \to \mathfrak{e}^*$, so that in the limit $l \to \infty$, $\mathfrak{e}^* = \frac{\frac{1}{2}\sigma_v^2 D\mathbb{J}_1(\mathfrak{e}^*) + \sigma_a^2}{\frac{1}{2}\sigma_v^2 D + \sigma_a^2}$. Since the RHS of this equation is a convex combination of $\mathbb{J}_1(\mathfrak{e}^*)$ and 1, and $\mathbb{J}_1(\mathfrak{e}^*) \geq \mathfrak{e}^*$ by the monotonicity of $\mathbb{J}_1$, the equality can hold only if $\mathbb{J}_1(\mathfrak{e}^*) = \mathfrak{e}^*$. The only such $\mathfrak{e}^*$ is 1.
- If $p = o(l^{\beta_w - \beta_b})$, then $c \to 1$, so that $\mathfrak{e}^* = \frac{\frac{1}{2}\sigma_v^2 D\mathbb{J}_1(1) + \sigma_a^2}{\frac{1}{2}\sigma_v^2 D + \sigma_a^2} = 1$.
- If $p \sim Cl^{\beta_w - \beta_b}$, then $c \to \frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2}$, and we have an equation $\mathfrak{e}^* = \frac{1}{\frac{1}{2}\sigma_v^2 D + \sigma_a^2}(\frac{1}{2}\sigma_v^2 D\mathbb{J}_1(\frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2}) + \sigma_a^2)$. By the monotonicity of $\mathbb{J}_1$, the RHS is at least $\frac{1}{\frac{1}{2}\sigma_v^2 D + \sigma_a^2}(\frac{1}{2}\sigma_v^2 D(\frac{\sigma_w^2 C\mathfrak{e}^* + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2}) + \sigma_a^2)$, which is a convex combination of $\mathfrak{e}^*$ and 1. Since $\mathfrak{e}^* \leq 1$, the equality can hold only if $\mathfrak{e}^* = 1$.

Suppose $\beta_v + \beta_w > 1$ and $\alpha > 1$.

By Thm A.6, $p = \Theta(1)$ and therefore $\gamma = \Theta(1)$. Both converge to fixed points p^* and γ^* (possibly dependent on initial data $p^{(0)}$ and $\gamma^{(0)}$) because they are monotonically increasing sequences. Thus $\mathfrak{e}^* = \gamma^*/p^*$.

Unwinding the proof of Thm A.6, we see that $p - \underline{p} = l^{-\alpha}$, so that $p/(p - \underline{p}) = \Theta(l^{\alpha})$. Since the RHS of Eq. ($\star$) cannot blow up, it must be the case that $(\epsilon - \underline{\epsilon}) = O(l^{-\alpha}) \implies \epsilon = O(l^{-\alpha+1})$. If $(\epsilon - \underline{\epsilon})p/(p - \underline{p}) = \mathcal{F} + o(1)$ for some constant $\mathcal{F}$, then the LHS becomes $\mathfrak{e}^* + \mathcal{F}$ in the limit.

Yet, unless $\gamma^{(0)} = p^{(0)}$, $\mathfrak{e}^* < 1$. Therefore, $\mathcal{F} > 0$ (or else, like in the case of $\beta_v + \beta_w \leq 1$ or $\alpha \leq 1$, $\mathfrak{e}^* = 1$) whenever $\gamma^{(0)} < p^{(0)}$, and $\epsilon = \Theta(l^{-\alpha+1})$.

□

Theorem A.8. *Suppose $\phi = \mathrm{ReLU}$. Then for $\alpha = \min(\beta_v + \beta_b, \beta_a)$ and $\delta = \frac{1}{2}\sigma_v^2\sigma_w^2$ as in Thm A.6:*

- *If $\beta_v + \beta_w > 1$, then $\daleth^{(0)}/\daleth^{(l)} = \Theta(1)$.*
  - *If $\alpha \in [0, 1)$*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\alpha-1+\beta_v}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{-\max(1-\alpha-\beta_w,-\beta_b)}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1).$$

  - *If $\alpha = 1$*
    - *If $\beta_w \leq \beta_b$*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\beta_w}/\log l) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1)$$

    - *Otherwise*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\beta_b}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1)$$

  - *If $\alpha > 1$*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\min(\beta_w,\beta_b)}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1)$$

- *If $\beta_v + \beta_w = 1$, then $\daleth^{(0)}/\daleth^{(l)} = \Theta(l^\delta)$.*
  - *If $\delta \neq 1 - \alpha$*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\delta+\beta_v-\max(\delta,1-\alpha)}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\delta+\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\delta-\max(\max(\delta,1-\alpha)-\beta_w,-\beta_b)}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(l^\delta)$$

  - *Otherwise*
    - *If $\beta_b \geq \beta_w - \delta$*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\delta+\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\beta_w}/\log l) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(l^\delta)$$

    - *Otherwise*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\delta+\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\delta+\beta_b}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(l^\delta)$$

- *If $\beta_v + \beta_w < 1$, then $\daleth^{(0)}/\daleth^{(l)} = \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + \tilde{\Theta}(l^{\max(1-2\beta_v-2\beta_w,0)}))$.*

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(1)$$
$$\chi_b^{(1)}/\chi_b^{(l)} = \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + \tilde{\Theta}(l^{\max(1-2\beta_v-2\beta_w,0)}))$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\beta_w})$$
$$\chi_a^{(1)}/\chi_a^{(l)} = \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + \tilde{\Theta}(l^{\max(1-2\beta_v-2\beta_w,0)})).$$

Proof. Using Thm A.2 and the fact that $\mathrm{V}\phi(0,q) = \frac{1}{2}q, \mathrm{V}\dot{\phi}(0,q) = \frac{1}{2}$, we get

$$\daleth = (\frac{1}{2}\sigma_v^2\sigma_w^2 l^{-\beta_v-\beta_w} + 1)\daleth, \qquad \chi_b = \frac{1}{2}\sigma_v^2 l^{-\beta_v}\daleth,$$
$$\chi_w = \frac{1}{2}\sigma_v^2 l^{-\beta_v}\daleth p, \qquad \chi_v = \frac{1}{2}\daleth q, \qquad \chi_a = \daleth,$$

so

$$\frac{\chi_b^{(1)}}{\chi_b^{(l)}} = \frac{\frac{1}{2}\sigma_v^2\daleth^{(1)}}{\frac{1}{2}\sigma_v^2 l^{-\beta_v}\daleth^{(l)}} = l^{\beta_v}\frac{\daleth^{(0)}}{\daleth^{(l)}}$$
$$\frac{\chi_w^{(1)}}{\chi_w^{(l)}} = \frac{\frac{1}{2}\sigma_v^2\daleth^{(1)}p^{(0)}}{\frac{1}{2}\sigma_v^2 l^{-\beta_v}\daleth^{(l)}p^{(l-1)}} = l^{\beta_v}\frac{p^{(0)}}{p^{(l-1)}}\frac{\daleth^{(0)}}{\daleth^{(l)}}$$
$$\frac{\chi_v^{(1)}}{\chi_v^{(l)}} = \frac{\frac{1}{2}\daleth^{(1)}q^{(1)}}{\frac{1}{2}\daleth^{(l)}q^{(l-1)}} = \frac{q^{(0)}}{q^{(l-1)}}\frac{\daleth^{(0)}}{\daleth^{(l)}}$$
$$\frac{\chi_a^{(1)}}{\chi_a^{(l)}} = \frac{\daleth^{(1)}}{\daleth^{(l)}}.$$

Suppose $\beta_v + \beta_w > 1$. Then $\frac{\daleth^{(0)}}{\daleth^{(l)}} = \Theta(1)$ by Lemma B.7. By Thm A.6:

- If $\beta_v + \beta_w > 1$, then $\frac{\daleth^{(0)}}{\daleth^{(l)}} = \Theta(1)$ by Lemma B.7.
  - If $\alpha \in [0,1)$
    - $p = \Theta(l^{1-\alpha})$ and $q = \Theta(l^{\max(1-\alpha-\beta_w,-\beta_b)})$. So

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\alpha-1+\beta_v}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{-\max(1-\alpha-\beta_w,-\beta_b)}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1).$$

  - If $\alpha = 1$, then $p = \Theta(\log l)$ and
    - If $\beta_w \leq \beta_b$
      - $q = \Theta(l^{-\beta_w}\log l)$ so that

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\beta_w}/\log l) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1)$$

    - Otherwise
      - $q = \Theta(l^{-\beta_b})$ so that

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\beta_b}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1)$$

  - If $\alpha > 1$
    - $p = \Theta(1)$ and $q = \Theta(l^{-\min(\beta_w,\beta_b)})$. So

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\beta_v}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\min(\beta_w,\beta_b)}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(1)$$

- If $\beta_v + \beta_w = 1$, then by Lemma B.7 $\daleth^{(0)}/\daleth^{(l)} = \Theta(l^{\frac{1}{2}\sigma_v^2\sigma_w^2}) = \Theta(l^{\delta})$.
  - If $\delta \neq 1-\alpha$
    - $p = \Theta(l^{\max(\delta,1-\alpha)})$, and $q = \Theta(l^{\max(\max(\delta,1-\alpha)-\beta_w,-\beta_b)})$. So

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\delta+\beta_v-\max(\delta,1-\alpha)}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\delta+\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\delta-\max(\max(\delta,1-\alpha)-\beta_w,-\beta_b)}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(l^{\delta})$$

  - Otherwise, $p = \Theta(l^{\delta}\log l)$, and

- If $\beta_b \geq \beta_w - \delta$
  - $q = \Theta(l^{\delta-\beta_w}\log l)$, so

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\delta+\beta_v-\delta}/\log l) = \Theta(l^{\beta_v/\log l}) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\delta+\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\delta+\beta_w-\delta}/\log l) = \Theta(l^{\beta_v/\log l}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(l^{\delta})$$

- Otherwise
  - $q = \Theta(l^{-\beta_b})$, so

$$\chi_w^{(1)}/\chi_w^{(l)} = \Theta(l^{\delta+\beta_v-\delta}/\log l) = \Theta(l^{\beta_v}/\log l) \qquad \chi_b^{(1)}/\chi_b^{(l)} = \Theta(l^{\delta+\beta_v})$$
$$\chi_v^{(1)}/\chi_v^{(l)} = \Theta(l^{\delta+\beta_b}) \qquad \chi_a^{(1)}/\chi_a^{(l)} = \Theta(l^{\delta})$$

- If $\beta_v + \beta_w < 1$, then by Lemma B.7, $\daleth^{(0)}/\daleth^{(l)} = \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + R(l;\delta,\beta_v+\beta_w))$, where $R = \tilde{\Theta}(l^{\max(1-2\beta_v-2\beta_w,0)})$ is as in Lemma B.7.
  - We have $p = \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + R(l;\delta,\beta_v+\beta_w) + O(1))$, where R is the same as above, and $q = \exp(\frac{\delta}{1-\beta} l^{1-\beta_v-\beta_w} + R(l;\delta,\beta_v+\beta_w) - \beta_w \log l + O(1))$. Thus

$$\begin{aligned}
\chi_w^{(1)}/\chi_w^{(l)} &= \exp(O(1)) = \Theta(1)\\
\chi_b^{(1)}/\chi_b^{(l)} &= \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + R + \beta_v \log l)\\
\chi_v^{(1)}/\chi_v^{(l)} &= \exp(\beta_w \log l + O(1)) = \Theta(l^{\beta_w})\\
\chi_a^{(1)}/\chi_a^{(l)} &= \exp(\frac{\delta}{1-\beta_v-\beta_w} l^{1-\beta_v-\beta_w} + R).
\end{aligned}$$

□

B Lemmas

In this section, we present many lemmas used in the proofs of the main theorems. In all cases, the lemmas here have already appeared in some form in Yang and Schoenholz (2017), and for completeness, we either include them and their proofs here or improve upon and extend them, with the blessing of the authors.

Lemma B.1. *Asymptotically,*

$$c = \frac{\sigma_w^2 l^{-\beta_w}\underline{\gamma} + \sigma_b^2 l^{-\beta_b}}{\sigma_w^2 l^{-\beta_w}\underline{p} + \sigma_b^2 l^{-\beta_b}}$$
$$= \begin{cases}
1 - \sigma_w^2\sigma_b^{-2} l^{\beta_b-\beta_w}\underline{p}(1-\underline{\gamma}/\underline{p}) + \Theta(l^{2(\beta_b-\beta_w)}\underline{p}^2(1-\underline{\gamma}/\underline{p}))\\
\qquad = 1 + O(l^{\beta_b-\beta_w}\underline{p}) & \text{if } \underline{p} = o(l^{\beta_w-\beta_b})\\
\underline{\gamma}/\underline{p} + \sigma_b^2\sigma_w^{-2} l^{\beta_w-\beta_b}\underline{p}^{-1}(\underline{\gamma}/\underline{p} - 1) + \Theta(l^{2(\beta_w-\beta_b)}\underline{p}^{-2}(\underline{\gamma}/\underline{p}-1))\\
\qquad = \underline{\gamma}/\underline{p} + O(l^{\beta_w-\beta_b}\underline{p}^{-1}) & \text{if } \underline{p} = \omega(l^{\beta_w-\beta_b})\\
\frac{\sigma_w^2 C\underline{\gamma}/\underline{p}+\sigma_b^2}{\sigma_w^2 C+\sigma_b^2} + \mathcal{E}Rl^{\beta_b-\beta_w}(\underline{\gamma}/\underline{p}-1) + \Theta((Rl^{\beta_b-\beta_w})^2(\underline{\gamma}/\underline{p}-1))\\
\qquad = \frac{\sigma_w^2 C\underline{\gamma}/\underline{p}+\sigma_b^2}{\sigma_w^2 C+\sigma_b^2} + o(1) & \text{if } \underline{p} = Cl^{\beta_w-\beta_b} + R \text{ for some constant } C \text{ and remainder } R = o(l^{\beta_w-\beta_b}).
\end{cases}$$

where $\mathcal{E} = \frac{\sigma_w^2\sigma_b^2}{(\sigma_w^2 C+\sigma_b^2)^2}$.

Proof. If $\underline{p} = o(l^{\beta_w - \beta_b})$, then

$$
\begin{aligned}
(\sigma_b^2 l^{-\beta_b} + \sigma_w^2 l^{-\beta_w}\underline{p})^{-1} &= \sigma_b^{-2} l^{\beta_b}(1 + \sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p})^{-1} \\
&= \sigma_b^{-2} l^{\beta_b}(1 - \sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p} + (\sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p})^2 - \cdots) \\
\frac{\sigma_w^2 l^{-\beta_w}\underline{\gamma} + \sigma_b^2 l^{-\beta_b}}{\sigma_w^2 l^{-\beta_w}\underline{p} + \sigma_b^2 l^{-\beta_b}} &= (\sigma_w^2 l^{-\beta_w}\underline{\gamma} + \sigma_b^2 l^{-\beta_b})\sigma_b^{-2} l^{\beta_b}(1 - \sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p} + (\sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p})^2 - \cdots) \\
&= (\sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{\gamma} + 1)(1 - \sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p} + (\sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}\underline{p})^2 - \cdots) \\
&= 1 - \sigma_w^2 \sigma_b^{-2} l^{\beta_b - \beta_w}(\underline{p} - \underline{\gamma}) + \Theta((l^{\beta_b - \beta_w})^2 \underline{p}(\underline{p} - \underline{\gamma})) \\
&= 1 + O(l^{\beta_b - \beta_w}\underline{p}).
\end{aligned}
$$

If $\underline{p} = \omega(l^{\beta_w - \beta_b})$, then

$$
\begin{aligned}
(\sigma_w^2 l^{-\beta_w}\underline{p} + \sigma_b^2 l^{-\beta_b})^{-1} &= (\sigma_w^2 l^{-\beta_w}\underline{p})^{-1}(1 + \sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-1})^{-1} \\
&= (\sigma_w^2 l^{-\beta_w}\underline{p})^{-1}(1 - \sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-1} + (\sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-1})^2 - \cdots) \\
\frac{\sigma_w^2 l^{-\beta_w}\underline{\gamma} + \sigma_b^2 l^{-\beta_b}}{\sigma_w^2 l^{-\beta_w}\underline{p} + \sigma_b^2 l^{-\beta_b}} &= (\underline{\gamma}/\underline{p} + \sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-1})(1 - \sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-1} + (\sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-1})^2 - \cdots) \\
&= \underline{\gamma}/\underline{p} + \sigma_b^2 \sigma_w^{-2} l^{\beta_w - \beta_b}\underline{p}^{-2}(\underline{\gamma} - \underline{p}) + \Theta(l^{2(\beta_w - \beta_b)}\underline{p}^{-3}(\underline{\gamma} - \underline{p})) \\
&= \underline{\gamma}/\underline{p} + O(l^{\beta_w - \beta_b}\underline{p}^{-1}).
\end{aligned}
$$

If $\underline{p} \sim C l^{\beta_w - \beta_b}$ for some C and $R = \underline{p} - C l^{\beta_w - \beta_b} = o(l^{\beta_w - \beta_b})$, then

$$
\begin{aligned}
\sigma_w^2 l^{-\beta_w}\underline{p} + \sigma_b^2 l^{-\beta_b} &= l^{-\beta_w}\underline{p}(\sigma_w^2 + \frac{\sigma_b^2 l^{-\beta_b}}{l^{-\beta_w}\underline{p}}) \\
&= l^{-\beta_w}\underline{p}(\sigma_w^2 + \frac{\sigma_b^2}{C + R l^{\beta_b - \beta_w}}) \\
&= \frac{l^{-\beta_w}\underline{p}}{C + R l^{\beta_b - \beta_w}}(\sigma_w^2(C + R l^{\beta_b - \beta_w}) + \sigma_b^2) \\
\sigma_w^2 l^{-\beta_w}\underline{\gamma} + \sigma_b^2 l^{-\beta_b} &= \frac{l^{-\beta_w}\underline{p}}{C + R l^{\beta_b - \beta_w}}(\sigma_w^2 \underline{\gamma}/\underline{p}(C + R l^{\beta_b - \beta_w}) + \sigma_b^2) \\
\frac{\sigma_w^2 l^{-\beta_w}\underline{\gamma} + \sigma_b^2 l^{-\beta_b}}{\sigma_w^2 l^{-\beta_w}\underline{p} + \sigma_b^2 l^{-\beta_b}} &= \frac{\sigma_w^2 \underline{\gamma}/\underline{p}(C + R l^{\beta_b - \beta_w}) + \sigma_b^2}{\sigma_w^2(C + R l^{\beta_b - \beta_w}) + \sigma_b^2} \\
&= (\frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2} + \frac{\sigma_w^2 \underline{\gamma}/\underline{p} R l^{\beta_b - \beta_w}}{\sigma_w^2 C + \sigma_b^2})(1 - \frac{\sigma_w^2 R l^{\beta_b - \beta_w}}{\sigma_w^2 C + \sigma_b^2} + \cdots) \\
&= \frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2} + R l^{\beta_b - \beta_w}(\frac{\sigma_w^2 \underline{\gamma}/\underline{p}}{\sigma_w^2 C + \sigma_b^2} - \frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2}\frac{\sigma_w^2}{\sigma_w^2 C + \sigma_b^2}) + \cdots \\
&= \frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2} + R l^{\beta_b - \beta_w}\frac{\sigma_w^2}{\sigma_w^2 C + \sigma_b^2}(\underline{\gamma}/\underline{p} - \frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2}) + \cdots \\
&= \frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2} + R l^{\beta_b - \beta_w}\frac{\sigma_w^2}{\sigma_w^2 C + \sigma_b^2}\frac{\sigma_b^2(\underline{\gamma}/\underline{p} - 1)}{\sigma_w^2 C + \sigma_b^2} + \Theta((R l^{\beta_b - \beta_w})^2(\underline{\gamma}/\underline{p} - 1)) \\
&= \frac{\sigma_w^2 \underline{\gamma}/\underline{p} C + \sigma_b^2}{\sigma_w^2 C + \sigma_b^2} + o(1)
\end{aligned}
$$

□

For any function f that is $(k+1)$-times differentiable in a neighborhood of 0, we have the asymptotic expansion

$$
f(z) = \sum_{n=0}^{k} \frac{d^n f}{dz^n}(0)\frac{z^n}{n!} + O(z^{k+1}), \text{as } z \to 0.
$$

Since

$$\frac{d^n}{d(1/q)^n} q^{1/2} \mathrm{V}\phi(0,q)\bigg|_{q\to\infty} = \frac{(-1)^n}{2^n\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi^2(z) z^{2n}\,\mathrm{d}z$$

whenever the RHS is integrable, we have

Lemma B.2. *Suppose $\phi^2(z)z^{2n}$ is integrable over $z \in \mathbb{R}$ for all $0 \le n \le N+1$. Then $\mathrm{V}\phi(0,q) = q^{-1/2}(\sum_{n=0}^{N} C_n q^{-n} + O(q^{-N-1}))$ as $q \to \infty$, where*

$$C_n := \frac{(-1)^n}{2^n n!\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi^2(z) z^{2n}\,\mathrm{d}z.$$

Note that $\operatorname{sech}^d(z) = \Theta(e^{-d|z|})$ for $z \to \infty$ as long as $d > 0$, so that C_n from the above result converges when $\phi = \operatorname{sech}^d$. Therefore

Lemma B.3. *Let $d > 0$. We have $\mathrm{V}\operatorname{sech}^d(0,q) \simeq q^{-1/2}\sum_{n\ge 0} C_n q^{-n}$, where*

$$C_n := \frac{(-1)^n}{2^n n!\sqrt{2\pi}} \int_{-\infty}^{\infty} \operatorname{sech}^{2d}(z) z^{2n}\,\mathrm{d}z.$$

As corollaries, we obtain the following asymptotics.

Lemma B.4. $\mathrm{V}\dot{\tanh}(0,q) = \frac{2}{3}\sqrt{\frac{2}{\pi}} q^{-1/2} + \Theta(q^{-3/2})$ *as $q \to \infty$.*

Proof. Use Lemma B.3 along with the fact that $\dot{\tanh}(z) = \operatorname{sech}^2(z)$ and $\int \operatorname{sech}^4 z\,\mathrm{d}z = \frac{2}{3}\tanh z + \frac{1}{2}\operatorname{sech}^2 z \tanh z$. □

Lemma B.5. $1 - \mathrm{V}\tanh(0,q) = \sqrt{\frac{2}{\pi}} q^{-1/2} + \Theta(q^{-3/2})$ *as $q \to \infty$.*

Proof. Use Lemma B.3 along with the fact that $1 - \tanh^2(z) = \operatorname{sech}^2(z)$ and $\int \operatorname{sech}^2 z\,\mathrm{d}z = \tanh z$. □

Lemma B.6. *Let $d \in \mathbb{R}$ and $1 < M < N$ with $N - M \in \mathbb{Z}^{\ge 0}$. Set $\Sigma(M,N,d) := \sum_{a=M}^{N} a^d$. If we fix M and let $N \to \infty$,*

$$\Sigma(M,N,d) = \begin{cases} \Theta(1) & \text{if } d < -1 \\ \log N + O(1) & \text{if } d = -1 \\ \frac{N^{d+1}}{d+1} + O(1) & \text{if } -1 < d < 0 \\ N - M + 1 & \text{if } d = 0 \\ \frac{1}{d+1}N^{d+1} + \frac{1}{2}N^d + O(N^{\max(0,d-1)}) & \text{if } d > 0 \end{cases}$$

Proof. Euler-MacLaurin formula. □

Lemma B.7. *Suppose $\epsilon^{(l)}$ satisfies the recurrence*

$$\epsilon^{(l)} = \epsilon^{(l-1)}(1 + \frac{\delta}{l^\beta}).$$

for some nonzero constant $\delta \in \mathbb{R}$ independent of l.

- *If $\beta > 1$, then $\epsilon^{(l)} = \Theta(1)$.*
- *If $\beta = 1$, then $\epsilon^{(l)} = \Theta(l^\delta)$.*
- *If $0 < \beta < 1$, then $\epsilon^{(l)} = \exp(\frac{\delta}{1-\beta} l^{1-\beta} + \tilde{\Theta}(l^{\max(0,1-2\beta)}))$. In particular, $\epsilon^{(l)} = \exp(\frac{\delta}{1-\beta} l^{1-\beta} + R)$, where remainder R is*

$$R = \begin{cases} \Theta(l^{1-2\beta}) & \text{if } \beta < 1/2 \\ \Theta(\log l) & \text{if } \beta = 1/2 \\ \Theta(1) & \text{if } \beta > 1/2 \end{cases}$$

Proof. We have

$$\begin{aligned}\log \epsilon^{(l)} &= \log \epsilon^{(l-1)} + \log(1 + \delta/l^{\beta}) \\ &= \log \epsilon^{(l-1)} + \delta/l^{\beta} + \Theta(\delta^2/l^{2\beta})\end{aligned}$$

for large l. If $\beta > 1$, then $\sum_l l^{-\beta}$ converges, and

$$\begin{aligned}\log \epsilon^{(l)} &= \log \epsilon^{(0)} - \Theta(1) \\ \epsilon^{(l)} &= \Theta(1).\end{aligned}$$

If $\beta = 1$, then

$$\begin{aligned}\log \epsilon^{(l)} &= \log \epsilon^{(0)} + \delta \log l + \Theta(1) \\ \epsilon^{(l)} &= \Theta(l^{\delta}).\end{aligned}$$

If $\beta < 1$, then

$$\log \epsilon^{(l)} = \log \epsilon^{(0)} + \frac{\delta}{1-\beta} l^{1-\beta} + R$$

for a remainder R that is

$$\begin{cases}\Theta(l^{1-2\beta}) & \text{if } \beta < 1/2 \\ \Theta(\log l) & \text{if } \beta = 1/2 \\ \Theta(1) & \text{if } \beta > 1/2\end{cases}$$

Exponentiating gives the desired result. □

Lemma B.8. *Suppose $\epsilon^{(l)} = Cl^{-\alpha} + \epsilon^{(l-1)}(1 + \delta/l^{\beta})$ for $\alpha \in \mathbb{R}$, $C \neq 0$, and $\delta \neq 0$. Then*

- *If $\beta > 1$, then*
  - *$\epsilon^{(l)} = \Theta(l^{1-\alpha})$ if $\alpha \in [0, 1)$;*
  - *$\epsilon^{(l)} = \Theta(\log l)$ if $\alpha = 1$;*
  - *$\epsilon^{(l)} = \Theta(1)$ if $\alpha > 1$.*
- *If $\beta = 1$, then*
  - *$\epsilon^{(l)} = \Theta(l^{\max(\delta, 1-\alpha)})$ if $1 - \delta \neq \alpha$.*
  - *$\epsilon^{(l)} = \Theta(l^{\delta} \log l)$ if $1 - \delta = \alpha$.*
- *If $\beta < 1$, then*
  - *$\epsilon^{(l)} = \exp(\frac{\delta}{1-\beta} l^{1-\beta} + \tilde{\Theta}(l^{\max(0, 1-2\beta)}))$. In particular, $\epsilon^{(l)} = \exp(\frac{\delta}{1-\beta} l^{1-\beta} + R(l; \delta, \beta) + O(1))$, where R is the same R as in Lemma B.7, depending on only l, δ, and β but not on α or C, with*

$$R = \begin{cases}\Theta(l^{1-2\beta}) & \text{if } \beta < 1/2 \\ \Theta(\log l) & \text{if } \beta = 1/2 \\ \Theta(1) & \text{if } \beta > 1/2.\end{cases}$$

Furthermore, for $\beta = -\delta = 1$: $\epsilon^{(l)} \sim l^{-1}$ if $\alpha > 2$, $\epsilon^{(l)} \sim l^{1-\alpha}$ if $\alpha < 2$, and $\epsilon^{(l)} \sim l^{\delta} \log l$ if $\alpha = 2$.

Proof. We can unwind the recurrence to get

$$\epsilon^{(l)} = C \sum_{m=1}^{l} m^{-\alpha} \prod_{n=m+1}^{l} (1 + \frac{\delta}{n^{\beta}}) + \epsilon^{(0)} \prod_{n=1}^{l} (1 + \frac{\delta}{n^{\beta}})$$

Suppose $\beta > 1$. By Lemma B.7, we get

$$\begin{aligned}\epsilon^{(l)} &= \Theta(1)\sum_{m=1}^{l} m^{-\alpha} + \epsilon^{(0)}\Theta(1) \\ &= \begin{cases}\Theta(l^{1-\alpha}) & \text{if } \alpha \in [0,1) \\ \Theta(\log l) & \text{if } \alpha = 1 \\ \Theta(1) & \text{if } \alpha > 1.\end{cases}\end{aligned}$$

Now suppose $\beta = 1$. By Lemma B.7, we get

$$\epsilon^{(l)} = \sum_{m=1}^{l} m^{-\alpha}\Theta(m^{-\delta}l^{\delta}) + \epsilon^{(0)}\Theta(l^{\delta})$$

where the constants hidden inside the Θ are the same in every term of the sum. If $\alpha > 1-\delta$, then $m^{-\delta-\alpha} = o(m^{-1})$, so that $\sum_{m=1}^{l} m^{-\delta-\alpha} = \Theta(1)$, and

$$\begin{aligned}\epsilon^{(l)} &= \Theta(l^{\delta}) + \epsilon^{(0)}\Theta(l^{\delta}) \\ &= \Theta(l^{\delta}).\end{aligned}$$

On the other hand, if $\alpha < 1-\delta$, then $\sum_{m=1}^{l} m^{-\delta-\alpha} = \Theta(l^{1-\delta-\alpha})$. So

$$\begin{aligned}\epsilon^{(l)} &= \Theta(l^{1-\alpha}) + \epsilon^{(0)}\Theta(l^{\delta}) \\ &= \Theta(l^{1-\alpha}).\end{aligned}$$

If $\alpha = 1-\delta$, then $\sum_{m=1}^{l} m^{-\delta-\alpha} = \Theta(\log l)$. So

$$\begin{aligned}\epsilon^{(l)} &= \Theta(l^{\delta}\log l) + \epsilon^{(0)}\Theta(l^{\delta}) \\ &= \Theta(l^{\delta}\log l).\end{aligned}$$

Finally, if $\beta \in (0,1)$, then by Lemma B.7,

$$\epsilon^{(l)} = Ce^{\frac{\delta}{1-\beta}l^{1-\beta}+R}\sum_{m=1}^{l} m^{-\alpha}e^{\frac{-\delta}{1-\beta}m^{1-\beta}+\tilde{\Theta}(l^{\max(0,1-2\beta)})} + e^{\frac{\delta}{1-\beta}l^{1-\beta}+R},$$

where R is the remainder as in Lemma B.7. The sum $\sum_{m=1}^{l} m^{-\alpha}e^{\frac{-\delta}{1-\beta}m^{1-\beta}+\tilde{\Theta}(l^{\max(0,1-2\beta)})}$ can be upper and lower bounded by integrals of the form $\int x^{-\alpha}e^{-\mu x^{1-\beta}}\,\mathrm{d}x$ for appropriate μs, which are finite (they are bounded by values of incomplete Gamma functions). Thus $\epsilon^{(l)} = \Theta(1)e^{\frac{\delta}{1-\beta}l^{1-\beta}+R}$ where

$$R' = R + O(1) = \begin{cases}\Theta(l^{1-2\beta}) & \text{if } \beta < 1/2 \\ \Theta(\log l) & \text{if } \beta = 1/2 \\ \Theta(1) & \text{if } \beta > 1/2.\end{cases}$$

A fortiori, $\epsilon^{(l)} = e^{\frac{\delta}{1-\beta}l^{1-\beta}+\tilde{\Theta}(l^{\max(0,1-2\beta)})}$.

For our "furthermore" claim: the case of $\delta = -1$ telescopes, so that the upper and lower constants hidden in Θ can both be taken to be 1. □